IEEE Signal Processing Society | IEEE Transactions on Signal Processing

Recovery of sparse translation-invariant signals with continuous basis pursuit

Journal:	*Transactions on Signal Processing*
Manuscript ID:	T-SP-11251-2010
Manuscript Type:	Regular Paper
Date Submitted by the Author:	31-Dec-2010
Complete List of Authors:	Ekanadham, Chaitanya; New York University, Courant Institute of Mathmatical Sciences Tranchina, Daniel; Courant Institute of Mathematical Sciences, Mathematics Simoncelli, Eero; Courant Institute of Mathematical Sciences/Howard Hughes Medical Institute, Mathematics & Neural Science
EDICS:	SSP-APPL Applications of statistical signal processing techniques < Statistical Signal Processing, SSP-DECO Deconvolution < Statistical Signal Processing

Recovery of sparse translation-invariant signals with continuous basis pursuit

Chaitanya Ekanadham, Daniel Tranchina, and Eero Simoncelli, *Fellow, IEEE*

Abstract

We consider the problem of decomposing a signal into a linear combination of features, each a continuously translated version of one of a small set of elementary features. Although these constituents are drawn from a *continuous family*, most current signal decomposition methods rely on a finite dictionary of discrete examples selected this family (e.g., a set of shifted copies of a set of basic waveforms), and apply sparse optimization methods to select and solve for the relevant coefficients. Here, we generate a dictionary that includes auxilliary interpolation functions that approximate local continuous translates of features via constrained adjustment of their coefficients. We formulate a constrained convex optimization problem, in which the full set of dictionary coefficients represent a linear approximation of the signal, the auxiliary coefficients are constrained so as to only represent translated features, and sparsity is imposed on the non-auxiliary coefficients using an L1 penalty. The well-known *basis pursuit denoising* (BP) method may be seen as a special case, in which the auxiliary interpolation functions are omitted, and we thus refer to our methodology as *continuous basis pursuit* (CBP). We develop two implementations of CBP for a one-dimensional translation-invariant source, one using a first-order Taylor approximation, and another using a form of trigonometric spline. We examine the tradeoff between sparsity and signal reconstruction accuracy in these methods, demonstrating empirically that trigonometric CBP significantly outperforms Taylor CBP, which in turn offers significant gains over ordinary BP. In addition, the CBP bases can generally achieve equally good or better approximations with much coarser sampling than BP, leading to a reduction in dictionary dimensionality.

C.E. is with the Courant Institute of Mathematical Sciences (CIMS), New York University, NY 10003 (e-mail: chaitu@math.nyu.edu).

D.T. is with CIMS, the Center for Neural Science (CNS), and the Department of Biology, NYU, NY 10003 (e-mail: tranchin@courant.nyu.edu).

E.P.S. is with the Howard Hughes Medical Institute (HHMI), CNS, and CIMS, NYU, NY 10012 (e-mail: eero.simoncelli@nyu.edu).

This work was partially funded by NYU through a McCracken Fellowship to C.E., and by an HHMI Investigatorship to E.P.S.

I. Introduction

The decomposition of a signal into a sparse linear combination of features is an important and well-studied problem, and plays a central role in many applications. A surge of recent effort focuses on representing a signal as a noisy superposition of the smallest possible subset of functions drawn from a large finite dictionary. The standard formulation tries to minimize the L_0 pseudonorm (number of nonzero elements) of the vector of weights corresponding to the dictionary elements.

The finite dictionary of basis functions $\{\phi_k(t)\}$ may be fixed in advance, or optimized (so as to best represent an ensemble of signals). In general, this objective can only be minimized via exhaustive search of all 2^d subsets of the dictionary, making it infeasible in practice. However, two broad classes of approximate solutions have been widely studied in the literature. The first consists of greedy methods, dating back to variable selection methods in the 1970s ([1]). These methods are exemplified by the well-known "matching pursuit" algorithm of Mallat and Zhang [2], and include a variety of more recent "iterative thresholding" methods [3], [4], [5], [6]. The general idea is to solve sequentially for the nonzero elements of $\vec{x}$, at each step choosing the element(s) that best explain the current residual. A second category of solutions arises from convex relaxations of the L_0 objective, and include the LASSO [7], the basis pursuit denoising (BP) algorthm [8], and the Dantzig selector [9], each of which employ the convex L_1 norm. Results by Tibshirani [7] and Chen et. al [8] show that substituting an L_1 penalty makes the problem solvable using quadratic programming and yields solutions with a high degree of sparsity. Recent publications [10], [11] provide conditions on the dictionary that guarantee this approximation to be near-optimal.

Most objective functions that have been utilized for sparse decomposition are constructed around the premise of linear superposition and additive noise, and make no assumptions about the structure of the dictionary. However, many real signals are generated by processes that obey natural invariances (e.g., translation-invariance, dilation-invariance, rotation-invariance). In this setting, the goal is to identify feature instances in the signal along with their associated amplitudes and transformation parameters. With a translation-invariant signal in time, for example, one aims to identify the amplitudes and timeshifts of the features. In the majority of published examples, the problem is solved by constructing a finite dictionary that reflects the invariant structure: one

discretely samples the transformation parameters and applies these to a finite set of elementary features. For example, dictionaries for sound processing, whether learned or hand-constructed, are commonly "convolutional", containing time-delayed copies of template waveforms (e.g., [12]).

Dictionaries for image representation typically contain features that are translated, and in some cases, dilated and rotated (e.g., [13]).[1] This discrete sampling approach replaces the full nonlinear problem with a more tractable linear inverse problem. However, the ability of the discrete dictionary to accurately represent signals depends critically on the spacing at which the dictionary was sampled. In general, a very fine sampling is required, resulting in a very large and ill-conditioned dictionary. This ill-conditioning, in turn, is unfavorable for the relaxation approximations mentioned above. Furthermore, given this representation, it is still unclear how to estimate the true amplitudes and transformation parameters associated with the recovered features.

Here, we propose an alternative linear approximation to the full nonlinear problem. We focus on the problem of translation-invariant one-dimensional signals (although the methods generalize to other transformations, and higher dimensions). We construct a group of functions that can span local translations of the feature templates via continuous variation of their coefficients. As a concrete example, consider the original templates and their derivatives, which can approximate local translations through a first-order Taylor approximation. The resulting dictionary can generally approximate the true set of scaled and translated templates more accurately than a dictionary of equal size containing only translated copies of the feature itself (i.e. the special case in which the interpolating group is just the template). A signal of interest is then represented in this dictionary by "block-sparse" coefficients, where each non-zero coefficient block represents an amplitude-scaled and translated template. We formulate an objective function in which the coefficients are constraind so as to only represent scaled/transformed templates, and use an L_1 penalty to impose sparsity on the blocks. The advantage of this approach over ordinary BP is three-fold: (1) better approximation of translation-invariant signals, (2) a smaller basis, which leads to sparser solutions via convex optimization, and (3) an explicit mapping from this representation to amplitudes and transformation parameters.

[1]Many examples of sparse decomposition on images have been applied to nonoverlapping square blocks of pixels (e.g., [14], [15], [16]), but the effective dictionary for representing the entire image is the union of dictionary elements for each block, and thus consists of translated copies of the block dictionary elements.

II. Problem formulation

We begin by formulating a simple generative model for translation-invariant signals, as well as the maximum a posteriori (MAP) estimation framework for inferring the most likely parameters given the observed signal. Assume we observe a signal that is a noisy superposition of scaled time-shifted copies of a single known elementary waveform $f(t)$ on a finite interval $[0, T]$:

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

where $\eta(t)$ is a Gaussian white noise process with power σ^2, the *event times* $\{\tau_j\}$ are drawn from a Poisson process with rate μ, and the *event amplitudes* $\{a_j\}$ are drawn independently from a density $P_A(a)$. The inverse (inference) problem is then to recover the most likely values of parameters $\{\tau_j, a_j\}$ given $y(t)$. This amounts to maximizing the posterior distribution $P(\{\tau_j, a_j\}|y(t))$, which reduces, on taking the negative log, to solving:

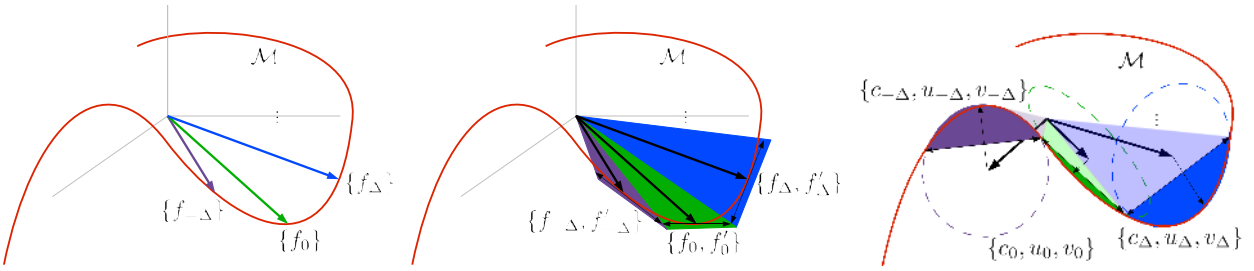
$$\min_{N,\{\tau_j,a_j\}} \frac{1}{2\sigma^2}\|y(t) - \sum_{j=1}^{N} a_j f(t - \tau_j)\|_2^2 + N\log(\mu) - \sum_{j=1}^{N} \log P_A(a_j) \tag{2}$$

This *sparse deconvolution* formulation has been used to describe many real-world problems including seismogram analysis [17], neural spike sorting [18], acoustic signal analysis [12], and image processing [13]. Unfortunately, solving Eq. (2) directly is intractable, due to the discrete nature of N and the nonlinearity embedding of the τ_j's within the argument of the waveform $f(\cdot)$. It is thus desirable to find alternative formulations that (i) approximate the signal posterior distribution well, (ii) have parameters that can be tractably estimated, and (iii) have an intuitive mapping back to the original representation.

III. Conventional solution: discretization and BP

A standard simplification of the problem is to discretize the event times at a spacing that is fine enough that the Poisson process is well-approximated by a Bernoulli process. The interval $[0, T]$ is divided into $N_\Delta = \lceil T/\Delta \rceil$ time bins of size Δ, where the probability of an event in each bin is $\mu\Delta$, for Δ sufficiently small.[2] This discrete process is represented by a vector $\vec{x} \in \mathbb{R}^{N_\Delta}$, whose

[2]The probability of two or more events is $O(\Delta^2)$, which is negligible for Δ small.



(a) (b) (c)

Fig. 1. Illustration of the three approximations of the manifold of translates of the waveform, $\mathcal{M}_{f,T}$. (a) The standard basis pursuit (BP) dictionary, F_Δ, as used in Eq. (6), consists of discrete time-shifts of the waveform $f(t)$. (b) Continuous basis pursuit with first-order Taylor interpolator (CBP-T), as specified by Eq. (12). Each pair of functions, $(f_{k\Delta}, f'_{k\Delta})$, with properly constrained coefficients, represents a triangular region of the space (shaded regions). (c) Continuous basis pursuit with polar interpolation (CBP-P), as specified by Eq. (18). Each triplet of functions, $(c_{k\Delta}, u_{k\Delta}, v_{k\Delta})$, represents the surface of a cone (see Fig. 3(b) for parameterization).

elements x_k are interpreted as the amplitude of any event in the interval $(\frac{(2n-1)\Delta}{2}, \frac{(2n+1)\Delta}{2})$. The corresponding prior probability distribution on each x_k is a mixture of a point mass at zero, and $P_A(\cdot)$:

$$P(\vec{x}) = \prod_{k=1}^{N_\Delta} [(1-\mu\Delta)\delta(x_k) + (\mu\Delta)P_A(x_k)] \tag{3}$$

The MAP estimate for this approximate model is obtained by solving:

$$\min_{\vec{x}} \frac{1}{2\sigma^2}\|y(t) - (F_\Delta \vec{x})(t)\|_2^2 - \log(\mu\Delta)\|\vec{x}\|_0 - \sum_{k:x_k\neq 0} \log P_A(x_k) \tag{4}$$

where F_Δ is a linear operator that contains a fixed dictionary of time-shifted copies of $f(\cdot)$:

$$(F_\Delta \vec{x})(t) := \sum_{k=1}^{N_\Delta} x_k f(t - k\Delta). \tag{5}$$

This convolutional dictionary is illustrated in Fig. 1(a).

The advantage of the time discretization is that the data fidelity term is now a quadratic function of the parameters (as compared with the nonlinear embedding of the τ_j's in the original formulation of Eq. (2)). But solving Eq. (4) exactly is NP-hard due to the L_0 term [19], and so we must either resort to approximate algorithms, or introduce further approximations in the

objective. For our purposes here, we choose the latter path. Specifically, we "convexify" the objective function of Eq. (4) by replacing any nonconvex terms with convex approximations. We use a well-known method known as the LASSO ([7]) or alternatively, basis pursuit denoising ([8]), to replace the L_0 penalty term with an L_1 penalty term.[3] The resulting modified optimization problem becomes:

$$\min_{\vec{x}} \ \frac{1}{2\sigma^2}\|y(t) - (F_\Delta \vec{x})(t)\|_2^2 + \lambda\|\vec{x}\|_1 \tag{6}$$

where $\lambda > 0$ is a parameter to be determined. Note that probabilistically, the second term indicates that we have effectively approximated the nonconvex mixture prior Eq. (3) by a Laplacian distribution, $P(x) \propto e^{-\lambda|x|}$.

The global optimum of Eq. (6) can be found using standard quadratic programming methods. A well-known article by Candès et al. [10] provides sufficient conditions under which L_1-minimizing solutions are good approximations for L_0-minimizing solutions (up to a term that is linear in the L_2 norm of the noise). Roughly speaking, the condition limits the correlations between small subsets of dictionary elements. However, this condition is rarely satisfied in the convolutional setting, because the validity of the discrete approximation typically requires Δ to be quite small, which leads to a highly correlated dictionary F_Δ.

In addition to these concerns about the tradeoff between the quality of discrete approximation and the convex relaxation, the discretization of event times also leaves us with no well-defined mapping between the solution of Eq. (4) and the continuous parameters $\{\tau_j, a_j\}$ that optimize the original objective of Eq. (2). One can understand the extent of these problems more concretely by focusing on a single time-shifted waveform $f(t-\tau)$ with $\tau \in (0, \Delta)$, and a two-element dictionary F_Δ containing $f(t)$ and $f(t-\Delta)$. Equation (6) is then a 2D problem, as illustated in Fig. 2(a). The solution is a point at which an elliptical level curve of the first (L_2) term is tangent to a straight-line level curve of the L_1 term. Note that the ellipses are stretched in the direction parallel to the L_1 level lines, because trading off amplitude between the two coefficients does not significantly change the reconstruction error when Δ is small (one can show mathematically that the two right singular vectors of the basis matrix are parallel and orthogonal to the L_1 level curves). Also shown in Fig. 2(a) is the family of solutions that are obtained as one varies λ from

[3]For the purposes of this paper, we assume that the last term in Eq. (4), accounting for the event amplitude probabilities, can also be replaced by an L_1 penalty term.

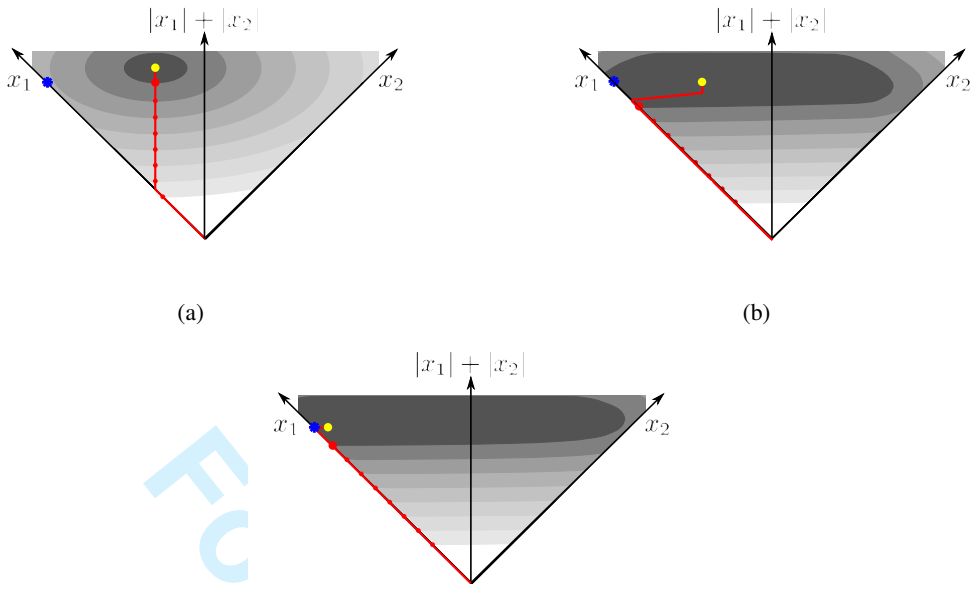


Fig. 2. Geometry of convex objective functions for BP, CBP-T, and CBP-P methods, illustrated in two dimensions. Signal consists of a single waveform, $f(t-0.65\Delta)$, and the dictionary contains two shifted copies of the waveform, $\{f(t), f(t-\Delta)\}$, along with interpolating functions appropriate for each method. Each plot shows the space of coefficients, $\{x_0, x_1\}$, associated with the two shifted waveforms. Grayscale regions indicate reconstruction accuracy (L_2 norm term of the objective function). The sparsity measure (L_1 norm term of the objective function) is the sum of the two coefficient values, corresponding to the vertical position on the plot. Red curve indicates family of solutions traced out as λ is increased fom 0 (yellow dot) to infinity. Red points along this path indicate equal increments in reconstruction accuracy, with enlarged point indicating a common value for comparison across all three methods. Blue dot indicates to the true L_0-minimizing solution. (a) Standard BP solution (Eq. (6)). (b) Continuous basis pursuit with Taylor interpolator (Eq. (12)), using a basis of $\{f_0, f_0', f_\Delta, f_\Delta'\}$. In this case, the iso-accuracy curves are computed by minimizing the L_2 error over all feasible values of the derivative coefficients, and are thus no longer elliptical. (c) CBP with polar interpolation (Eq. (18)).

0 to $+\infty$ (red path). Notice that these solutions are not sparse in the L_0 sense (i.e., they do not intersect either of the two axes) until λ is so large that the signal reconstruction is quite poor. This is the case regardless of how small Δ is, and is due not only to the failure of the L_1 norm to approximate the L_0 pseudonorm, but also to the inability of the discrete model to account for continuous event times. In the following sections, we develop and demonstrate a proposed solution for these problems.

IV. Continuous basis pursuit

The motivation for the discretization of the inference problem was tractability. Equation (4) approximates the original model well in the limit as Δ goes to 0, but it is only tractably solvable in regimes where Δ is large enough so that correlations are limited and L_1-based relaxations can be employed. In this section, we augment the discrete model by adding variables to account for the continuous nature of the event times, and adapt the LASSO to solve this augmented representation. By accounting for the continuous timeshifts, the augmented model not only can approximate the original model better (and for a larger range of Δ), but also admits sparser solutions via an L_1-based recovery method, which we refer to as *continuous basis pursuit* (CBP).

We return to the original continuous problem formulation of Eq. (2), but we assume (without loss of generality) that the waveform is normalized, $\|f(t)\|_2 = 1$, and that the amplitudes, $\{a_j\}$, are all nonnegative. Our noisy observation arises from a linear superposition of time-shifted waveforms, $f(t-\tau)$, which we will abbreviate as $f_\tau(t)$. The set of all time-shifted and amplitude-scaled waveforms forms a 2D nonlinear manifold:

$$\mathcal{M}_{f,T} := \{af(t-\tau) : a \geq 0, \tau \in [0,T]\} \subset L_2([0,T]). \tag{7}$$

The discretized dictionary, F_Δ, provides a linear subspace approximation of this manifold, as illustrated in Fig. 1(a). But the representation of a single element of the manifold (corresponding to a translated scaled copy of the waveform) will typically be approximated by the superposition of several, if not many, elements from the dictionary F_Δ. We can remedy this by augmenting the dictionary to include *interpolation* functions, that allow better approximation of the continuously shifted waveforms. We describe two specific examples of this method, and then provide a general form.

A. Taylor interpolation

If $f(t)$ is differentiable, one can approximate local shifts of $f(t)$ by linearly combining $f(t)$ and its derivative via a first-order Taylor expansion:

$$f_\tau(t) = f(t) - \tau f'(t) + O(\tau^2) \tag{8}$$

This motivates a dictionary consisting of the original shifted waveforms, $\{f_{k\Delta}(t)\}$, and their derivatives, $\{f'_{k\Delta}(t)\}$. We choose a basis function spacing, Δ, as twice the maximal timeshift

such that the first-order Taylor approximation holds within a desired accuracy, δ:

$$\Delta := \max\{\Delta' : \max_{|\tau|<\frac{\Delta'}{2}} \|f_\tau(t) - (f(t) - \tau f'(t))\|_2 \leq \delta\} \tag{9}$$

We can then approximate the manifold of scaled and time-shifted waveforms using *constrained* linear combinations of dictionary elements:

$$\mathcal{M}_{f,T} \approx \left\{ \begin{array}{lcl} & & x \geq 0, \\ x f_{k\Delta}(t) + d f'_{k\Delta}(t) & : & |d| \leq \frac{\Delta}{2} x, \\ & & k = 1, ..., N_\Delta \end{array} \right\} \tag{10}$$

There is a one-to-one correspondence between sums of points on the manifold $\mathcal{M}_{f,T}$ and their respective approximations with this dictionary:

$$\sum_k x_k f_{k\Delta}(t) + d_k f'_{k\Delta}(t) \approx \sum_k x_k f_{(k\Delta - d_k/x_k)}(t). \tag{11}$$

This holds as long as $|d_k/x_k| \neq \frac{\Delta}{2}$ (which corresponds to the situation where the the waveform is displaced exactly halfway in between two lattice points, and can thus be equally well represented by the basis function and associated derivative on either side). This is illustrated in Fig. 1(b).

The inference problem is now solved by optimizing a *constrained* convex objective function:

$$\min_{\vec{x},\vec{d}} \ \frac{1}{2\sigma^2} \left\| y(t) - (F_\Delta \vec{x})(t) - (F'_\Delta \vec{d})(t) \right\|_2^2 + \lambda \|\vec{x}\|_1$$

$$\text{s.t.} \quad \left\{ \begin{array}{l} x_k \geq 0, \\ |d_k| \leq \frac{\Delta}{2} x_k \end{array} \right\} \text{ for k} = 1, ..., \text{N}_\Delta \tag{12}$$

where the dictionary F_Δ is defined as in Eq. (5), and F'_Δ is a dictionary of time-shifted waveform derivatives $\{f'_{k\Delta}(t)\}$. Equation (11) provides an explicit mapping from appropriately constrained coefficient configurations to event amplitudes and timeshifts. Figure 2(b) illustrates this objective function for the same single-waveform example described previously. The shaded regions are the level sets of the L_2 term of Eq. (12) visualized in the (x_1, x_2)-plane by minimizing over the derivative coefficients (d_1, d_2). Note that unlike the corresponding BP level sets shown in Fig. 2(a), these are no longer elliptical, and that they allow sparse solutions (i.e., points on the $x_1 = 0$ axis) with low reconstruction error. As a result, for λ sufficiently large, the solution of Eq. (12) is not only sparse in the L_0 sense, but also provides a good reconstruction of the signal.

B. Polar interpolation

Although the Taylor series provides the most intuitive and well-known method of approximating time-shifts, we have developed an alternative interpolator that is significantly more accurate. The solution is motivated by the observation that the manifold of time-shifted waveforms, $f_\tau(t)$, must lie on the surface of a unit hypersphere (because the waveform L_2-norm is preserved under time shifting), and furthermore, must have a constant curvature (by symmetry). This leads to the notion that it might be well-approximated by an arc of a circle. As such, we approximate a segment of the manifold, $\{f_\tau : |\tau| \leq \frac{\Delta}{2}\}$, by the unique circular arc that contains the three points $\{f_{-\Delta/2}, f_0, f_{\Delta/2}\}$, as illustrated in Fig. 3(a). The resulting interpolator is an example of a trigonometric spline [20], in which the three time-shifted functions are linearly combined using trigonometric coefficients to approximate intermediate translates of $f(t)$:

$$f_\tau(t) \approx c(t) + r\cos(\frac{2\tau}{\Delta}\theta)u(t) + r\sin(\frac{2\tau}{\Delta}\theta)v(t) \tag{13}$$

where the functions $\{c(t), u(t), v(t)\}$ are computed from linear combinations of $\{f_{-\Delta/2}, f_0, f_{\Delta/2}\}$:

$$\begin{pmatrix} f_{-\frac{\Delta}{2}}(t) \\ f_0(t) \\ f_{\frac{\Delta}{2}}(t) \end{pmatrix} = \begin{pmatrix} 1 & r\cos(\theta) & -r\sin(\theta) \\ 1 & r & 0 \\ 1 & r\cos(\theta) & r\sin(\theta) \end{pmatrix} \begin{pmatrix} c(t) \\ u(t) \\ v(t) \end{pmatrix} \tag{14}$$

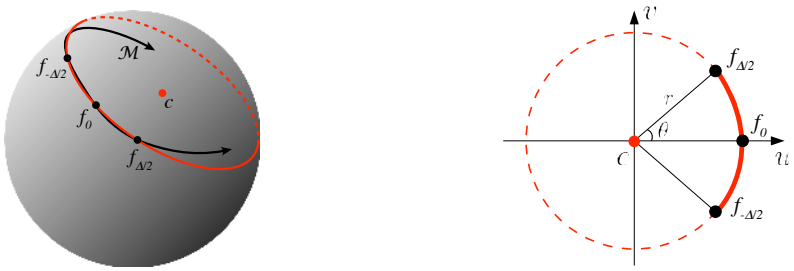
The constant r is the radius of the circular arc, and θ is half the angle subtended by the arc, both of which depend on $f(t)$ and can be computed in closed form. These relationships are illustrated in Fig. 3(b). The approximation can be easily expressed in the frequency domain, by taking the Fourier transform of both sides of Eq. (13) and using Eq. (14):

$$\begin{aligned} e^{-i\omega\tau} \approx\ & (1 - 2a(\tau)) \ + \\ & e^{i\omega\frac{\Delta}{2}}(a(\tau) - b(\tau)) \ + \\ & e^{-i\omega\frac{\Delta}{2}}(a(\tau) + b(\tau)) \end{aligned} \tag{15}$$

where

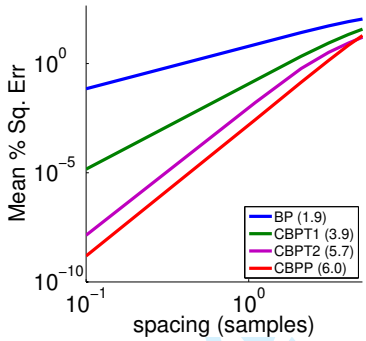
$$a(\tau) = \frac{\cos(\frac{2\tau\theta}{\Delta}) - 1}{2(\cos(\theta) - 1)} \quad \text{and} \quad \mathrm{b}(\tau) = \frac{\sin(\frac{2\tau\theta}{\Delta})}{2\sin(\theta)}.$$

Figure. 4(a) compares nearest neighbor (as is implicitly used in BP), first-order Taylor, and polar interpolation in terms of their accuracy in approximating timeshifts of a Gaussian derivative



(a) (b)

Fig. 3. Illustration of the polar interpolator. (a) The manifold of time shifts of $f(t)$ (black line) lies on the surface of a hypersphere. We approximate a segment of this manifold, for time shifts $\tau \in [-\frac{\Delta}{2}, \frac{\Delta}{2}]$, with a portion of a circle (red), with center defined by $c(t)$. (b) Parameterization of the circular arc approximation.



(a)

Fig. 4. Comparison of the nearest neighbor, first-order Taylor, and polar interpolators (as used in BP, CBP-T, and CBP-P, respectively) for a waveform $f(t) \propto te^{-\alpha t^2}$. Sinc and 2nd-order Taylor interpolation are also shown. The estimated slopes (asymptotic rates of convergence) are shown in the legend.

waveform, $f(t) \propto te^{-\alpha t^2}$. For reference, the second-order Taylor interpolator is also included. The polar interpolator is seen to be significantly more accurate than nearest-neighbor and 1st-order Taylor, and even surpasses 2nd-order Taylor by an order of magnitude (although they have the same asymptotic rate of convergence). This allows one to choose a much larger Δ for a given desired accuracy.

We now construct a dictionary of time-shifted copies of the functions used to represent the polar interpolation, $\{c_{k\Delta}, u_{k\Delta}, v_{k\Delta}\}$, and form a convex set from these to approximate the

manifold:

$$\mathcal{M}_{f,T} \approx \left\{ \begin{array}{ll} xc_{k\Delta}(t) & x \geq 0, \\ & y^2 + z^2 \leq x^2 r^2, \\ + \; yu_{k\Delta}(t) \quad : & \\ & xr\cos(\theta) \leq y \leq xr, \\ + \; zv_{k\Delta}(t) & \\ & k = 1, ...N_\Delta \end{array} \right\} \tag{16}$$

The constraints on the coefficients (x, y, z) ensure that they represent a scaled translate of $f(t)$, except for the second, which is a convex relaxation of the true constraint, $y^2 + z^2 = x^2 r^2$ (see below). As with the Taylor approximation, we have a one-to-one correspondence between event amplitudes/timeshifts and the constrained coefficients:

$$\sum_k x_k c_{k\Delta}(t) + y_k u_{k\Delta}(t) + z_k v_{k\Delta}(t) \tag{17}$$
$$\approx \sum_k x_k f_{(k\Delta - \frac{\Delta}{2\theta}\tan^{-1}(z_k/y_k))}(t)$$

as long as $z_k/y_k \neq \tan(\theta)$ for all k. The inference problem again boils down to minimizing a constrained convex objective function:

$$\min_{\vec{x},\vec{y},\vec{z}} \; \frac{1}{2\sigma^2} \left\| y(t) - (C_\Delta \vec{x})(t) - (U_\Delta \vec{y})(t) - (V_\Delta \vec{z})(t) \right\|_2^2 + \lambda \|\vec{x}\|_1$$

$$\text{s.t.} \quad \left\{ \begin{array}{l} x_k \geq 0, \\ \sqrt{y_k^2 + z_k^2} \leq x_k r, \\ x_k r\cos(\theta) \leq y_k \leq x_k r, \end{array} \right\} \text{ for k} = 1, ...\text{N}_\Delta \tag{18}$$

where $C_\Delta, U_\Delta, V_\Delta$ are dictionaries containing Δ-shifted copies of $c(t), u(t), v(t)$, respectively. Equation (18) is an example of a "second-order cone program" for which efficient solvers exist ([21]). After the optimum values for $\{\vec{x}, \vec{y}, \vec{z}\}$ are obtained, timeshifts and amplitudes can be inferred by first projecting the solution back to the original constraint set:

$$(x_k, y_k, z_k) \leftarrow (x_k, \frac{y_k x_k r}{\sqrt{y_k^2 + z_k^2}}, \frac{z_k x_k r}{\sqrt{y_k^2 + z_k^2}}) \tag{19}$$

and then using Eq. (17) to solve for the event times.

Figure 2(c) illustrates the optimization of Eq. (18) for the simple example described in the previous section. Notice that the solution corresponding to $\lambda = 0$ (yellow dot) is significantly sparser relative to both the CBP-T and BP solutions, and that the solution becomes L_0 sparse if λ is increased by just a small amount, giving up very little reconstruction accuracy.

C. General interpolation

We can generalize the CBP approach to use any linear interpolation scheme. Suppose we have a set of basis functions $\{\phi_n(t)\}_1^m$ in $L_2([0,T])$ (for simplicity, assume they are orthonormal) and a corresponding interpolation map $\vec{D}(\cdot)$ such that local shifts can be approximated as:

$$f_\tau(t) \approx \sum_{n=1}^{m} D_n(\tau)\phi_n(t), \quad |\tau| \leq \frac{\Delta}{2}. \tag{20}$$

Let S be the set of all nonnegative scalings of the image of $[-\frac{\Delta}{2}, \frac{\Delta}{2}]$ under the interpolator:

$$S = \{a\vec{D}(\tau) : a \geq 0, \tau \leq \frac{\Delta}{2}\}.$$

If the interpolator $\vec{D}(\tau)$ is invertible, we have a one-to-one correspondence as before:

$$\begin{aligned} &\sum_{k=1}^{N_\Delta}\sum_{n=1}^{m} x_{kn}\phi_n(t-k\Delta) \\ &\qquad \approx \sum_{k=1}^{N_\Delta} \|\vec{x}_k\|_2 f_{(k\Delta - \vec{D}^{(-1)}(\vec{x}_k/\|\vec{x}_k\|_2))}(t) \end{aligned} \tag{21}$$

where each group $\vec{x}_k := [x_{k1}, ..., x_{km}]$ is in S and $|\vec{D}^{(-1)}(\vec{x}_k/\|\vec{x}_k\|_2)| \neq \frac{\Delta}{2}$ for all k. Note that in this general form, the L_2 norm of each group $\vec{x}_k$ governs the amplitude of the corresponding time-shifted waveform.[4] As we saw in the previous examples, S may or may not be convex, so we relax to its convex hull, denoted by $\overline{S}$, keeping in mind that we must project our solution back onto S at the end, using an operator $P_S(\cdot)$.

Finally, we can write obtain the representation using this interpolation by solving:

$$\begin{aligned} \min_{\vec{x}} \ & \frac{1}{2\sigma^2}\|y(t) - (\Phi_\Delta \vec{x})(t)\|_2^2 + \lambda \sum_{k=1}^{N_\Delta} \|\vec{x}_k\|_2 \\ \text{s.t.} \quad & \vec{x}_k \in \overline{S} \quad \text{for} \quad k = 1, ..., N_\Delta \end{aligned} \tag{22}$$

where the linear operator Φ_Δ is defined as:

$$(\Phi_\Delta \vec{x})(t) := \sum_{k=1}^{N_\Delta}\sum_{n=1}^{m} x_{kn}\phi_n(t-k\Delta)$$

[4]Our specific examples used the amplitude of a single coefficient as opposed to the group L_2 norm. However, the constraints in these examples make the two formulations equivalent up to $O(\Delta)$. For the Taylor interpolator, $x_k^2 \approx x_k^2 + d_k^2$. For the polar interpolator, $c_k^2 + u_k^2 + v_k^2 \approx 2c_k^2$.

Equation (22) can be solved efficiently using standard convex optimization methods (e.g., interior point methods [21]). It is similar to the objective functions used to recover so-called "block-sparse" signals (e.g., [16], [22]), but includes auxilliary constraints on the coefficients to ensure that only signals close to $span(\mathcal{M}_{f,T})$ are represented. Table I summarizes the Taylor and polar interpolation examples within the general framework, along with the case of nearest-neighbor interpolation (which corresponds to standard BP).

Property	BP (nearest-neighbor)	CBP - Taylor interp	CBP polar interp
basis: $\{\phi_n(t)\}_{n=1}^m$	$[f(t)]$	$[f(t), f'(t)]$	$[c(t), u(t), v(t)]$
interpolator: $\vec{D}(\tau)$	1	$[1, \tau]^T$	$[1, r\cos(\theta\frac{2\tau}{\Delta}), r\sin(\theta\frac{2\tau}{\Delta})]^T$
constrained coefficient set: S	$\{x_1 \geq 0\}$	$\{x_1 \geq 0, \lvert x_2\rvert \leq x_1\frac{\Delta}{2}\}$	$\{x_1 \geq 0, x_2^2 + x_3^2 = r^2x_1^2, rx_1\cos(\theta) \leq x_2 \leq rx_1\}$
convex relaxation: $\overline{S}$	S	S	$\{x_1 \geq 0, x_2^2 + x_3^2 \leq r^2x_1^2, rx_1\cos(\theta) \leq x_2 \leq rx_1\}$
projection operator: $P_S(\vec{x})$	$\vec{x}$	$\vec{x}$	$[x_1, rx_1\frac{x_2}{\sqrt{x_2^2+x_3^2}}, rx_1\frac{x_3}{\sqrt{x_2^2+x_3^2}}]^T$

TABLE I

The quality of the solution relies on the accuracy of the interpolator, the convex approximation $\overline{S} \approx S$, and the ability of the block-L_1 based penalty term in Eq. (22) to achieve L_0-sparse solutions that reconstruct the signal accurately. The first two of these are relatively straightforward, since they depend solely on the properties of the interpolator (see Fig. 4(a)). The last is difficult to predict, even for the simple examples illustrated in Figure 2. The level sets of the L_2 term can have a complicated form when taking the constraints into account, and it is not clear a priori whether this will facilitate or hinder the L_1 term in achieving sparse solutions. Nevertheless, our empirical results clearly indicate that solving Eq. (22) with Taylor and polar interpolators yields significantly sparser solutions than those achieved with standard BP.

V. Empirical results

We evaluate our method on data simulated according to the generative model of Eq. (1). We chose the probability density on event amplitudes, $P_A(\cdot)$, to be uniform on the interval $[a, b]$ with $0 < a < b$. We used a single template waveform $f(t) \propto te^{-\alpha t^2}$ (normalized, so that $\|f\|_2 = 1$), for which the interpolator performances are plotted in Fig. 4(a). We compared solutions of Eqs. (6), (12), and (18). In all recovery methods, amplitudes were constrained to be

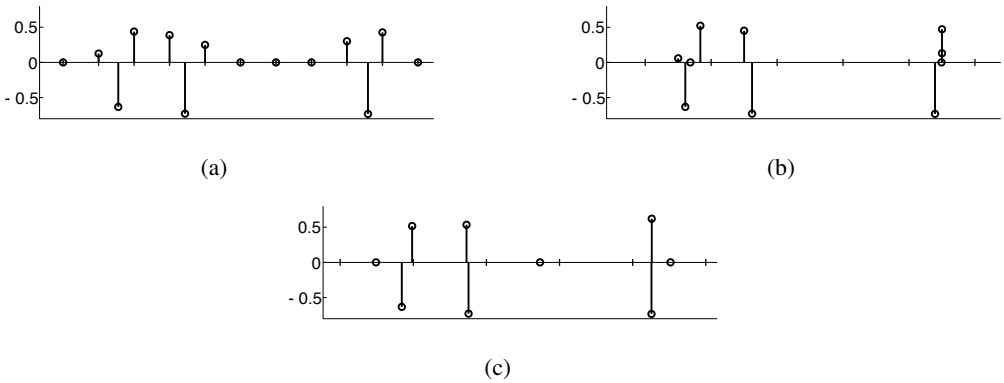


Fig. 5. Example of sparse signal recovery for (a) BP (Eq. (6)), (b) CBP-T (Eq. (12)), and (c) CBP-P (Eq. (18)). For each method, the values of Δ and λ were chosen to minimize the average sum of squares of the two types of error. Upward stems indicate the estimated magnitudes placed at locations determined by the interpolation coefficients via Eq. (21). Ticks denote the location of the basis functions corresponding to each upward-pointing stem. Downward stems indicate the locations and magnitudes of the true signal. SNR was 12 (identical signal and noise for all three examples).

nonnegative (this is already assumed for the CBP methods, and amounts to an additional linear inequality constraint for BP). Each method has two free parameters: Δ controls the spacing of the basis, and λ controls the tradeoff between reconstruction error and sparsity. We varied these parameters systematically and measured performance in terms of two quantities: (1) signal reconstruction error (which decreases as λ increases or Δ decreases), and (2) sparsity of the estimated event amplitudes ($\{\|\vec{x}_j\|_2\}$), which increases as λ increases. The former is simply the first term in the objective function (for all three methods). For the latter, to ensure numerical stability, we used the L_p norm with $p = 0.1$ (results were stable with respect to the choice of p, as long as $p < 1$ and p was not below the numerical precision of the optimizations. Computations were performed numerically, by sampling the functions $f(t)$ and $y(t)$ at a fine constant spacing. We used the convex solver package CVX [23] to to obtain numerical solutions.

A small temporal window of the events recovered by the three methods is provided in Figure 5. The three plots show the estimated event times and amplitudes for BP, CBP-T, and CBP-P (upward stems) compared to the true event times/amplitudes (downward stems). The figure demonstrates that CBP, equipped with either Taylor or polar interpolators, is able to recover the event train more accurately, and with a larger spacing between basis functions (indicated by the tick marks on the x-axis). As predicted by the reasoning laid out in Figure 2(a), basis pursuit tends to split events across two or more adjacent low-amplitude coefficients, thus producing less

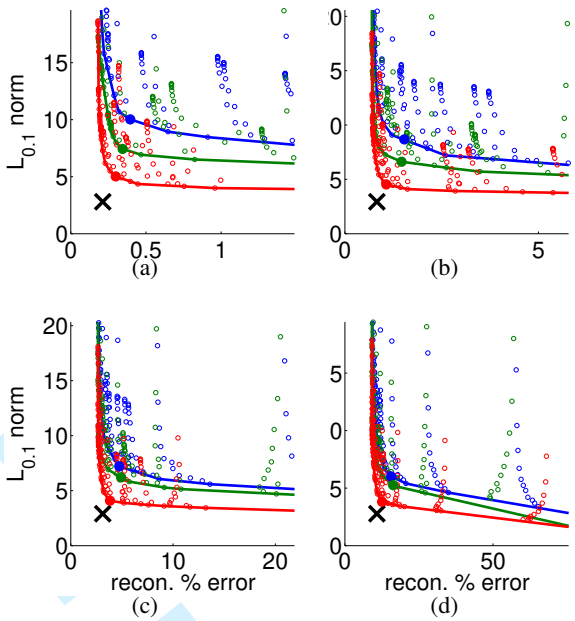
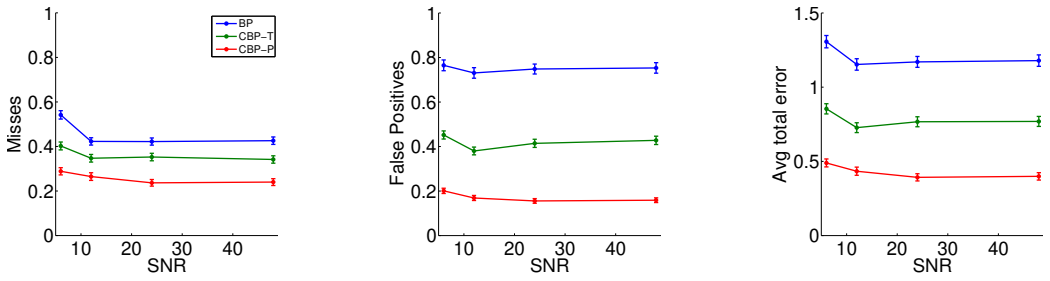


Fig. 6. Error plots for four noise levels: (a) SNR = 48dB (b) SNR = 24dB (c) SNR = 12dB (d) and SNR = 6dB, where SNR is defined as $\|f\|_\infty/\sigma$). Each graph shows the tradeoff between the average reconstruction error (vertical axis) and the sparsity (horizontal axis, measured as average $L_{0.1}$ norm of estimated amplitudes). Each point represents the error values for one of the methods, applied with a particular setting of (Δ, λ), averaged over 500 trials. Colors indicate the method used (BP-blue,CBP-T-green CBP-P-red). Bold lines denote the convex hulls of all points for each method. The large dots indicate the "best" solution as measured by Euclidean distance from the correct solution (indicated by black X's).

sparse solutions and making it hard to infer the number of events and their respective amplitudes and times. Sparsity can be improved by increasing λ, but at the expense of a substantial increase in approximation error.

Figure 6 illustrates the tradeoff between sparsity and approximation error for each of the methods. Each panel corresponds to a different noise level. The individual points, color-coded for each method, are obtained by running the associated method 500 times for a given (Δ, λ) combination, and averaging the errors over these trials. The solid curves are the (numerically computed) convex hulls of all points obtained for each method, and clearly indicate the tradeoff between the two types of error. We can see that the performance of BP is strictly dominated by that of CBP-T: For every BP solution, there is a CBP-T solution that has lower values for both error types. Similarly, CBP-T is strictly dominated by CBP-P, which can be seen to come close to the error values of the ground truth answer (which is indicated by a black X).

We performed a signal detection analysis of the performance of these methods, classifying



(a) (b) (c)

Fig. 7. Signal detection analysis of solutions (see text). (a) Average miss rate (as a fraction of the mean number of events per trial) computed over 500 trials for each method, and for each SNR (defined as $\|f\|_\infty/\sigma$). (b) Average false positive rate. (c) Total error (sum of misses and false positives)

identification errors as misses and false positives. Given a true and estimated event train, we say that an event is matched across the two if (1) the estimated event's amplitude is within some threshold $\alpha > 0$ of the true amplitude and (2) the estimated event time is within some threshold $\nu > 0$ of the true event time, and (3) no other estimated event has been matched to the true event. We evaluated the three methods using these criteria, using a value of $\alpha = \frac{1}{\sqrt{12}}$ (one standard deviation of the amplitude distribution Unif$[0.5, 1.5]$)) and $\nu = 3$ samples. We found that results were relatively stable with respect to these threshold choices. For each method and noise level we chose the (λ, Δ) combination yielding a solution closest to ground truth (corresponding to the large dots in Figure 6). Figure 7 shows the errors as a function of the noise level. We see that performance of all methods is surprisingly stable across SNR levels. We also see that BP performance is dominated at all noise levels by CBP-T, which has fewer misses as well as fewer false positives, and CBP-T is similarly dominated by CBP-P.

Finally, we examined the distribution of the amplitudes estimated by each algorithm, and compare with the distribution of the source, as given by Eq. (3). Figure 8 shows the amplitude histogram for each method. We see that CBP-P produces amplitude distributions that are far better-matched to the correct distribution of amplitudes.

A. Multiple features

All of the methods we've described can be easily extended to the case of multiple templates, by taking as a dictionary the union of dictionaries associated with each individual template. We

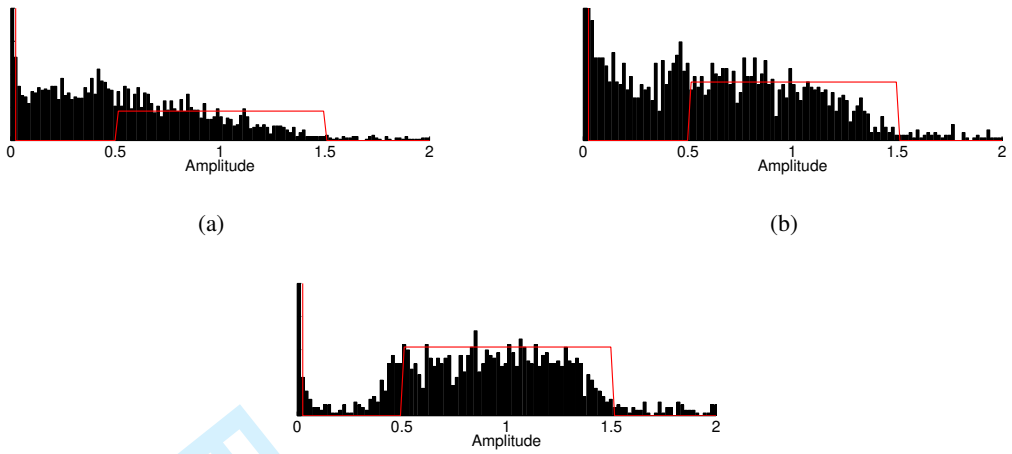
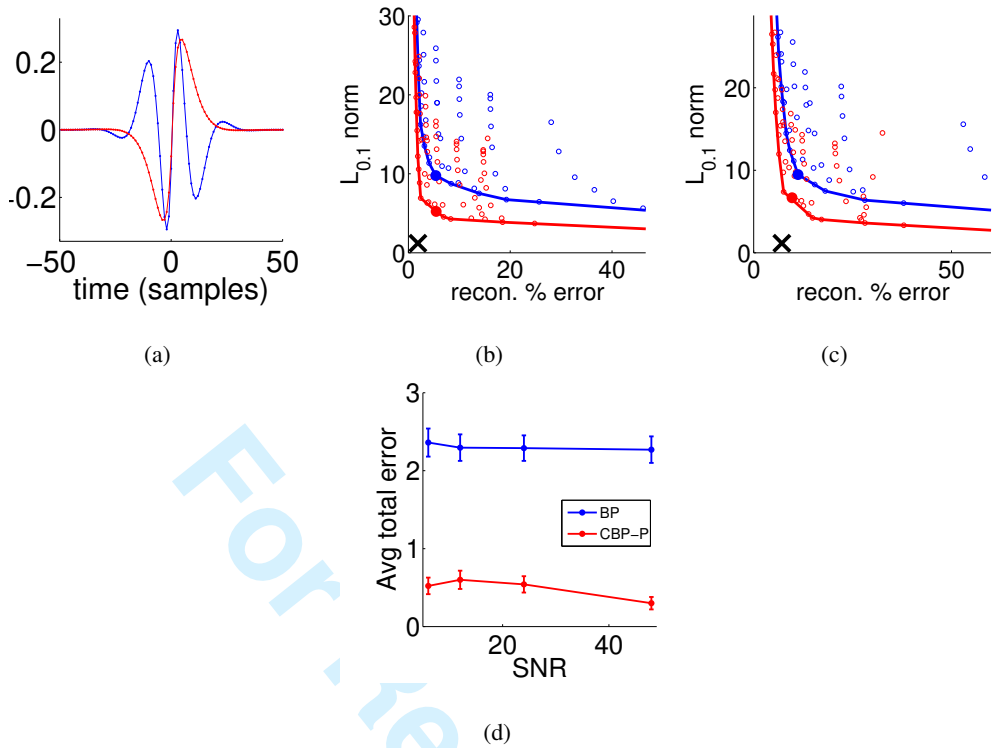


Fig. 8. Histograms of the estimated amplitudes for (a) BP, (b) CBP-T, and (c) CBP-P. All methods were constrained to estimate only nonnegative amplitudes, but no upper bound was imposed. The true distribution of amplitudes is given by Eq. (3), and is indicated in red.

performed a final set of experiments for the case of two features (waveforms shown in Fig. 9(a)) that are "gammatone" filters, as commonly used in audio processing. Data were generated by constructing two correlated Poisson processes with the same marginal rate λ and a correlation of $\rho = 0.5$. These were generated by independently creating 2 Poisson process with rate $\lambda(1-\rho)$ and then superimposing a randomly jittered "common" Poisson process with rate $\lambda\rho$. As before, event amplitudes were drawn independently from a uniform distribution on $[a, b]$ with $a > 0$.

We examined and compared performance of BP and CBP-P. Both methods used dictionaries formed from the union of dictionaries for each template, but we forced the two individual dictionaries to use a common spacing, Δ, for both waveforms. In general, the spacing could be chosen differently for each waveform, providing more flexibility, at the expense of additional parameters that must be selected or optimized. Figures 9(b) and 9(c) show the error tradeoff for different settings of (Δ, λ) at SNR levels of 24 and 12, respectively (the results were qualitatively unchanged for SNR values of 48 and 6). Figure 9(d) shows the total number of event identification errors (misses plus false positives) for each method as a function of SNR at each methods optimal (Δ, λ) setting.



(a) (b) (c)

(d)

Fig. 9. (a) Two gammatone features of the form $f_i(t) = at^{n-1}e^{-2\pi bt}\cos(2\pi\omega_i t)$ for $i = 1, 2$. (b) and (c) show the sparsity and reconstruction errors for BP (blue) and CBP-P (red), as in Figure 6, with SNRs of 24 and 12, respectively. (d) plots the total number of misses and false positives (with same thresholds as in Figure 7(c)) for each method.

VI. Discussion

We have introduced a novel methodology for sparse signal decomposition in terms of continously shifted features. The method can be seen as a continuous form of the well-known basis pursuit method, and we thus have dubbed it *Continuous Basis Pursuit*. The method overcomes the limitation of basis pursuit in the convolutional setting casued by the tradeoff between discretization error and the effectiveness of the L_1 relaxation for obtaining sparse solutions. In particular, our method employs an alternative discrete basis (not necessarily the features themselves) which can explicitly account for the continuous timeshifts present in the signal. We derived a general convex objective function that can be used with any such basis. The coefficients are constrained so as to represent only transformed versions of the templates, and the objective function uses an L_1 norm to penalize the amplitudes of these transformed versions . We showed empirically that using simple first-order (Taylor) and second-order (polar) interpolation schemes yields superior solutions in the sense that (1) they are sparser, with improved reconstruction accuracy, (2) they

produce substantially better identification of events (fewer misses and false positives), and (3) the amplitude statistics are a better match to those of the true generative model. We showed that these results are stable across a wide range of noise levels. We conclude that an interpolating basis, coupled with appropriate constraints on coefficients, provides a powerful and tractable tool for modeling and decomposing translation-invariant signals.

We believe our method can be extended for use with other types of signal, as well as to tranformations other than translation. For example, for one dimensional signals such as audio, one might also include dilation or frequency-modulation of the templates. For two-dimensional signals, such as photographic images, one could include rotation. For each of these, the primary hurdles are to specify (1) the form of the linear interpolation (for joint variables, this might be done separably, or using a multi-dimensional interpolator), (2) the constraints on coefficients (and a convex relaxation of these constraints), and (3) a means of inverting the interpolator so as to obtain transformation parameters from recovered coefficients. Another natural extension is to use CBP in the context of learning optimal templates for decomposing an ensemble of signals, as has been previously done with BP (e.g., [13], [14], [12], [15], [24]).

ACKNOWLEDGMENT

The authors would like to thank Sinan Güntürk for helpful discussions in the early stages of this work.

REFERENCES

[1] J. H. Friedman and J. W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Trans Computers*, C-23(9):881–890, 1974.

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

[3] J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory*, 50(10):2231–2242, 2004.

[4] T. Blumensath and M. E. Davies. Iterative thresholding for sparse approximations. *J. Fourier Analysis and Applications*, 2004.

[5] M. Elad. Why simple shrinkage is still relevant for redundant representations. *IEEE Trans Info Theory*, 52:5559–5569, 2006.

[6] J. Bioucas-Dias and M. Figueiredo. A new TwIST: Two-step iterative shrinkage/thresholding algorithms for image restoration. *IEEE Trans Image Processing*, 16(12):2992–3004, 2007.

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

[8] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

[9] Emmanuel C and Terence Tao. The dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351, 2005.

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

[11] Martin J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). *IEEE Trans. Inf. Theor.*, 55:2183–2202, May 2009.

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

[13] Phil Sallee and Bruno A. Olshausen. Learning sparse multiscale image representations. In *NIPS*, pages 1327–1334, 2002.

[14] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, Jun 1996.

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

[16] 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.

[17] J. Mendel. *Optimal Seismic Deconvolution: An Estimation Based Approach*. Academic Pr, 1983.

[18] M. S. Lewicki. A review of methods for spike sorting: the detection and classification of neural action potentials. *Network*, 9(4):R53–R78, Nov 1998.

[19] G. Davis, S. Mallat, and M. Avellaneda. Adaptive greedy approximations. *Constructive Approximation*, 13:57–98, 1997. 10.1007/BF02678430.

[20] I.J. Schoenberg. On trigonometric spline interpolation. *Journal of Math Mech.*, 13:795–825, 1964.

[21] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[22] 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.

[23] M. Grant and S. Boyd. Cvx: Matlab software for disciplined convex programming, version 1.21. http://cvxr.com/cvx, October 2010.

[24] P. Berkes, R. E. Turner, and M. Sahani. A structured model of video reproduces primary visual cortical organisation. *PLoS Computational Biology*, 5(9), 2009.

PLACE PHOTO HERE

Chaitanya Ekanadham Biography text here.

PLACE PHOTO HERE

Daniel Tranchina Biography text here.

PLACE PHOTO HERE

Eero P. Simoncelli received the BS degree (summa cum laude) in physics from Harvard University in 1984 and the MS and PhD degrees in electrical engineering from Massachusetts Institute of Technology in 1988 and 1993, respectively. He studied applied mathematics at Cambridge University for a year and a half. He was an assistant professor in the Computer and Information Science Department at the University of Pennsylvania from 1993 to 1996. In September 1996, he moved to New York University, where he is currently an associate professor in neural science and mathematics. In August 2000, he became an associate investigator at the Howard Hughes Medical Institute under their new program in computational biology. His research interests span a wide range of topics in the representation and analysis of visual images, in both machine and biological systems. He is a fellow of the IEEE.

Recovery of sparse translation-invariant signals with continuous basis pursuit

Chaitanya Ekanadham, Daniel Tranchina, and Eero Simoncelli, *Fellow, IEEE*

***Abstract*—We consider the problem of decomposing a signal into a linear combination of features, each a continuously translated version of one of a small set of elementary features. Although these constituents are drawn from a *continuous family*, most current signal decomposition methods rely on a finite dictionary of discrete examples selected this family (e.g., a set of shifted copies of a set of basic waveforms), and apply sparse optimization methods to select and solve for the relevant coefficients. Here, we generate a dictionary that includes auxilliary interpolation functions that approximate local continuous translates of features via constrained adjustment of their coefficients. We formulate a constrained convex optimization problem, in which the full set of dictionary coefficients represent a linear approximation of the signal, the auxiliary coefficients are constrained so as to only represent translated features, and sparsity is imposed on the non-auxiliary coefficients using an L1 penalty. The well-known *basis pursuit denoising* (BP) method may be seen as a special case, in which the auxiliary interpolation functions are omitted, and we thus refer to our methodology as *continuous basis pursuit* (CBP). We develop two implementations of CBP for a one-dimensional translation-invariant source, one using a first-order Taylor approximation, and another using a form of trigonometric spline. We examine the tradeoff between sparsity and signal reconstruction accuracy in these methods, demonstrating empirically that trigonometric CBP significantly outperforms Taylor CBP, which in turn offers significant gains over ordinary BP. In addition, the CBP bases can generally achieve equally good or better approximations with much coarser sampling than BP, leading to a reduction in dictionary dimensionality.**

I. Introduction

THE decomposition of a signal into a sparse linear combination of features is an important and well-studied problem, and plays a central role in many applications. A surge of recent effort focuses on representing a signal as a noisy superposition of the smallest possible subset of functions drawn from a large finite dictionary. The standard formulation tries to minimize the L_0 pseudonorm (number of nonzero elements) of the vector of weights corresponding to the dictionary elements.

The finite dictionary of basis functions $\{\phi_k(t)\}$ may be fixed in advance, or optimized (so as to best represent an ensemble of signals). In general, this objective can only be minimized via exhaustive search of all 2^d subsets of the dictionary, making it infeasible in practice. However, two broad classes of approximate solutions have been widely studied in the literature. The first consists of greedy methods, dating back to variable selection methods in the 1970s ([1]). These methods are exemplified by the well-known "matching pursuit" algorithm of Mallat and Zhang [2], and include a variety of more recent "iterative thresholding" methods [3], [4], [5], [6]. The general idea is to solve sequentially for the nonzero elements of $\vec{x}$, at each step choosing the element(s) that best explain the current residual. A second category of solutions arises from convex relaxations of the L_0 objective, and include the LASSO [7], the basis pursuit denoising (BP) algorithm [8], and the Dantzig selector [9], each of which employ the convex L_1 norm. Results by Tibshirani [7] and Chen et. al [8] show that substituting an L_1 penalty makes the problem solvable using quadratic programming and yields solutions with a high degree of sparsity. Recent publications [10], [11] provide conditions on the dictionary that guarantee this approximation to be near-optimal.

Most objective functions that have been utilized for sparse decomposition are constructed around the premise of linear superposition and additive noise, and make no assumptions about the structure of the dictionary. However, many real signals are generated by processes that obey natural invariances (e.g., translation-invariance, dilation-invariance, rotation-invariance). In this setting, the goal is to identify feature instances in the signal along with their associated amplitudes and transformation parameters. With a translation-invariant signal in time, for example, one aims to identify the amplitudes and timeshifts of the features. In the majority of published examples, the problem is solved by constructing a finite dictionary that reflects the invariant structure: one discretely samples the transformation parameters and applies these to a finite set of elementary features. For example, dictionaries for sound processing, whether learned or hand-constructed, are commonly "convolutional", containing time-delayed copies of template waveforms (e.g., [12]).

Dictionaries for image representation typically contain features that are translated, and in some cases, dilated and rotated (e.g., [13]).[1] This discrete sampling approach replaces the full nonlinear problem with a more tractable linear inverse problem. However, the ability of the discrete dictionary to accurately represent signals depends critically on the spacing at which the dictionary was sampled. In general, a very

C.E. is with the Courant Institute of Mathematical Sciences (CIMS), New York University, NY 10003 (e-mail: chaitu@math.nyu.edu).

D.T. is with CIMS, the Center for Neural Science (CNS), and the Department of Biology, NYU, NY 10003 (e-mail: tranchin@courant.nyu.edu).

E.P.S. is with the Howard Hughes Medical Institute (HHMI), CNS, and CIMS, NYU, NY 10012 (e-mail: eero.simoncelli@nyu.edu).

This work was partially funded by NYU through a McCracken Fellowship to C.E., and by an HHMI Investigatorship to E.P.S.

[1]Many examples of sparse decomposition on images have been applied to nonoverlapping square blocks of pixels (e.g., [14], [15], [16]), but the effective dictionary for representing the entire image is the union of dictionary elements for each block, and thus consists of translated copies of the block dictionary elements.

fine sampling is required, resulting in a very large and ill-conditioned dictionary. This ill-conditioning, in turn, is unfavorable for the relaxation approximations mentioned above. Furthermore, given this representation, it is still unclear how to estimate the true amplitudes and transformation parameters associated with the recovered features.

Here, we propose an alternative linear approximation to the full nonlinear problem. We focus on the problem of translation-invariant one-dimensional signals (although the methods generalize to other transformations, and higher dimensions). We construct a group of functions that can span local translations of the feature templates via continuous variation of their coefficients. As a concrete example, consider the original templates and their derivatives, which can approximate local translations through a first-order Taylor approximation. The resulting dictionary can generally approximate the true set of scaled and translated templates more accurately than a dictionary of equal size containing only translated copies of the feature itself (i.e. the special case in which the interpolating group is just the template). A signal of interest is then represented in this dictionary by "block-sparse" coefficients, where each non-zero coefficient block represents an amplitude-scaled and translated template. We formulate an objective function in which the coefficients are constraind so as to only represent scaled/transformed templates, and use an L_1 penalty to impose sparsity on the blocks. The advantage of this approach over ordinary BP is three-fold: (1) better approximation of translation-invariant signals, (2) a smaller basis, which leads to sparser solutions via convex optimization, and (3) an explicit mapping from this representation to amplitudes and transformation parameters.

II. Problem formulation

We begin by formulating a simple generative model for translation-invariant signals, as well as the maximum a posteriori (MAP) estimation framework for inferring the most likely parameters given the observed signal. Assume we observe a signal that is a noisy superposition of scaled time-shifted copies of a single known elementary waveform $f(t)$ on a finite interval $[0, T]$:

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

where $\eta(t)$ is a Gaussian white noise process with power σ^2, the *event times* $\{\tau_j\}$ are drawn from a Poisson process with rate μ, and the *event amplitudes* $\{a_j\}$ are drawn independently from a density $P_A(a)$. The inverse (inference) problem is then to recover the most likely values of parameters $\{\tau_j, a_j\}$ given $y(t)$. This amounts to maximizing the posterior distribution $P(\{\tau_j, a_j\}|y(t))$, which reduces, on taking the negative log, to solving:

$$\min_{N,\{\tau_j,a_j\}} \frac{1}{2\sigma^2}\|y(t) - \sum_{j=1}^{N} a_j f(t-\tau_j)\|_2^2 + N\log(\mu) - \sum_{j=1}^{N} \log P_A(a_j) \tag{2}$$

This *sparse deconvolution* formulation has been used to describe many real-world problems including seismogram analysis [17], neural spike sorting [18], acoustic signal analysis [12], and image processing [13]. Unfortunately, solving Eq. (2) directly is intractable, due to the discrete nature of N and the nonlinearity embedding of the τ_j's within the argument of the waveform $f(\cdot)$. It is thus desirable to find alternative formulations that (i) approximate the signal posterior distribution well, (ii) have parameters that can be tractably estimated, and (iii) have an intuitive mapping back to the original representation.

III. Conventional solution: discretization and BP

A standard simplification of the problem is to discretize the event times at a spacing that is fine enough that the Poisson process is well-approximated by a Bernoulli process. The interval $[0, T]$ is divided into $N_\Delta = \lceil T/\Delta \rceil$ time bins of size Δ, where the probability of an event in each bin is $\mu\Delta$, for Δ sufficiently small.[2] This discrete process is represented by a vector $\vec{x} \in \mathbb{R}^{N_\Delta}$, whose elements x_k are interpreted as the amplitude of any event in the interval $(\frac{(2n-1)\Delta}{2}, \frac{(2n+1)\Delta}{2})$. The corresponding prior probability distribution on each x_k is a mixture of a point mass at zero, and $P_A(\cdot)$:

$$P(\vec{x}) = \prod_{k=1}^{N_\Delta} [(1-\mu\Delta)\delta(x_k) + (\mu\Delta)P_A(x_k)] \tag{3}$$

The MAP estimate for this approximate model is obtained by solving:

$$\min_{\vec{x}} \frac{1}{2\sigma^2}\|y(t) - (F_\Delta \vec{x})(t)\|_2^2 - \log(\mu\Delta)\|\vec{x}\|_0 - \sum_{k:x_k\neq 0} \log P_A(x_k) \tag{4}$$

where F_Δ is a linear operator that contains a fixed dictionary of time-shifted copies of $f(\cdot)$:

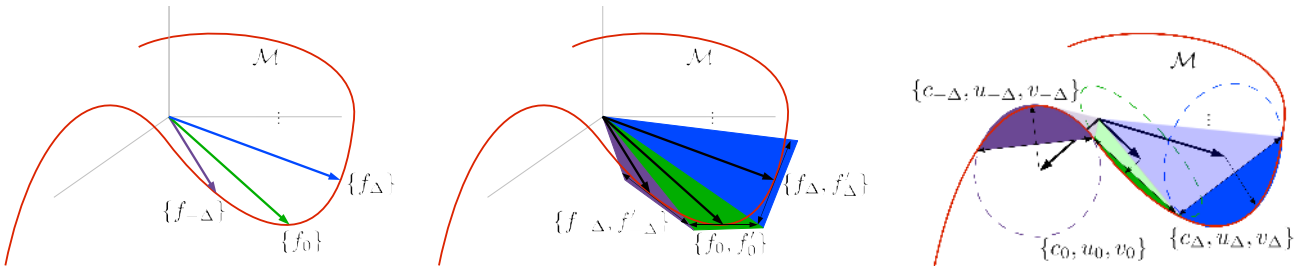
$$(F_\Delta \vec{x})(t) := \sum_{k=1}^{N_\Delta} x_k f(t - k\Delta). \tag{5}$$

This convolutional dictionary is illustrated in Fig. 1(a).

The advantage of the time discretization is that the data fidelity term is now a quadratic function of the parameters (as compared with the nonlinear embedding of the τ_j's in the original formulation of Eq. (2)). But solving Eq. (4) exactly is NP-hard due to the L_0 term [19], and so we must either resort to approximate algorithms, or introduce further approximations in the objective. For our purposes here, we choose the latter path. Specifically, we "convexify" the objective function of Eq. (4) by replacing any nonconvex terms with convex approximations. We use a well-known method known as the LASSO ([7]) or alternatively, basis pursuit denoising ([8]), to replace the L_0 penalty term with an L_1 penalty term.[3] The resulting modified optimization problem

[2] The probability of two or more events is $O(\Delta^2)$, which is negligible for Δ small.

[3] For the purposes of this paper, we assume that the last term in Eq. (4), accounting for the event amplitude probabilities, can also be replaced by an L_1 penalty term.



(a) (b) (c)

Fig. 1. Illustration of the three approximations of the manifold of translates of the waveform, $\mathcal{M}_{f,T}$. (a) The standard basis pursuit (BP) dictionary, F_Δ, as used in Eq. (6), consists of discrete time-shifts of the waveform $f(t)$. (b) Continuous basis pursuit with first-order Taylor interpolator (CBP-T), as specified by Eq. (12). Each pair of functions, $(f_{k\Delta}, f'_{k\Delta})$, with properly constrained coefficients, represents a triangular region of the space (shaded regions). (c) Continuous basis pursuit with polar interpolation (CBP-P), as specified by Eq. (18). Each triplet of functions, $(c_{k\Delta}, u_{k\Delta}, v_{k\Delta})$, represents the surface of a cone (see Fig. 3(b) for parameterization).

becomes:

$$\min_{\vec{x}} \ \frac{1}{2\sigma^2}\|y(t) - (F_\Delta \vec{x})(t)\|_2^2 + \lambda\|\vec{x}\|_1 \qquad (6)$$

where $\lambda > 0$ is a parameter to be determined. Note that probabilistically, the second term indicates that we have effectively approximated the nonconvex mixture prior Eq. (3) by a Laplacian distribution, $P(x) \propto e^{-\lambda|x|}$.

The global optimum of Eq. (6) can be found using standard quadratic programming methods. A well-known article by Candès et al. [10] provides sufficient conditions under which L_1-minimizing solutions are good approximations for L_0-minimizing solutions (up to a term that is linear in the L_2 norm of the noise). Roughly speaking, the condition limits the correlations between small subsets of dictionary elements. However, this condition is rarely satisfied in the convolutional setting, because the validity of the discrete approximation typically requires Δ to be quite small, which leads to a highly correlated dictionary F_Δ.

In addition to these concerns about the tradeoff between the quality of discrete approximation and the convex relaxation, the discretization of event times also leaves us with no well-defined mapping between the solution of Eq. (4) and the continuous parameters $\{\tau_j, a_j\}$ that optimize the original objective of Eq. (2). One can understand the extent of these problems more concretely by focusing on a single time-shifted waveform $f(t-\tau)$ with $\tau \in (0, \Delta)$, and a two-element dictionary F_Δ containing $f(t)$ and $f(t-\Delta)$. Equation (6) is then a 2D problem, as illustated in Fig. 2(a). The solution is a point at which an elliptical level curve of the first (L_2) term is tangent to a straight-line level curve of the L_1 term. Note that the ellipses are stretched in the direction parallel to the L_1 level lines, because trading off amplitude between the two coefficients does not significantly change the reconstruction error when Δ is small (one can show mathematically that the two right singular vectors of the basis matrix are parallel and orthogonal to the L_1 level curves). Also shown in Fig. 2(a) is the family of solutions that are obtained as one varies λ from 0 to $+\infty$ (red path). Notice that these solutions are not sparse in the L_0 sense (i.e., they do not intersect either of the two axes) until λ is so large that the signal reconstruction is quite poor. This is the case regardless of how small Δ is, and is due not only to the failure of the L_1 norm to approximate the L_0 pseudonorm, but also to the inability of the discrete model to account for continuous event times. In the following sections, we develop and demonstrate a proposed solution for these problems.

IV. Continuous basis pursuit

The motivation for the discretization of the inference problem was tractability. Equation (4) approximates the original model well in the limit as Δ goes to 0, but it is only tractably solvable in regimes where Δ is large enough so that correlations are limited and L_1-based relaxations can be employed. In this section, we augment the discrete model by adding variables to account for the continuous nature of the event times, and adapt the LASSO to solve this augmented representation. By accounting for the continuous timeshifts, the augmented model not only can approximate the original model better (and for a larger range of Δ), but also admits sparser solutions via an L_1-based recovery method, which we refer to as *continuous basis pursuit* (CBP).

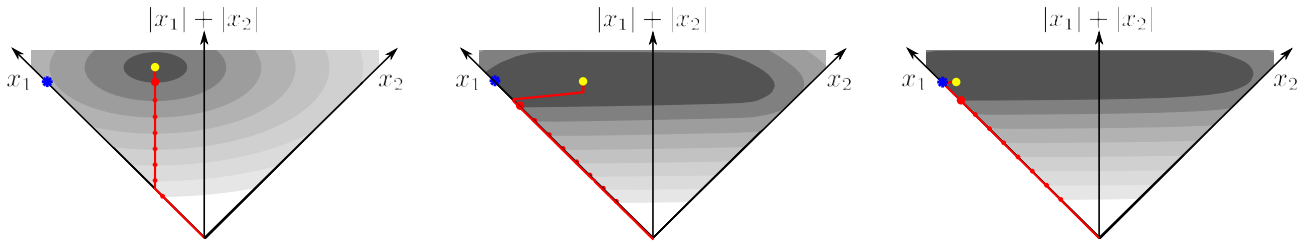
We return to the original continuous problem formulation of Eq. (2), but we assume (without loss of generality) that the waveform is normalized, $\|f(t)\|_2 = 1$, and that the amplitudes, $\{a_j\}$, are all nonnegative. Our noisy observation arises from a linear superposition of time-shifted waveforms, $f(t-\tau)$, which we will abbreviate as $f_\tau(t)$. The set of all time-shifted and amplitude-scaled waveforms forms a 2D nonlinear manifold:

$$\mathcal{M}_{f,T} := \{af(t-\tau) : a \geq 0, \tau \in [0,T]\} \subset L_2([0,T]). \quad (7)$$

The discretized dictionary, F_Δ, provides a linear subspace approximation of this manifold, as illustrated in Fig. 1(a). But the representation of a single element of the manifold (corresponding to a translated scaled copy of the waveform) will typically be approximated by the superposition of several, if not many, elements from the dictionary F_Δ. We can remedy this by augmenting the dictionary to include *interpolation* functions, that allow better approximation of the continuously shifted waveforms. We describe two specific examples of this method, and then provide a general form.

A. Taylor interpolation

If $f(t)$ is differentiable, one can approximate local shifts of $f(t)$ by linearly combining $f(t)$ and its derivative via a first-



(a) (b) (c)

Fig. 2. Geometry of convex objective functions for BP, CBP-T, and CBP-P methods, illustrated in two dimensions. Signal consists of a single waveform, $f(t - 0.65\Delta)$, and the dictionary contains two shifted copies of the waveform, $\{f(t), f(t-\Delta)\}$, along with interpolating functions appropriate for each method. Each plot shows the space of coefficients, $\{x_0, x_1\}$, associated with the two shifted waveforms. Grayscale regions indicate reconstruction accuracy (L_2 norm term of the objective function). The sparsity measure (L_1 norm term of the objective function) is the sum of the two coefficient values, corresponding to the vertical position on the plot. Red curve indicates family of solutions traced out as λ is increased fom 0 (yellow dot) to infinity. Red points along this path indicate equal increments in reconstruction accuracy, with enlarged point indicating a common value for comparison across all three methods. Blue dot indicates to the true L_0-minimizing solution. (a) Standard BP solution (Eq. (6)). (b) Continuous basis pursuit with Taylor interpolator (Eq. (12)), using a basis of $\{f_0, f_0', f_\Delta, f_\Delta'\}$. In this case, the iso-accuracy curves are computed by minimizing the L_2 error over all feasible values of the derivative coefficients, and are thus no longer elliptical. (c) CBP with polar interpolation (Eq. (18)).

order Taylor expansion:

$$f_\tau(t) = f(t) - \tau f'(t) + O(\tau^2) \tag{8}$$

This motivates a dictionary consisting of the original shifted waveforms, $\{f_{k\Delta}(t)\}$, and their derivatives, $\{f'_{k\Delta}(t)\}$. We choose a basis function spacing, Δ, as twice the maximal timeshift such that the first-order Taylor approximation holds within a desired accuracy, δ:

$$\Delta := \max\{\Delta' : \max_{|\tau|<\frac{\Delta'}{2}} \|f_\tau(t) - (f(t) - \tau f'(t))\|_2 \leq \delta\} \tag{9}$$

We can then approximate the manifold of scaled and time-shifted waveforms using *constrained* linear combinations of dictionary elements:

$$\mathcal{M}_{f,T} \approx \left\{ x f_{k\Delta}(t) + d f'_{k\Delta}(t) \quad : \quad \begin{array}{l} x \geq 0, \\ |d| \leq \frac{\Delta}{2} x, \\ k = 1, ..., N_\Delta \end{array} \right\} \tag{10}$$

There is a one-to-one correspondence between sums of points on the manifold $\mathcal{M}_{f,T}$ and their respective approximations with this dictionary:

$$\sum_k x_k f_{k\Delta}(t) + d_k f'_{k\Delta}(t) \approx \sum_k x_k f_{(k\Delta - d_k/x_k)}(t). \tag{11}$$

This holds as long as $|d_k/x_k| \neq \frac{\Delta}{2}$ (which corresponds to the situation where the the waveform is displaced exactly halfway in between two lattice points, and can thus be equally well represented by the basis function and associated derivative on either side). This is illustrated in Fig. 1(b).

The inference problem is now solved by optimizing a *constrained* convex objective function:

$$\min_{\vec{x},\vec{d}} \frac{1}{2\sigma^2} \left\| y(t) - (F_\Delta \vec{x})(t) - (F'_\Delta \vec{d})(t) \right\|_2^2 + \lambda \|\vec{x}\|_1$$

$$\text{s.t.} \quad \left\{ \begin{array}{l} x_k \geq 0, \\ |d_k| \leq \frac{\Delta}{2} x_k \end{array} \right\} \text{ for k} = 1, ..., \text{N}_\Delta \tag{12}$$

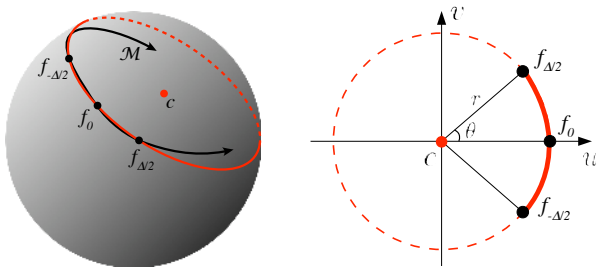
where the dictionary F_Δ is defined as in Eq. (5), and F'_Δ is a dictionary of time-shifted waveform derivatives $\{f'_{k\Delta}(t)\}$. Equation (11) provides an explicit mapping from appropriately constrained coefficient configurations to event amplitudes and timeshifts. Figure 2(b) illustrates this objective function for the same single-waveform example described previously. The shaded regions are the level sets of the L_2 term of Eq. (12) visualized in the (x_1, x_2)-plane by minimizing over the derivative coefficients (d_1, d_2). Note that unlike the corresponding BP level sets shown in Fig. 2(a), these are no longer elliptical, and that they allow sparse solutions (i.e., points on the $x_1 = 0$ axis) with low reconstruction error. As a result, for λ sufficiently large, the solution of Eq. (12) is not only sparse in the L_0 sense, but also provides a good reconstruction of the signal.

B. Polar interpolation

Although the Taylor series provides the most intuitive and well-known method of approximating time-shifts, we have developed an alternative interpolator that is significantly more accurate. The solution is motivated by the observation that the manifold of time-shifted waveforms, $f_\tau(t)$, must lie on the surface of a unit hypersphere (because the waveform L_2-norm is preserved under time shifting), and furthermore, must have a constant curvature (by symmetry). This leads to the notion that it might be well-approximated by an arc of a circle. As such, we approximate a segment of the manifold, $\{f_\tau : |\tau| \leq \frac{\Delta}{2}\}$, by the unique circular arc that contains the three points $\{f_{-\Delta/2}, f_0, f_{\Delta/2}\}$, as illustrated in Fig. 3(a). The resulting interpolator is an example of a trigonometric spline [20], in which the three time-shifted functions are linearly combined using trigonometric coefficients to approximate intermediate translates of $f(t)$:

$$f_\tau(t) \approx c(t) + r\cos(\frac{2\tau}{\Delta}\theta)u(t) + r\sin(\frac{2\tau}{\Delta}\theta)v(t) \tag{13}$$

where the functions $\{c(t), u(t), v(t)\}$ are computed from linear combinations of $\{f_{-\Delta/2}, f_0, f_{\Delta/2}\}$:



(a) (b)

Fig. 3. Illustration of the polar interpolator. (a) The manifold of time shifts of $f(t)$ (black line) lies on the surface of a hypersphere. We approximate a segment of this manifold, for time shifts $\tau \in [-\frac{\Delta}{2}, \frac{\Delta}{2}]$, with a portion of a circle (red), with center defined by $c(t)$. (b) Parameterization of the circular arc approximation.

$$\begin{pmatrix} f_{-\frac{\Delta}{2}}(t) \\ f_0(t) \\ f_{\frac{\Delta}{2}}(t) \end{pmatrix} = \begin{pmatrix} 1 & r\cos(\theta) & -r\sin(\theta) \\ 1 & r & 0 \\ 1 & r\cos(\theta) & r\sin(\theta) \end{pmatrix} \begin{pmatrix} c(t) \\ u(t) \\ v(t) \end{pmatrix} \quad (14)$$

The constant r is the radius of the circular arc, and θ is half the angle subtended by the arc, both of which depend on $f(t)$ and can be computed in closed form. These relationships are illustrated in Fig. 3(b). The approximation can be easily expressed in the frequency domain, by taking the Fourier transform of both sides of Eq. (13) and using Eq. (14):

$$\begin{aligned} e^{-i\omega\tau} \approx (1 - 2a(\tau)) \; + \\ e^{i\omega\frac{\Delta}{2}}(a(\tau) - b(\tau)) \; + \\ e^{-i\omega\frac{\Delta}{2}}(a(\tau) + b(\tau)) \end{aligned} \quad (15)$$

where

$$a(\tau) = \frac{\cos(\frac{2\tau\theta}{\Delta}) - 1}{2(\cos(\theta) - 1)} \quad \text{and} \quad \mathrm{b}(\tau) = \frac{\sin(\frac{2\tau\theta}{\Delta})}{2\sin(\theta)}.$$

Figure. 4(a) compares nearest neighbor (as is implicitly used in BP), first-order Taylor, and polar interpolation in terms of their accuracy in approximating timeshifts of a Gaussian derivative waveform, $f(t) \propto te^{-\alpha t^2}$. For reference, the second-order Taylor interpolator is also included. The polar interpolator is seen to be significantly more accurate than nearest-neighbor and 1st-order Taylor, and even surpasses 2nd-order Taylor by an order of magnitude (although they have the same asymptotic rate of convergence). This allows one to choose a much larger Δ for a given desired accuracy.

We now construct a dictionary of time-shifted copies of the functions used to represent the polar interpolation, $\{c_{k\Delta}, u_{k\Delta}, v_{k\Delta}\}$, and form a convex set from these to approximate the manifold:

$$\mathcal{M}_{f,T} \approx \left\{ \begin{array}{l} xc_{k\Delta}(t) \\ \quad + yu_{k\Delta}(t) \\ \quad + zv_{k\Delta}(t) \end{array} \; : \; \begin{array}{l} x \geq 0, \\ y^2 + z^2 \leq x^2 r^2, \\ xr\cos(\theta) \leq y \leq xr, \\ k = 1, ...N_\Delta \end{array} \right\} \quad (16)$$

The constraints on the coefficients (x, y, z) ensure that they represent a scaled translate of $f(t)$, except for the second, which is a convex relaxation of the true constraint, $y^2 + z^2 = x^2 r^2$ (see below). As with the Taylor approximation, we have a one-to-one correspondence between event amplitudes/timeshifts and the constrained coefficients:

$$\begin{aligned} \sum_k x_k c_{k\Delta}(t) + y_k u_{k\Delta}(t) + z_k v_{k\Delta}(t) \\ \approx \sum_k x_k f_{(k\Delta - \frac{\Delta}{2\theta}\tan^{-1}(z_k/y_k))}(t) \end{aligned} \quad (17)$$

as long as $z_k/y_k \neq \tan(\theta)$ for all k. The inference problem again boils down to minimizing a constrained convex objective function:

$$\min_{\vec{x},\vec{y},\vec{z}} \frac{1}{2\sigma^2} \|y(t) - (C_\Delta \vec{x})(t) - (U_\Delta \vec{y})(t) - (V_\Delta \vec{z})(t)\|_2^2 + \lambda \|\vec{x}\|_1$$

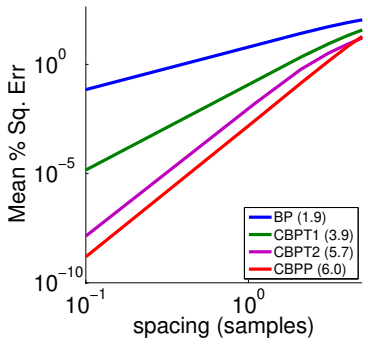
$$\text{s.t.} \quad \left\{ \begin{array}{l} x_k \geq 0, \\ \sqrt{y_k^2 + z_k^2} \leq x_k r, \\ x_k r\cos(\theta) \leq y_k \leq x_k r, \end{array} \right\} \text{ for } \mathrm{k} = 1, ...\mathrm{N}_\Delta \quad (18)$$

where $C_\Delta, U_\Delta, V_\Delta$ are dictionaries containing Δ-shifted copies of $c(t), u(t), v(t)$, respectively. Equation (18) is an example of a "second-order cone program" for which efficient solvers exist ([21]). After the optimum values for $\{\vec{x}, \vec{y}, \vec{z}\}$ are obtained, timeshifts and amplitudes can be inferred by first projecting the solution back to the original constraint set:

$$(x_k, y_k, z_k) \leftarrow (x_k, \frac{y_k x_k r}{\sqrt{y_k^2 + z_k^2}}, \frac{z_k x_k r}{\sqrt{y_k^2 + z_k^2}}) \quad (19)$$

and then using Eq. (17) to solve for the event times.

Figure 2(c) illustrates the optimization of Eq. (18) for the simple example described in the previous section. Notice that the solution corresponding to $\lambda = 0$ (yellow dot) is significantly sparser relative to both the CBP-T and BP solutions, and that the solution becomes L_0 sparse if λ is increased by just a small amount, giving up very little reconstruction accuracy.



(a)

Fig. 4. Comparison of the nearest neighbor, first-order Taylor, and polar interpolators (as used in BP, CBP-T, and CBP-P, respectively) for a waveform $f(t) \propto te^{-\alpha t^2}$. Sinc and 2nd-order Taylor interpolation are also shown. The estimated slopes (asymptotic rates of convergence) are shown in the legend.

C. General interpolation

We can generalize the CBP approach to use any linear interpolation scheme. Suppose we have a set of basis functions

$\{\phi_n(t)\}_1^m$ in $L_2([0,T])$ (for simplicity, assume they are orthonormal) and a corresponding interpolation map $\vec{D}(\cdot)$ such that local shifts can be approximated as:

$$f_\tau(t) \approx \sum_{n=1}^{m} D_n(\tau)\phi_n(t), \quad |\tau| \leq \frac{\Delta}{2}. \tag{20}$$

Let S be the set of all nonnegative scalings of the image of $[-\frac{\Delta}{2}, \frac{\Delta}{2}]$ under the interpolator:

$$S = \{a\vec{D}(\tau) : a \geq 0, \tau \leq \frac{\Delta}{2}\}.$$

If the interpolator $\vec{D}(\tau)$ is invertible, we have a one-to-one correspondence as before:

$$\sum_{k=1}^{N_\Delta}\sum_{n=1}^{m} x_{kn}\phi_n(t-k\Delta) \approx \sum_{k=1}^{N_\Delta} \|\vec{x}_k\|_2 f_{(k\Delta - \vec{D}^{(-1)}(\vec{x}_k/\|\vec{x}_k\|_2))}(t) \tag{21}$$

where each group $\vec{x}_k := [x_{k1}, ..., x_{km}]$ is in S and $|\vec{D}^{(-1)}(\vec{x}_k/\|\vec{x}_k\|_2)| \neq \frac{\Delta}{2}$ for all k. Note that in this general form, the L_2 norm of each group $\vec{x_k}$ governs the amplitude of the corresponding time-shifted waveform.[4] As we saw in the previous examples, S may or may not be convex, so we relax to its convex hull, denoted by $\overline{S}$, keeping in mind that we must project our solution back onto S at the end, using an operator $P_S(\cdot)$.

Finally, we can write obtain the representation using this interpolation by solving:

$$\min_{\vec{x}} \ \frac{1}{2\sigma^2}\|y(t) - (\Phi_\Delta \vec{x})(t)\|_2^2 + \lambda \sum_{k=1}^{N_\Delta} \|\vec{x}_k\|_2 \tag{22}$$
$$\text{s.t.} \quad \vec{x}_k \in \overline{S} \quad \text{for} \quad k = 1, ..., N_\Delta$$

where the linear operator Φ_Δ is defined as:

$$(\Phi_\Delta \vec{x})(t) := \sum_{k=1}^{N_\Delta}\sum_{n=1}^{m} x_{kn}\phi_n(t-k\Delta)$$

Equation (22) can be solved efficiently using standard convex optimization methods (e.g., interior point methods [21]). It is similar to the objective functions used to recover so-called "block-sparse" signals (e.g., [16], [22]), but includes auxilliary constraints on the coefficients to ensure that only signals close to $span(\mathcal{M}_{f,T})$ are represented. Table I summarizes the Taylor and polar interpolation examples within the general framework, along with the case of nearest-neighbor interpolation (which corresponds to standard BP).

The quality of the solution relies on the accuracy of the interpolator, the convex approximation $\overline{S} \approx S$, and the ability of the block-L_1 based penalty term in Eq. (22) to achieve L_0-sparse solutions that reconstruct the signal accurately. The first two of these are relatively straightforward, since they depend solely on the properties of the interpolator (see Fig. 4(a)). The last is difficult to predict, even for the simple examples illustrated in Figure 2. The level sets of the L_2 term can have a complicated form when taking the constraints into account, and it is not clear a priori whether this will facilitate or hinder the L_1 term in achieving sparse solutions. Nevertheless, our empirical results clearly indicate that solving Eq. (22) with Taylor and polar interpolators yields significantly sparser solutions than those achieved with standard BP.

[4]Our specific examples used the amplitude of a single coefficient as opposed to the group L_2 norm. However, the constraints in these examples make the two formulations equivalent up to $O(\Delta)$. For the Taylor interpolator, $x_k^2 \approx x_k^2 + d_k^2$. For the polar interpolator, $c_k^2 + u_k^2 + v_k^2 \approx 2c_k^2$.

V. Empirical results

We evaluate our method on data simulated according to the generative model of Eq. (1). We chose the probability density on event amplitudes, $P_A(\cdot)$, to be uniform on the interval $[a, b]$ with $0 < a < b$. We used a single template waveform $f(t) \propto te^{-\alpha t^2}$ (normalized, so that $\|f\|_2 = 1$), for which the interpolator performances are plotted in Fig. 4(a). We compared solutions of Eqs. (6), (12), and (18). In all recovery methods, amplitudes were constrained to be nonnegative (this is already assumed for the CBP methods, and amounts to an additional linear inequality constraint for BP). Each method has two free parameters: Δ controls the spacing of the basis, and λ controls the tradeoff between reconstruction error and sparsity. We varied these parameters systematically and measured performance in terms of two quantities: (1) signal reconstruction error (which decreases as λ increases or Δ decreases), and (2) sparsity of the estimated event amplitudes ($\{\|\vec{x}_j\|_2\}$), which increases as λ increases. The former is simply the first term in the objective function (for all three methods). For the latter, to ensure numerical stability, we used the L_p norm with $p = 0.1$ (results were stable with respect to the choice of p, as long as $p < 1$ and p was not below the numerical precision of the optimizations. Computations were performed numerically, by sampling the functions $f(t)$ and $y(t)$ at a fine constant spacing. We used the convex solver package CVX [23] to obtain numerical solutions.

A small temporal window of the events recovered by the three methods is provided in Figure 5. The three plots show the estimated event times and amplitudes for BP, CBP-T, and CBP-P (upward stems) compared to the true event times/amplitudes (downward stems). The figure demonstrates that CBP, equipped with either Taylor or polar interpolators, is able to recover the event train more accurately, and with a larger spacing between basis functions (indicated by the tick marks on the x-axis). As predicted by the reasoning laid out in Figure 2(a), basis pursuit tends to split events across two or more adjacent low-amplitude coefficients, thus producing less sparse solutions and making it hard to infer the number of events and their respective amplitudes and times. Sparsity can be improved by increasing λ, but at the expense of a substantial increase in approximation error.

Figure 6 illustrates the tradeoff between sparsity and approximation error for each of the methods. Each panel corresponds to a different noise level. The individual points, color-coded for each method, are obtained by running the associated method 500 times for a given (Δ, λ) combination, and averaging the errors over these trials. The solid curves are the (numerically

Property	BP (nearest-neighbor)	CBP - Taylor interp	CBP polar interp
basis: $\{\phi_n(t)\}_{n=1}^m$	$[f(t)]$	$[f(t), f'(t)]$	$[c(t), u(t), v(t)]$
interpolator: $\vec{D}(\tau)$	1	$[1, \tau]^T$	$[1, r\cos(\theta\frac{2\tau}{\Delta}), r\sin(\theta\frac{2\tau}{\Delta})]^T$
constrained coefficient set: S	$\{x_1 \geq 0\}$	$\{x_1 \geq 0, \lvert x_2\rvert \leq x_1\frac{\Delta}{2}\}$	$\{x_1 \geq 0, x_2^2 + x_3^2 = r^2x_1^2, rx_1\cos(\theta) \leq x_2 \leq rx_1\}$
convex relaxation: $\overline{S}$	S	S	$\{x_1 \geq 0, x_2^2 + x_3^2 \leq r^2x_1^2, rx_1\cos(\theta) \leq x_2 \leq rx_1\}$
projection operator: $P_S(\vec{x})$	$\vec{x}$	$\vec{x}$	$[x_1, rx_1\frac{x_2}{\sqrt{x_2^2+x_3^2}}, rx_1\frac{x_3}{\sqrt{x_2^2+x_3^2}}]^T$

TABLE I

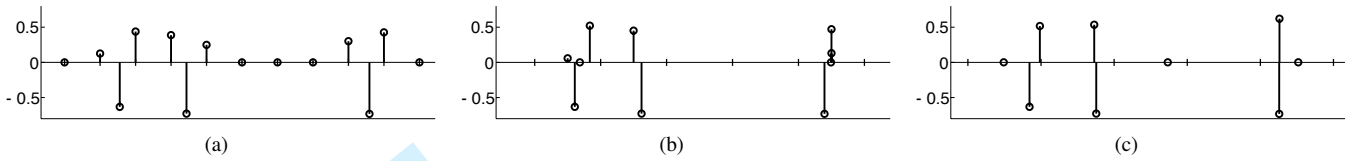


Fig. 5. Example of sparse signal recovery for (a) BP (Eq. (6)), (b) CBP-T (Eq. (12)), and (c) CBP-P (Eq. (18)). For each method, the values of Δ and λ were chosen to minimize the average sum of squares of the two types of error. Upward stems indicate the estimated magnitudes placed at locations determined by the interpolation coefficients via Eq. (21). Ticks denote the location of the basis functions corresponding to each upward-pointing stem. Downward stems indicate the locations and magnitudes of the true signal. SNR was 12 (identical signal and noise for all three examples).

computed) convex hulls of all points obtained for each method, and clearly indicate the tradeoff between the two types of error. We can see that the performance of BP is strictly dominated by that of CBP-T: For every BP solution, there is a CBP-T solution that has lower values for both error types. Similarly, CBP-T is strictly dominated by CBP-P, which can be seen to come close to the error values of the ground truth answer (which is indicated by a black X).

We performed a signal detection analysis of the performance of these methods, classifying identification errors as misses and false positives. Given a true and estimated event train, we say that an event is matched across the two if (1) the estimated event's amplitude is within some threshold $\alpha > 0$ of the true amplitude and (2) the estimated event time is within some threshold $\nu > 0$ of the true event time, and (3) no other estimated event has been matched to the true event. We evaluated the three methods using these criteria, using a value of $\alpha = \frac{1}{\sqrt{12}}$ (one standard deviation of the amplitude distribution Unif$[0.5, 1.5]$)) and $\nu = 3$ samples. We found that results were relatively stable with respect to these threshold choices. For each method and noise level we chose the (λ, Δ) combination yielding a solution closest to ground truth (corresponding to the large dots in Figure 6). Figure 7 shows the errors as a function of the noise level. We see that performance of all methods is surprisingly stable across SNR levels. We also see that BP performance is dominated at all noise levels by CBP-T, which has fewer misses as well as fewer false positives, and CBP-T is similarly dominated by CBP-P.

Finally, we examined the distribution of the amplitudes estimated by each algorithm, and compare with the distribution of the source, as given by Eq. (3). Figure 8 shows the amplitude histogram for each method. We see that CBP-P produces amplitude distributions that are far better-matched to the correct distribution of amplitudes.

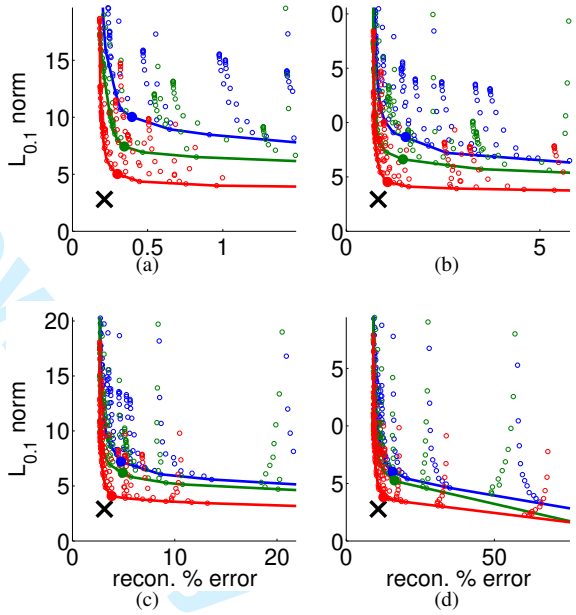
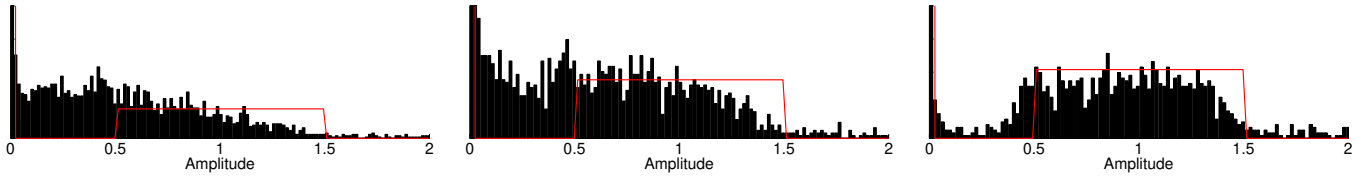


Fig. 6. Error plots for four noise levels: (a) SNR = 48dB (b) SNR = 24dB (c) SNR = 12dB (d) and SNR = 6dB, where SNR is defined as $\|f\|_\infty/\sigma$). Each graph shows the tradeoff between the average reconstruction error (vertical axis) and the sparsity (horizontal axis, measured as average $L_{0.1}$ norm of estimated amplitudes). Each point represents the error values for one of the methods, applied with a particular setting of (Δ, λ), averaged over 500 trials. Colors indicate the method used (BP-blue,CBP-T-green CBP-P-red). Bold lines denote the convex hulls of all points for each method. The large dots indicate the "best" solution as measured by Euclidean distance from the correct solution (indicated by black X's).

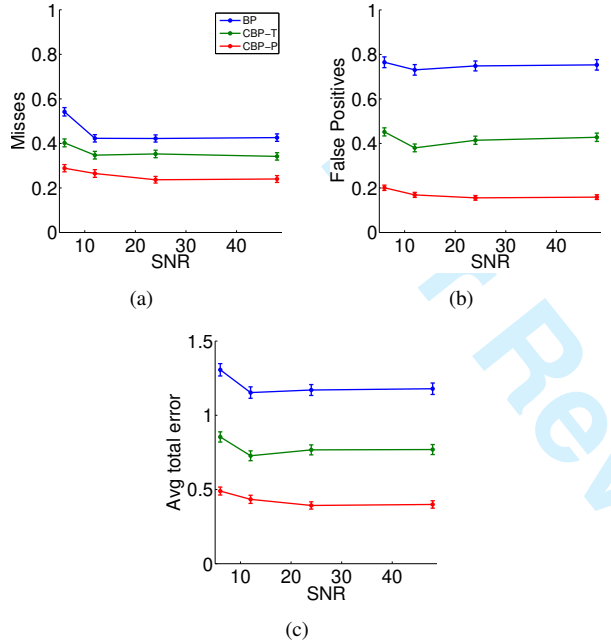
A. *Multiple features*

All of the methods we've described can be easily extended to the case of multiple templates, by taking as a dictionary the union of dictionaries associated with each individual template. We performed a final set of experiments for the case of two features (waveforms shown in Fig. 9(a)) that are "gammatone" filters, as commonly used in audio processing. Data were



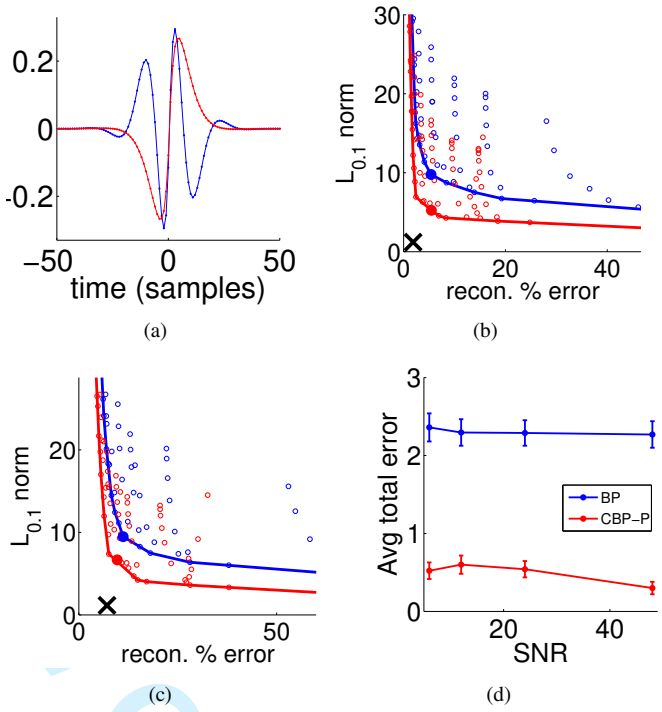
(a) (b) (c)

Fig. 8. Histograms of the estimated amplitudes for (a) BP, (b) CBP-T, and (c) CBP-P. All methods were constrained to estimate only nonnegative amplitudes, but no upper bound was imposed. The true distribution of amplitudes is given by Eq. (3), and is indicated in red.



(a) (b) (c)

Fig. 7. Signal detection analysis of solutions (see text). (a) Average miss rate (as a fraction of the mean number of events per trial) computed over 500 trials for each method, and for each SNR (defined as $\|f\|_\infty/\sigma$). (b) Average false positive rate. (c) Total error (sum of misses and false positives)



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

Fig. 9. (a) Two gammatone features of the form $f_i(t) = at^{n-1}e^{-2\pi bt}\cos(2\pi\omega_i t)$ for $i = 1, 2$. (b) and (c) show the sparsity and reconstruction errors for BP (blue) and CBP-P (red), as in Figure 6, with SNRs of 24 and 12, respectively. (d) plots the total number of misses and false positives (with same thresholds as in Figure 7(c)) for each method.

generated by constructing two correlated Poisson processes with the same marginal rate λ and a correlation of $\rho = 0.5$. These were generated by independently creating 2 Poisson process with rate $\lambda(1-\rho)$ and then superimposing a randomly jittered "common" Poisson process with rate $\lambda\rho$. As before, event amplitudes were drawn independently from a uniform distribution on $[a, b]$ with $a > 0$.

We examined and compared performance of BP and CBP-P. Both methods used dictionaries formed from the union of dictionaries for each template, but we forced the two individual dictionaries to use a common spacing, Δ, for both waveforms. In general, the spacing could be chosen differently for each waveform, providing more flexibility, at the expense of additional parameters that must be selected or optimized. Figures 9(b) and 9(c) show the error tradeoff for different settings of (Δ, λ) at SNR levels of 24 and 12, respectively (the results were qualitatively unchanged for SNR values of 48 and 6). Figure 9(d) shows the total number of event identification errors (misses plus false positives) for each method as a function of SNR at each methods optimal (Δ, λ) setting.

VI. Discussion

We have introduced a novel methodology for sparse signal decomposition in terms of continously shifted features. The method can be seen as a continuous form of the well-known basis pursuit method, and we thus have dubbed it *Continuous Basis Pursuit*. The method overcomes the limitation of basis pursuit in the convolutional setting casued by the tradeoff between discretization error and the effectiveness of the L_1 relaxation for obtaining sparse solutions. In particular, our method employs an alternative discrete basis (not necessarily the features themselves) which can explicitly account for the continuous timeshifts present in the signal. We derived a general convex objective function that can be used with any such basis. The coefficients are constrained so as to represent only transformed versions of the templates, and the objective function uses an L_1 norm to penalize the amplitudes of these transformed versions . We showed empirically that using simple first-order (Taylor) and second-order (polar) in-

terpolation schemes yields superior solutions in the sense that (1) they are sparser, with improved reconstruction accuracy, (2) they produce substantially better identification of events (fewer misses and false positives), and (3) the amplitude statistics are a better match to those of the true generative model. We showed that these results are stable across a wide range of noise levels. We conclude that an interpolating basis, coupled with appropriate constraints on coefficients, provides a powerful and tractable tool for modeling and decomposing translation-invariant signals.

We believe our method can be extended for use with other types of signal, as well as to tranformations other than translation. For example, for one dimensional signals such as audio, one might also include dilation or frequency-modulation of the templates. For two-dimensional signals, such as photographic images, one could include rotation. For each of these, the primary hurdles are to specify (1) the form of the linear interpolation (for joint variables, this might be done separably, or using a multi-dimensional interpolator), (2) the constraints on coefficients (and a convex relaxation of these constraints), and (3) a means of inverting the interpolator so as to obtain transformation parameters from recovered coefficients. Another natural extension is to use CBP in the context of learning optimal templates for decomposing an ensemble of signals, as has been previously done with BP (e.g., [13], [14], [12], [15], [24]).

Acknowledgment

The authors would like to thank Sinan Güntürk for helpful discussions in the early stages of this work.

References

[1] J. H. Friedman and J. W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Trans Computers*, C-23(9):881–890, 1974.

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

[3] J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory*, 50(10):2231–2242, 2004.

[4] T. Blumensath and M. E. Davies. Iterative thresholding for sparse approximations. *J. Fourier Analysis and Applications*, 2004.

[5] M. Elad. Why simple shrinkage is still relevant for redundant representations. *IEEE Trans Info Theory*, 52:5559–5569, 2006.

[6] J. Bioucas-Dias and M. Figueiredo. A new TwIST: Two-step iterative shrinkage/thresholding algorithms for image restoration. *IEEE Trans Image Processing*, 16(12):2992–3004, 2007.

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

[8] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

[9] Emmanuel C and Terence Tao. The dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351, 2005.

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

[11] Martin J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). *IEEE Trans. Inf. Theor.*, 55:2183–2202, May 2009.

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

[13] Phil Sallee and Bruno A. Olshausen. Learning sparse multiscale image representations. In *NIPS*, pages 1327–1334, 2002.

[14] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, Jun 1996.

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

[16] 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.

[17] J. Mendel. *Optimal Seismic Deconvolution: An Estimation Based Approach*. Academic Pr, 1983.

[18] M. S. Lewicki. A review of methods for spike sorting: the detection and classification of neural action potentials. *Network*, 9(4):R53–R78, Nov 1998.

[19] G. Davis, S. Mallat, and M. Avellaneda. Adaptive greedy approximations. *Constructive Approximation*, 13:57–98, 1997. 10.1007/BF02678430.

[20] I.J. Schoenberg. On trigonometric spline interpolation. *Journal of Math Mech.*, 13:795–825, 1964.

[21] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[22] 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.

[23] M. Grant and S. Boyd. Cvx: Matlab software for disciplined convex programming, version 1.21. http://cvxr.com/cvx, October 2010.

[24] P. Berkes, R. E. Turner, and M. Sahani. A structured model of video reproduces primary visual cortical organisation. *PLoS Computational Biology*, 5(9), 2009.

PLACE PHOTO HERE

Chaitanya Ekanadham Biography text here.

PLACE PHOTO HERE

Daniel Tranchina Biography text here.

PLACE PHOTO HERE

Eero P. Simoncelli received the BS degree (summa cum laude) in physics from Harvard University in 1984 and the MS and PhD degrees in electrical engineering from Massachusetts Institute of Technology in 1988 and 1993, respectively. He studied applied mathematics at Cambridge University for a year and a half. He was an assistant professor in the Computer and Information Science Department at the Univer- sity of Pennsylvania from 1993 to 1996. In September 1996, he moved to New York University, where he is currently an associate professor in neural science and mathematics. In August 2000, he became an associate investigator at the Howard Hughes Medical Institute under their new program in computational biology. His research interests span a wide range of topics in the representation and analysis of visual images, in both machine and biological systems. He is a fellow of the IEEE.