Performance of Compressive Parameter Estimation via K-Median Clustering ☆

Dian Mo[a], Marco F. Duarte[a]

[a]*Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, MA 01003*

Abstract

Compressive sensing (CS) has attracted significant attention in parameter estimation tasks, where parametric dictionaries (PDs) collect signal observations for a sampling of the parameter space and yield sparse representations for signals of interest when the sampling is dense. While this sampling also leads to high dictionary coherence, one can leverage structured sparsity models to prevent highly coherent dictionary elements from appearing simultaneously in the recovered signal. However, the resulting approaches depend heavily on the careful setting of the maximum allowable coherence; furthermore, their guarantees are not concerned with general parameter estimation performance. We propose the use of earth mover's distance (EMD), as applied to a pair of true and estimated PD coefficient vectors, to measure the parameter estimation error. We formally analyze the connection between the EMD and the parameter estimation error and show that the EMD provides a better-suited metric for parameter estimation performance than the Euclidean distance. Additionally, we analyze the previously described relationship between K-median clustering and EMD-optimal sparse approximation and leverage it to develop improved PD-based parameter estimation algorithms. Finally, our numerical experiments verify our theoretical results and show that the proposed compressive parameter estimation algorithms have similar performance as the state-of-the-art algorithms while featuring simpler implementation and broader applicability.

Keywords: compressive sensing; parameter estimation; parametric dictionary; earth mover's distance; K-median clustering

1. Introduction

Compressive sensing (CS) simultaneously acquires and compresses signals via random projections, and recovers a signal if it is sparse or if there exists a basis or dictionary in which it can be expressed sparsely [3–5].

☆Early versions of this work appeared at Proceedings of SPIE Wavelets and Sparsity XV, August 2013 [1] and 2016 Annual Conference on Information Science and Systems[2]

Email addresses: mo@umass.edu (Dian Mo), mduarte@ecs.umass.edu (Marco F. Duarte)

Recently, the application of CS has been extended from signal recovery to parameter estimation through the design of parametric dictionaries (PDs) that contain signal observations for a sampling set of the parameter space [6–17]. The resulting connection between parameter estimation and sparse signal recovery has made it possible for compressive parameter estimation to be implemented via standard CS recovery algorithms, where the PD coefficients obtained from recovery algorithms can be interpreted by matching the locations of the nonzero coefficients with the parameter estimates. This CS approach has been previously formulated for many landmark compressive parameter estimation problems, including localization and bearing estimation [6–11], time delay estimation [12, 13], and frequency estimation (also known as line spectral estimation when the frequencies lie in a one-dimensional space) [14–18].

Unfortunately, only in the contrived case when the unknown parameters are all contained in the sampling set of the parameter space can the PD-based parameter estimation be perfect. In frequency estimation, the mismatch between the unknown frequencies and the discretized frequency samples results in loss of sparsity due to spectral leakage, therefore reducing recovery performance [19]. Additionally, the guarantees of bounded error between the true and the estimated PD coefficient vectors, measured via the Euclidean distance in almost all existing methods, have very limited impact on the performance of compressive parameter estimation given that the Euclidean distance does not relate to the parameter estimation error. In contrast, the earth mover's distance (EMD) [20–22] is a very attractive option due to the fact that the EMD of two sparse PD coefficient vectors is indicative of the parameter estimation error when the elements of the PD are sorted in increasing order of the values of the corresponding parameters.

1.1. Prior Work

The relevant literature includes several approaches that rely on convex optimization to avoid parameter discretization issues. Recently, total variation norm minimization has been proposed to estimate frequency-sparse signals from observations whose spectrum is confined within some set of low frequencies, a setup that is now known as the super-resolution problem [23]. This algorithm allows for accurate frequency estimation when the unknown frequencies are sufficiently separated, even under additive noise [24]. Inspired by this approach, when the frequencies are well separated, atomic norm minimization has been developed to estimate frequency-sparse signals with line spectra from samples obtained at random times [25, 26]. An extension to two-dimensional frequency estimation is available in [27]. Another method uses matrix completion to recover signals with arbitrary frequencies from random samples when a separation condition is met [28]. The main result is that the sparse recovery problem can be solved via low-rank matrix completion of a matrix possessing Hankel structure derived from the sparse signal. Similar performance guarantees have been established as

well for signal recovery from linear measurements [29]. The main drawback of optimization methods is that the performance is highly dependent on convex optimization solvers, which usually can be very time-consuming. Additionally, such methods are hard to extend to other parameter estimation problems such as time delay estimation.

For PD-based parameter estimation, a dense sampling of the parameter space is needed to improve the estimation resolution [7]. However, increasing the sampling density introduces highly coherent elements in the PDs, which might affect the performance and complexity of sparse recovery algorithms. Previous approaches to address the resulting coherence issues utilized a coherence-inhibiting structured sparse approximation where the resulting nonzero entries of the reconstructed coefficient vector correspond to the PD elements that have sufficiently low coherence [7, 16, 18]. Based on structured sparse recovery, an auxiliary interpolation has been proposed as an additional step to break through the limitation of discretization and improve the estimation performance [13, 17]. Those approaches need a careful setting of the value of the maximum allowable coherence among the chosen PD elements in the sparse representation. Setting the maximum allowable coherence to be too large restricts the minimum separation between any two parameters that can be observed simultaneously in a signal, and can potentially exclude a large class of observable signals from consideration.

EMD has been proposed as an alternative measure for sparse recovery [22]. Instead of performing constrained EMD minimization directly, a pyramid transform that connects a sparse model and a tree-sparse model allows for the implementation of sparse recovery with error measured by EMD using a constrained ℓ_1-norm minimization problem. However, the special requirements of the measurement matrix limit the use of such EMD-optimal sparse recovery in many practical applications.

1.2. Contribution

In this paper, we proposed a new PD-based method for compressive parameter estimation that replaces the hard thresholding operator, a widely used sparse approximation operator in sparse recovery algorithms to solve the sparse approximation problem under Euclidean distance, with K-median clustering, a novel sparse approximation operator that returns the optimal sparse approximation vector in terms of the EMD [21, 22]. Our main contributions can be detailed as follows. First, we show theoretically that the EMD between the sparse PD coefficient vectors provides an upper bound for the parameter estimation error, motivating the use of algorithms minimizing the EMD between coefficient vectors to achieve small parameter estimation error. Second, we formulate theorems that provide performance guarantees for parameter estimation with our proposed clustering methods; these guarantees are based on the correlation between PD elements. Third,

3

we analyze the effect of the decay of the correlation between PD elements on the performance of our proposed method, and elaborate on our guarantees when CS compression and signal noise are involved. Finally, we introduce and analyze the joint use of thresholding and clustering methods to address performance loss resulting from compression and noise. Although this paper focuses on one-dimensional parameters, our work can be easily extended to multidimensional parameter estimation by using K-median clustering for multidimensional spaces.

One of the essential advantages of our proposed algorithm is that it can be applied to arbitrary sparsity-based parameter estimation problems. As shown in the sequel, existing greedy algorithms require an additional parameter to prevent highly coherent PD elements from appearing in the signal representation simultaneously and estimate the correct parameters. As shown in our prior work, such algorithms are very sensitive to the chosen value of the additional parameter: an improper value of the additional parameter is significantly harmful to the estimation performance [1]. However, to the best of our knowledge, there is no guarantee that existing greedy algorithms with a specific parameter value will have good performance for parameter estimation under suitable conditions. Furthermore, we do not know of a procedure to set the parameter values so that the existing algorithms will have good performance under the given conditions except for a grid search that chooses the one that performs best experimentally. In contrast, our proposed method does not rely on additional parameters in the most amenable cases. In the presence of a slowly decaying correlation function, compression, or signal noise, where it is more difficult for our proposed algorithm to obtain accurate estimates, our theoretical analysis provides guidelines to choose a proper value of the threshold level and guarantees that the proposed algorithm will have small estimation error. In terms of broad applicability, our proposed method can be used for time delay estimation, frequency estimation and other sparsity-based parameter estimation problems; additionally, the observation model is arbitrary as well, with example applications using subsampled signals or randomly projected measurements. This is in contrast with existing approaches that obtain state-of-the-art performance, which are specialized to a single problem and observation model.

1.3. Organization

This paper is organized as follows. In Section 2, we provide a summary of compressive parameter estimation and the issues in existing work and introduce some important background for the proposed algorithm. In Section 3, we present and analyze the use of EMD and clustering methods for PD-based parameter estimation; furthermore, we formulate and analyze an algorithm for EMD-optimal sparse recovery that employs K-median clustering. In Section 4, we present numerical simulations that verify our results

for the clustering method in the example applications of time delay estimation and frequency estimation. Finally, we provide a discussion and conclusions in Section 5.

2. Background

2.1. One-Dimensional Parameter Estimation

One-dimensional parameter estimation problems are usually defined in terms of a parametric signal class, which is defined via a mapping $\psi : \Theta \to X$ from the parameter space $\Theta \subset \mathbb{R}$ to the signal space $X \subset \mathbb{C}^N$. The signal observed in a parameter estimation problem contains K unknown parametric components $x = \sum_{i=1}^{K} c_i \psi(\theta_i)$, and the goal is to obtain estimates of the parameters $\widehat{\theta}_i$ from the signal x. For example, in time delay estimation, the parameter $\theta_i = \tau_i$ refers to the time delay and the parametric signal $\psi(\theta_i)$ corresponds to a known waveform $g(t)$ with time delay τ_i, i.e.,

$$\psi(\theta_i) = [g(-\tau_i), g(T - \tau_i), \ldots, g((N-1)T - \tau_i)]^T, \tag{1}$$

when T represents the sampling period. Similarly, in line spectral estimation the parameter $\theta_i = f_i$ is the frequency and the parametric signal corresponds to the complex exponential of the frequency f_i, i.e.,

$$\psi(\theta_i) = \mathbf{e}(f_i) = \left[1, e^{j 2\pi f_i \frac{1}{N}}, \ldots, e^{j 2\pi f_i \frac{N-1}{N}} \right]^T. \tag{2}$$

Due to the demands in storage, communication, and computation, for large-scale signal processing it is often desirable to work with a lower-dimensional observation y rather than the signal x itself. The most common observations $y = x_S$ contain only the signal samples at random times $S \subset \{1, 2, \ldots, N\}$ with $|S| = M$. Alternatively, the observations $y = \Phi x$ can also be obtained by compressing the signal with a dimension-reducing measurement matrix $\Phi \in \mathbb{R}^{M \times N}$.

Although we are not aware of convex optimization approaches for generic time delay estimation, it is possible to transform a time delay estimation problem into a frequency estimation problem when the signal observations are linear measurements of the signal [30, 31], which can then itself be solved using convex optimization [25–29]. Assume that G is a diagonal matrix whose diagonal entries are the discrete Fourier transform coefficients of the waveform samples $g = [g(0), g(T), \ldots, g((N-1)T)]^T$. Then X, the discrete Fourier transform of x, has the sparsity structure as $X = Gx'$, where $x' = \sum_{i=1}^{K} c_i \mathbf{e}(\tau_i)$ is a frequency-sparse

5

signal. Thus, the linear measurements can be expressed as

$$y = \Phi x = \Phi F^{-1} X = \Phi F^H G x'. \tag{3}$$

Therefore, the time delay estimation of x becomes the frequency estimation of x' with the new measurement matrix $\Phi' = \Phi F^H G$. In summary, to estimate the time delay indirectly via frequency estimation, the observations are required to be the linear measurements of the signals, and the diagonal matrix G should have nonzero diagonal entries so that the new measurements matrix Φ' satisfies the properties required by sparse recovery algorithms. We will see that obtaining such G is difficult in practice in Section 4.

2.2. Parametric Dictionaries

The parameter estimation methods based on PDs model the estimation problem in a generic form and can be used both in frequency estimation and time delay estimation directly. One can obtain a PD as a collection of samples from the parametric signal space

$$\Psi = \left[\psi\left(\bar{\theta}_1\right), \psi\left(\bar{\theta}_2\right), \ldots, \psi\left(\bar{\theta}_L\right) \right] \subseteq \psi(\Theta), \tag{4}$$

which corresponds to a set of samples from the parameter space $\bar{\Theta} = \left\{ \bar{\theta}_1, \bar{\theta}_2, \ldots, \bar{\theta}_L \right\} \subseteq \Theta$. In this way, the signal can be expressed as a linear combination of the PD elements $x = \Psi c$ when all the unknown parameters are contained in the sampling set of the parameter space, i.e., $\theta_i \in \bar{\Theta}$ for each $i = 1, \ldots, K$. Therefore, finding the unknown parameters reduces to finding the PD elements appearing in the signal representation or, equivalently, finding the nonzero entries or support of the sparse PD coefficient vector c for which we indeed have $y = \Phi \Psi c$. The search for the vector c can be performed using CS recovery.

PD-based compressive parameter estimation can be perfect only if the parameter sample set $\bar{\Theta}$ is dense and large enough to contain all of the unknown parameters $\{\theta_1, \theta_2, \ldots, \theta_K\}$. If this stringent case is not met for some unknown parameter θ_k, a denser sampling of the parameter space decreases the difference between the unknown parameter θ_k and the nearest parameter sample $\bar{\theta}_l$, so that we can better approximate the parametric signal $\psi(\theta_k)$ with the parametric signal $\psi\left(\bar{\theta}_l\right)$. However, highly dense sampling increases the similarity between adjacent PD elements and, by extension, the PD coherence [32], corresponding to the maximum normalized inner product of PD elements:

$$\mu(\Psi) = \max_{1 \leq i \neq j \leq L} \frac{\left| \langle \psi(\bar{\theta}_i), \psi(\bar{\theta}_j) \rangle \right|}{\left\| \psi(\bar{\theta}_i) \right\|_2 \left\| \psi(\bar{\theta}_j) \right\|_2}. \tag{5}$$

6

Additionally, denser sampling increases the difficulty of distinguishing between PD elements and severely hampers the performance of compressive parameter estimation [33, 34]. Prior work addressed such issues by using a coherence-inhibiting structured sparse approximation, where the resulting K nonzero entries of the coefficient vector correspond to PD elements that have sufficiently low coherence. This approach is known as band exclusion [7, 13, 16–18] and inhibits the highly coherent PD elements from appearing in signal representation simultaneously. Choosing a adequate value for the maximum allowed coherence ν that defines the restriction on the choice of PD elements is essential to successful performance: setting its value too large results in the selection of coherent PD elements, while setting its value too small tightens up requirements on the minimum separation of the parameters. Unfortunately, a guideline for the chosen of ν has not been fully established.

Another issue is that existing CS recovery algorithms commonly used in this setting can only guarantee stable recovery of the sparse PD coefficient vectors when the error is measured by the ℓ_2 norm; in other words, we can only predict that the estimated coefficient vector will be close to the true coefficient vector in Euclidean distance. Such a guarantee is linked to the core hard thresholding operator, which returns the optimal sparse approximation to the input vector with respect to the ℓ_2 norm. However, the guarantee provides control on the performance of parameter estimation only in the most demanding case of exact recovery, i.e., when the parameter estimation is perfect. Otherwise, such a recovery guarantee is meaningless for parameter estimation since the ℓ_2 norm cannot precisely measure the difference between the supports of the sparse PD coefficient vectors. For an illustrative example, consider a simple frequency estimation problem where the PD collects complex sinusoids at frequencies $\{f_1, f_2, \ldots, f_L\}$ and there is only one unknown frequency f_i ($i \in \{1, 2, \ldots, L\}$). In other words, the coefficient vector is $c = e_i$, where e_i denotes the canonical vector that is equal to 1 at its ith entries and 0 elsewhere. Although two estimated coefficient vectors $\widehat{c}_1 = e_j$ and $\widehat{c}_2 = e_k$ ($j, k \in \{1, 2, \ldots, L\}$ and $j \neq k$) have the same ℓ_2 distance to the true coefficient vector c, the frequency estimate f_j from $\widehat{c}_1$ and f_k from $\widehat{c}_2$ have unequal estimation error $|f_i - f_j|$ and $|f_i - f_k|$, respectively.

Alternatively, the earth mover's distance (EMD) has recently been used in CS to measure the distance between coefficient vectors in terms of the similarity between their supports [21, 22]. In particular, the EMD between two vectors with the same ℓ_1 norm optimizes the work of the flow (i.e., the amount of the flow and the distance of flow) applied to one vector in order to obtain the other vector. In our illustrative example, if the values of the frequency samples of the PD increase monotonically, the EMDs between the true coefficient vector c and the estimated coefficient vectors $\widehat{c}_1$ and $\widehat{c}_2$ are proportional to the frequency

errors $|f_i - f_j|$ and $|f_i - f_k|$, respectively. Based on the fact that the work of the flow between any two entries of a PD coefficient vector is proportional to the distance between the two corresponding parameters, it follows that the EMD between pairs of PD coefficient vectors efficiently measures the corresponding error of parameter estimation. We elaborate our study of this property in Section 3.1.

2.3. K-Median Clustering

Cluster analysis partitions a set of data points based on the similarity information between each pair of points, which is usually expressed in terms of a distance [35, 36]. Clustering is the task of partitioning a set of points into different groups in such a way that the points in the same group, which is called a cluster, are more similar to each other than to those in other groups. The greater the similarity within a group is and the greater that the difference among groups is, the better or more distinct the clustering is.

The goal of clustering L points $\{p_1, p_2, \ldots, p_L\}$ associated with positive weights $w_1, w_2, \ldots, w_L$ and pairwise distances $d(p_i, p_j)$ into K clusters is to find the K centroids $\{q_1, q_2, \ldots, q_K\}$ of the clusters such that each cluster contains all points that are closer to their centroid than to other centroids:

$$C_i = \{p_l : d(p_l, q_i) \leq d(p_l, q_j), i \neq j\}. \tag{6}$$

One can define a clustering cost as a weighted sum of the distances between each point and their corresponding centroid:

$$J = \sum_{i=1}^{K} \sum_{p_j \in C_i} w_j d(q_i, p_j). \tag{7}$$

Different choices of pairwise distances $d(p_i, p_j)$ can result in different procedures to obtain the centroid assignments that minimize the cost J [36]. If the squared Euclidean distance is used, i.e., $d(p_i, p_j) = \|p_i - p_j\|_2^2$, then each cluster's centroid will be the weighted mean of its elements, and so the clustering is called K-means clustering. If the Manhattan distance is used, i.e., $d(p_i, p_j) = \|p_i - p_j\|_1$, then each cluster's centroid will be the weighted median of its elements, and so the clustering is called K-median clustering. In the special case where $p_i \in \mathbb{R}$, i.e., all the points are along a line, and the absolute value is used as a distance, i.e., $d(p_i, p_j) = |p_i - p_j|$, the cost J is equal to

$$J = \sum_{i=1}^{K} \sum_{p_j \in C_i} w_j |q_i - p_j|. \tag{8}$$

Therefore, one can solve for the centroids by setting the subgradient of the measure function (8) with respect

to each centroid to zero; the result is

$$\sum_{p_j \in C_i} w_j \text{sign}(q_i - p_j) = 0, \tag{9}$$

for $i = 1, \ldots, K$, where $\text{sign}(x)$ returns the sign of x. Equation (9) illustrates that the resulting centroids are the weighted medians of the elements in each cluster, and implies that the points on the two sides of the centroids have maximally balanced weight. That is, for each $i = 1, \ldots, K$,

$$\sum_{j:p_j \in C_i, p_j \leq q_i} w_j \geq \sum_{j:p_j \in C_i, p_j > q_i} w_j,$$
$$\sum_{j:p_j \in C_i, p_j < q_i} w_j \leq \sum_{j:p_j \in C_i, p_j \geq q_i} w_j. \tag{10}$$

The main difference between K-median clustering and the more predominant K-means clustering is that the former minimizes a sum of pairwise Manhattan distances, while the latter minimizes a sum of squared Euclidean distances. This difference makes the K-median clustering more robust to noise and outliers since the mean of a cluster deviates from the center of the cluster when outliers are present, while the median stays close to the center and is less impacted by the outliers. As shown in the experiments of Section 4, K-median clustering performs better than K-means clustering in parameter estimation.

2.4. Earth Mover's Distance

The earth mover's distance (EMD) between two vectors $(c, \widehat{c})$ relies on the notion of mass assigned to each entry of the involved vectors, with the mass of each entry being equal to its magnitude. The goal of EMD is to compute the lowest transfer of mass among the entries of the first vector needed in order to match the entries of the second vector. We assume that the vectors c and $\widehat{c}$ are two K-sparse vectors with supports $S = \{s_1, s_2, \ldots, s_K\}$ and $\widehat{S} = \{\widehat{s}_1, \widehat{s}_2, \ldots, \widehat{s}_K\}$, respectively. EMD $(c, \widehat{c})$ represents the distance between the two vectors by finding the minimum sum of mass flows $f_{i,j}$ from entry c_{s_i} to entry $\widehat{c}_{\widehat{s}_j}$ multiplied by the distance $d_{i,j} = |s_i - \widehat{s}_j|$ that can be applied to the first vector c to yield the second vector $\widehat{c}$.[1] This is a

---

[1]The standard definition of EMD assumes that the two vectors have the same ℓ_1 norm. Nonetheless, one can add additional entries on both vectors to account for the norm mismatch [37]. More specifically, we denote these additional entries as c_{N+1} and $\widehat{c}_{N+1}$, i.e., both are at the end of the vectors. Their magnitudes are $c_{N+1} = c_{\min} + \max\{0, \|\widehat{c}\|_1 - \|c\|_1\}$ and $\widehat{c}_{N+1} = c_{\min} + \max\{0, \|c\|_1 - \|\widehat{c}\|_1\}$, where $c_{\min}$ is the minimum magnitude among all nonzero entries of c and $\widehat{c}$, and the flow distance between the new entries is $d_{N+1,N+1} = 0$. Furthermore, the flow distance between all other entries and the new ones are $d_{i,N+1} = d_{N+1,j} = N$ for both vectors so that the propagation of mass to these new nodes is discouraged and limited to the mismatch; note that the flow between the two new entries is $f_{N+1,N+1} = c_{\min}$.

typical linear programming problem that can be written as

$$\text{EMD}\,(c,\widehat{c}) = \quad \min_{f} \qquad \sum_{i,j} f_{i,j} d_{i,j}$$

$$\text{such that} \quad \sum_{j=1}^{K} f_{i,j} = |c_i|, \quad i = 1, 2, \ldots, K,$$

$$\sum_{i=1}^{K} f_{i,j} = |\widehat{c}_j|, \quad j = 1, 2, \ldots, K, \tag{11}$$

$$f_{i,j} \geq 0, \qquad i, j = 1, 2, \ldots K.$$

3. Clustering Methods for Parameter Estimation

In this section, we present our approach for parameter estimation based on K-median clustering. We begin by studying the relationship between the EMD of PD coefficient vectors and the error of parameter estimation in order to justify the use of EMD in parameter estimation. Next, we show that the K-median clustering can be used as EMD-optimal sparse approximation to replace the hard thresholding operator in standard CS recovery algorithms. Afterward, we provide a theoretical performance analysis that features an upper bound on the parameter estimation error relying on the correlation function of the PD elements. Finally, we analyze the effect of compression and noise on the performance of the proposed algorithm and describe the use of thresholding to mitigate the performance reduction.

3.1. Estimation Errors

We wish to obtain parameter estimates that minimize the total estimation error, i.e., the sum of the absolute differences $\left| \theta_k - \widehat{\theta}_k \right|$ between each unknown parameter θ_k and the corresponding estimated parameter $\widehat{\theta}_k$. Computing the estimation error between a set of K one-dimensional true parameters $\theta = \{\theta_1, \theta_2, \ldots, \theta_K\} \subset \mathbb{R}$ and a set of K one-dimensional estimates $\widehat{\theta} = \left\{ \widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_K \right\} \subset \mathbb{R}$ involves an assignment problem that minimizes the cost of assigning each true parameter to a parameter estimate. The resulting parameter estimation error can be obtained by solving the following linear program, which is

a relaxation of the formal integer program optimization [38, Corollary 19.2a]:

$$PEE(\theta, \widehat{\theta}) = \min_{g} \quad \sum_{i,j} g_{i,j} t_{i,j}$$

$$\text{such that} \quad \sum_{j=1}^{K} g_{i,j} = 1, \quad i = 1, 2, \ldots, K;$$

$$\sum_{i=1}^{K} g_{i,j} = 1, \quad j = 1, 2, \ldots, K; \tag{12}$$

$$g_{i,j} \geq 0, \qquad i, j = 1, 2, \ldots, K,$$

where $t_{i,j} = \left| \theta_i - \widehat{\theta}_j \right|$. The minimizer g for the optimization above encodes the matching (between the true parameter values and their estimates) that minimizes the total parameter estimation error,

When the sampling interval of the parameter space that generates by PD is constant and equal to Δ (so that $\bar{\theta}_i = \Delta \cdot (i-1) + \theta_1$, where i denotes the index for the corresponding PD element $\psi(\bar{\theta}_i)$), it is easy to see that for a pair of true and estimated coefficient vectors and the corresponding true and estimated parameter values, the computations of the parameter estimation error (12) and the EMD (11) are a match under the transformation $t_{ij} = \Delta \cdot d_{ij}$. This straightforward comparison demonstrates the close relationship between the EMD and parameter estimation error (PEE), which is formally stated in the following theorem and proven in Appendix A.

Theorem 1. *Assume that Δ is the sampling interval of the parameter space that generates the PD used for parameter estimation. If c and $\widehat{c}$ are two K-sparse PD coefficient vectors with supports corresponding to two sets of parameters θ and $\widehat{\theta}$, then the EMD between the two coefficient vectors provides an upper bound to the parameter estimation error between the two sets of parameters:*

$$PEE(\theta, \widehat{\theta}) \leq \frac{\Delta}{c_{\min}} EMD(c, \widehat{c}), \tag{13}$$

where $c_{\min}$ is the smallest component magnitude among the nonzero entries of c and $\widehat{c}$.

Theorem 1 clearly shows that the EMD between sparse PD coefficient vectors provides an upper bound of the parameter estimation error for the corresponding parameters, with a scaling factor proportional to the PD parameter sampling rate Δ. For a particular estimation problem where the PD sampling interval is fixed, Theorem 1 gives the intuition that, though minimizing the EMD of PD coefficient vectors doesn't always minimize the parameter estimation error, it does minimize its upper bound in (13) and is likely

11

to return a smaller estimation error than non-EMD based approaches, for which no parameter estimation performance guarantee exists. Thus, finding an EMD-optimal sparse approximation algorithm is the next important step for the proposed parameter estimation method.

It is worth noting from the proof of Theorem 1 that the relationship between EMD and parameter estimation error is met with equality when $c_{\min} = c_{\max}$, i.e., all component magnitudes have the same values. If the dynamic range of component magnitudes is defined as

$$r = \max_{i \neq j} \frac{|c_i|}{|c_j|}, \tag{14}$$

then r reflects how close to equality (13) can be. The dynamic range of component magnitudes is an important condition in parameter estimation, as we will show in the sequel.

3.2. EMD-Optimal Sparse Approximation

As we have just shown, EMD-optimal sparse approximation plays a crucial role in our proposed compressive parameter estimation approach to enable small estimation error. In a similar way that the hard thresholding operator return the best sparse approximation vector in terms of Euclidean distance, we present an EMD-optimal sparse approximation that will return a sparse approximation vector $\widehat{c} \in \mathbb{C}$ that is optimal (among all sparse vectors) in terms of the EMD $(\widehat{c}, v)$ for any input vector $v \in \mathbb{C}$. This approach makes it possible to integrate EMD into a CS framework and to formulate a new compressive parameter estimation algorithm by replacing the hard thresholding operation by the new EMD-optimal sparse approximation.

Assume that $I = \{1, 2, \ldots, L\}$ and $S = \{s_1, s_2, \ldots, s_K\}$ is a K-element subset of I. Consider the problem of finding the K-sparse vector $\widehat{c}$ with fixed support S that has the smallest EMD to an arbitrary vector v. The minimum flow work defined in the EMD is achieved if and only if the flow is active between each entry of the vector v and its nearest (in terms of index) nonzero entry s_i of the vector $\widehat{c}$. In other words, the nonzero entries of $\widehat{c}$ partition the entries of v into K different groups. Denote by V_i the set of the entry indices of v that are matched to the nonzero entry s_i of $\widehat{c}$; this set can be written as

$$V_i = \{l \in I : |l - s_i| \leq |l - s_j|, s_j \neq s_i, s_j \in S\}. \tag{15}$$

The EMD defined in (11) can then be written as

$$\text{EMD}(v, \widehat{c}) = \sum_{i=1}^{K} \sum_{j \in V_i} |v_j||j - s_i|. \tag{16}$$

12

It is important to note that (16) has the same formula as (8), which is the objective function used in K-median clustering. Thus, one can pose a K-median clustering problem to minimize the value of (16) over all possible supports S. To that end, define L points in a one-dimensional space with weights $|v_1|, |v_2|, \ldots, |v_L|$ and locations $1, 2, \ldots, L$, respectively. It is easy to see that if we denote the set of centroid positions obtained by performing K-median clustering for this problem as S, then the set S corresponds to the support of the K-sparse signal that is closest to the vector v when measured with the EMD. One can then simply compute the sets in (15) and define the EMD-optimal K-sparse approximation $\widehat{c}$ to the vector v as $\widehat{c}_{s_i} = \sum_{j \in V_i} v_j$ for $s_i \in S$, with all other entries equal to zero.

3.3. Correlation Function in PD-based Parameter Estimation

We follow the convention of greedy algorithms for sparse approximation and CS, where a proxy of the coefficient vector is obtained via the correlation of the observations with the PD elements, i.e., $v = \Psi^H x$, where Ψ^H denotes the Hermitian (conjugate transpose) of Ψ. The resulting proxy vector v can be expressed as a linear combination of shifted correlation functions. The magnitudes of the components of this linear combination will be proportional to the magnitudes of the corresponding entries of the sparse coefficient vector c.

The correlation value between the PD elements corresponding to parameters θ_1 and θ_2 is defined as

$$\lambda(\omega) = \lambda(\theta_2 - \theta_1) = |\langle \psi(\theta_2), \psi(\theta_1) \rangle| = \left| \psi^H(\theta_1) \psi(\theta_2) \right|, \tag{17}$$

where $\omega = \theta_1 - \theta_2$ measures the difference between parameters, and we assume for simplicity that the correlation depends only on the parameter difference ω. In many parameter estimation problems, such as frequency estimation and time delay estimation, the correlation function has bounded variation such that the cumulative correlation function, defined as $\Lambda(\theta) = \sum_{\omega \in \bar{\Theta}: \omega \le \theta} \lambda(\omega)$, is bounded. The cumulative correlation function is a nondecreasing function with infimum $\Lambda(-\infty) = 0$ and supremum $\Lambda(\infty) = \sum_{\omega \in \bar{\Theta}} \lambda(\omega)$.

As shown in Figure 1, the correlation function $\lambda(\omega)$ achieves its maximum when the parameter difference $\omega = 0$ and decreases as $|\omega|$ increases, finally vanishing when $|\omega| \to \infty$. In words, the larger the parameter difference is, the smaller the similarity of the corresponding PD elements is. Due to the even nature of the correlation function, i.e., $\lambda(\omega) = \lambda(-\omega)$, the cumulative correlation function is rotationally symmetric, i.e., $\Lambda(\theta) + \Lambda(-\theta) = 2\Lambda(0) = \Lambda(\infty)$, as shown in Figure 1. Both figures also indicate that the correlation function for time delay estimation decays much faster than that for frequency estimation, which is indicative of the increased difficulty for frequency estimation with respect to time delay estimation, as shown in the sequel.
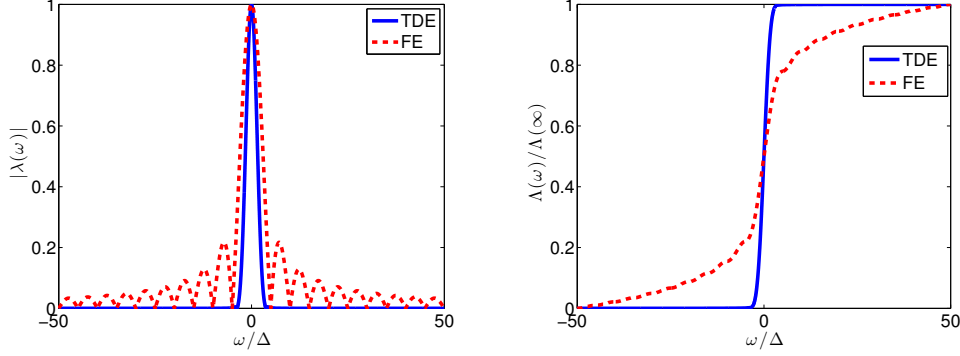
13

Figure 1: *Examples of correlation in PD constructions. (Left) Correlation function $|\lambda(\omega)|$ and (right) normalized cumulative correlation function $\Lambda(\omega)/\Lambda(\infty)$ as a function of the discretized parameter ω/Δ for time delay estimation (TDE) and frequency estimation (FE).*

For convenience of analysis, we assume that the correlation function is real and nonnegative, while noting that the experimental results match our theory even when the assumption does not hold. When the signal is measured directly without CS compression (i.e., the measurement matrix is the identity or $\Phi = I$) in a noiseless setting, the observations exactly match the sparse signal and can be written as

$$y = x = \sum_{i=1}^{K} c_i \psi(\theta_i). \tag{18}$$

Therefore, the proxy entries $v[j] = \psi(\bar{\theta}_j)^H x$ correspond to inner products between the observation vector y and the PD elements $\psi(\bar{\theta}_j)$ corresponding to the sampled parameters $\bar{\theta}_j \in \bar{\Theta}$. Thus, the proxy vector can be expressed as a linear combination of shifted correlation functions. For $j = 1, 2, \cdots, L$, the jth entry of the proxy vector is equal to

$$v[j] = \psi(\bar{\theta}_j)^H x = \sum_{i=1}^{K} c_i \psi(\bar{\theta}_j)^H \psi(\theta_i) = \sum_{i=1}^{K} c_i \lambda(\bar{\theta}_j - \theta_i). \tag{19}$$

It is easy to see that the proxy vector will feature local maxima at the indices for the sampled parameters in the PD that are closest to the true parameter. Thus, the goal of the parameter estimation finally reduces to the search of local maxima in the proxy vector over all parameter samples represented in the PD, which often is addressed via the sparse approximation operator on the proxy, e.g., via the hard thresholding operator or EMD-optimal sparse approximation.

When the correlation function has fast decay (usually the case when the coherence of the PD is very small), it is possible to find the local maxima of the proxy via the hard thresholding operator, as used in standard greedy algorithms for sparse signal recovery. However, when the correlation function decays slowly

Algorithm 1 *EMD-Optimal Sparse Parameter Estimator $\widehat{\theta} = \mathbb{C}(v, \bar{\Theta}, K)$*

Input: PD proxy vector v, set of parameter samples $\bar{\Theta}$, target sparsity K
Output: parameter estimates $\widehat{\theta}$, sampled indices S
1: **Initialize**: set $L = |\bar{\Theta}|$, choose S as a random K-element subset of $\{1, \ldots, L\}$
2: **repeat**
3:    $g_i = \arg\min_{j=1,\ldots,K} |i - s_j|$ for each $i = 1, \ldots, L$ {label parameter samples}
4:    $s_j = \text{median}\{(i, |v_i|) : g_i = j\}$ for each $j = 1, \ldots, K$ {update cluster medians}
5:    $\widehat{\theta} = \bar{\Theta}_S$ {find parameter estimates}
6: **until** S does not change or maximum number of iterations is reached

---

(usually the case when the PD elements are highly coherent), the thresholding operator will unavoidably focus its search around the peak of the proxy with the largest magnitude, unless additional approaches like band exclusion are implemented. In contrast, EMD-optimal sparse approximation identifies the local maxima of the proxy directly by exploiting the fact that these local maxima correspond to the K-median clustering centroids, when certain conditions (to be defined in Theorem 2 in the next subsection) are met. Thus, we can propose an EMD-optimal sparse approximation algorithm, shown as Algorithm 1, which leverages a standard iterative, Lloyd-style K-median clustering algorithm [35] and obtains the optimal sparse approximation of the proxy in the EMD sense. This algorithm will repeatedly assign each entry of the proxy vector to the cluster whose median is closest to the entry and then update each median by finding the weighted median of the cluster using the balance property (10). At the end, the parameter estimates corresponding to the centroids are returned.

3.4. Performance Analysis

There are some conditions that the signal x should satisfy to minimize the estimation error when using Algorithm 1.

- *Minimum Parameter Separation:* If two parameters θ_i and θ_j are too close to each other, the similarity of $\psi(\theta_i)$ and $\psi(\theta_j)$ makes it difficult to distinguish them. Therefore, our first condition considers the minimum separation distance:

$$\zeta = \min_{i \neq j} |\theta_i - \theta_j|. \tag{20}$$

- *Parameter Range:* It is often convenient to restrict the feasible parameters and sampled parameters in a small range, i.e, $\bar{\Theta}$ is bounded. Any parameter observed should be sufficiently far away from the bounds of the parameter range. According to (10), which implies a balance of the proxy v around the local maxima when using a clustering method, there will be a bias in the estimation due to the missed portion of the symmetric correlation function λ when the unknown parameter is too close to the bound

15

of the parameter range. Therefore, the condition should also consider the minimum off-bound distance ϵ, formally written as

$$\epsilon = \min_{1 \leq i \leq K} \left\{ \min \left(\theta_i - \min \left(\bar{\Theta} \right), \max \left(\bar{\Theta} \right) - \theta_i \right) \right\}. \tag{21}$$

- *Dynamic Range:* If the magnitudes of some components in the signal are too small, they may be dwarfed by larger components and ignored by the greedy algorithms. Thus, we need to pose an additional condition on the dynamic range of the component magnitudes as defined in (14).

With these conditions, we can formulate the following theorem that guarantees the performance of our clustering-based method for parameter estimation.

Theorem 2. *Assume that the sampling interval of the parameter space $\Delta \to 0$, and that the signal x given in (18) involving K parameters $\theta_1, \theta_2, \ldots, \theta_K$ has a dynamic range of the component magnitudes as defined in (14). For any allowed error $\sigma > 0$, if the minimum separation distance (20) satisfies*

$$\zeta \geq 2\Lambda^{-1} \left(2\Lambda(0) \left(1 - \frac{\Lambda(\sigma)/\Lambda(0) - 1}{(2K-2)r + 1} \right) \right) + 2\sigma, \tag{22}$$

and the minimum off-bound distance (21) satisfies

$$\epsilon \geq \Lambda^{-1} \left(2\Lambda(0) \left(1 - \frac{\Lambda(\sigma)/\Lambda(0) - 1}{2Kr} \right) \right), \tag{23}$$

then the estimates $\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_K$ returned from Algorithm 1 have estimation error $PEE\left(\theta, \widehat{\theta}\right) \leq K\sigma$.

The full proof of Theorem 2 is provided in Appendix B. Theorem 2 provides a guarantee on the parameter estimation error obtained from K-median clustering of the proxy v for the PD coefficients. Theorem 2 also provides a connection between the conditions described earlier and the performance of PD-based parameter estimation, providing a guide for the design and evaluation of PDs that can achieve the required estimation error for practical problems. For example, in time delay estimation problems, instead of increasing the minimum separation distance required for recovery, one can try to design a transmitted waveform that improves the other conditions cited above (such as one that increases the rate of decay of the cumulative correlation functions, as will be discussed in the sequel) to achieve small estimation error.

Theorem 2 also makes explicit the linear relationship between the normalized cumulative correlation for the minimum separation distance $\Lambda(\zeta/2)/\Lambda(\infty)$ and the normalized cumulative correlation for the maximum observed error $\Lambda(\sigma)/\Lambda(0)$, which can be obtained by applying $\Lambda(\cdot)$ on both sides of (22) while ignoring the

small term σ. Additionally, the required minimum separation distance will be dependent on the specific parameter estimation problem, even when the maximum allowed error is kept constant. This illustrates the wide difference in performances in parameter estimation, e.g., between time delay estimation and frequency estimation: the minimum separation distance required by time delay estimation is much smaller than that of frequency estimation due to the contrasting rates of decay of the function Λ, as shown in Figure 1. Once more, we emphasize that although Theorem 2 is derived for the case of a nonnegative real correlation function and is asymptotic on the parameter sampling interval Δ, our numerical simulations in the sequel show that the predicted relationship between ζ and σ are observed in practical problems of modest sizes.

3.5. Effect of Compression

The addition of CS and measurement noise make the estimation problem more difficult, since in both of these cases there is a decrease in the rate of decay of the cumulative correlation function Λ, as shown in the sequel. When the measurement matrix Φ is used to obtain the observed measurements y from the signal x such that $y = \Phi x = \sum_{i=1}^{K} c_i \Phi \psi(\theta_i)$, the proxy becomes

$$v[j] = \langle \Phi^H y, \psi(\bar{\theta}_j) \rangle = \langle y, \Phi \psi(\bar{\theta}_j) \rangle = \sum_{i=1}^{K} c_i \langle \Phi \psi(\theta_i), \Phi \psi(\bar{\theta}_j) \rangle. \tag{24}$$

Only if $\Phi^H \Phi = I$ can (24) be identical to (19). We define the compressed correlation function as

$$\lambda_\Phi(\omega) = \lambda_\Phi(\theta_1 - \theta_2) = \langle \Phi \psi(\theta_2), \Phi \psi(\theta_1) \rangle = \psi^H(\theta_1) \Phi^H \Phi \psi(\theta_2). \tag{25}$$

The proxy can again be expressed as the linear combination of shifted copies of the redefined correlation function (25):

$$v[j] = v(\bar{\theta}_j) = \sum_{i=1}^{K} c_i \lambda_\Phi(\bar{\theta}_j - \theta_i). \tag{26}$$

Although in general we will have $\lambda \neq \lambda_\Phi$, we can use the preservation property of inner products through random projections [39]. That is, when Φ has independent and identically distributed (i.i.d.) Gaussian random entries and sufficiently many rows, there exists a constant $\delta > 0$ such that for all pairs (θ_i, θ_j) of interest we have

$$(1 - \delta)\lambda(\theta_i - \theta_j) \leq \lambda_\Phi(\theta_i - \theta_j) \leq (1 + \delta)\lambda(\theta_i - \theta_j). \tag{27}$$

The parameter δ decays as the compression rate $\kappa = M/N$ increases, and the manifolds $\lambda(\cdot)$ we consider here are known to be amenable to large amounts of compression [40]. Such a relationship indicates that the

compression can affect the correlation function and, by extension, the performance of clustering methods for compressive parameter estimation.

3.6. Quantifying the Role of Correlation Decay

We choose to focus on simple bounds for the correlation function λ to analyze its role in the performance of EMD and PD-based parameter estimation. Similarly to [41], we use bounding functions to measure and control the decay rate of the correlation function λ. We approximate the correlation function $\lambda_\Phi(\omega)$ with an exponential function $\bar{\lambda}(\omega) = \exp(-a|\omega|)$ that provides an upper bound of the actual correlation function, i.e., $\lambda_\Phi(\omega) \leq \bar{\lambda}(\omega)$. The performance obtained from the exponential function approximation $\bar{\lambda}(\cdot)$ provides an upper bound of the performance from the real correlation function $\lambda(\cdot)$. In the exponential function, a is the parameter that controls the decay rate: the larger a is, the faster that the correlation function decays. The decay rate of the compressed correlation function (25) will be smaller than that of the original correlation function (17). We assume that $\lambda(\omega) = \exp(-a|\omega|)$ and that a bound $\lambda_\Phi(\omega) \leq \exp(-b|\omega|)$ exists; in this case, $b < a$ due to the fact that (27) provides us with the following upper bound:

$$\lambda_\Phi(\omega) \leq (1+\delta)\lambda(\omega) \leq (1+\delta)\exp(-a|\omega|) \leq \exp\left(-\left(a - \frac{\ln(1+\delta)}{|\omega|}\right)|\omega|\right), \tag{28}$$

where $\ln(1+\delta)/|\omega| > 0$ when $\delta > 0$. This shows that CS reduces the decay speed of the correlation function and increases the necessary minimum separation distance and minimum off-bound distance to guarantee the preservation of parameter estimation performance. This dependence is also manifested in the experimental results of Section 4 when the correlation function does not follow an exact exponential decay.

We observe in practice that the issues with slow-decaying correlation functions arise whenever the sum of the copies of the correlation functions far from their peaks becomes comparable to the peak of any given copy. Thus, one can use operators such as thresholding functions to remove this effect from appearing in Algorithm 1. We can write a hard-thresholded version of the proxy from (26) as

$$v_t(\theta_i) = \begin{cases} v(\theta_i) & |v(\theta_i)| > t, \\ 0 & v(\theta_i)| \leq t, \end{cases} \tag{29}$$

where t is the value of the thresholding. The following theorem is proven in Appendix C and focuses on the minimum separation ζ by assuming that ϵ is sufficiently large. The theorem demonstrates that the thresholding operator reduces the required minimum separation distance for accurate estimation.

Theorem 3. *Under the setup of Theorem 2, assume that the correlation function defined in (25) is given by* $\lambda_\Phi(\omega) = \exp(-a|\omega|)$. *For any allowed error $\sigma > 0$, when the minimum off-bound distance ϵ is large enough, if the dynamic range given in (14) is equal to r, the threshold given in (29) satisfies $rc_{\min}\exp(-a\sigma) \leq t \leq c_{\min}$, and the minimum separation distance given in (20) satisfies*

$$\zeta \geq \frac{1}{a}\ln\left(\sqrt{\frac{8r^2}{t^2/(rc_{\min})^2 - \exp(-2a\sigma)} + 1}\right) \tag{30}$$

where $c_{\min}$ is the minimum component magnitude, then the estimates $\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_K$ returned from performing Algorithm 1 on the thresholded proxy v_t in (29) have estimation error $PEE\left(\theta, \widehat{\theta}\right) \leq K\sigma$.

Theorem 3 extends Theorem 2 by including the use of thresholding as a tool to combat the slow decay of the correlation function, due to an ill-posed estimation problem or the use of CS. One can also instinctively see that the presence of noise in the measurements will also slow the decay of the correlation function, which according to the theorem will require larger minimum separation or careful thresholding. In practice, the decay coefficient a can usually be obtained by finding the minimum value such that the exponential function $\exp(-a|\omega|)$ provides a tight upper bound for the correlation function $\lambda(\omega)$.

Perhaps the most important application of Theorem 3 focuses on the choice of threshold t for the particular problem of interest, which can improve the performance of the clustering method in compressive parameter estimation. To deal with the problems of slow decay or large dynamic range, one can try to increase the threshold value on the proxy rather than increasing the minimum separation to improve the estimation performance. Again, although Theorem 3 is based on an approximation of the actual compressive parameter estimation problem setup, our numerical results in the sequel show its validation in practical settings for time delay estimation and frequency estimation.

It is worth noting the similarities between our analysis above and that performed in [41]. Both analyses rely on the assumption that the cumulative correlation function is symmetric, nondecreasing and variation-bounded. Additionally, both analyses use an upper bound of the correlation function to simplify the analysis. Furthermore, both analyses derive the requirement that the separation between any two parameters needs to be large enough. The main difference between the analyses is that the guarantees of [41] are focused on exact parameter estimation, while our analysis considers the quality of parameter approximation when the problem setup does not provide exact estimation.

4. Numerical Experiments

In order to test the performance of the clustering parameter estimation method on different problems, we present a number of numerical simulations involving time delay estimation and frequency estimation.[2] Before detailing our experimental setups, we define the parametric signals and the PDs involved in these two example applications.

The parametric signal model for time delay estimation uses a chirp signal $g(t) = \exp\left(j2\pi\left(f_c + f_a\frac{t}{T}\right)t\right)p(t)$, where

$$
p(t) = \begin{cases} \sqrt{\dfrac{2}{3Tf_s}}\left(1 + \cos\left(2\pi\dfrac{t}{T}\right)\right), & t \in [0, T], \\ 0, & \text{otherwise.} \end{cases}
\tag{31}
$$

The length of the chirp and the chirp's starting frequency are fixed to be $T = 1$ μs and $f_c = 1$ MHz across all experiments. The chirp's frequency sweep f_a will have different values in different experiments. The sampling frequency of the chirp signal is $f_s = \frac{1}{T_s}$, and N samples are taken for each signal. The delay parameter space range goes from $\theta_{\min} = 0$ to $\theta_{\max} = NT_s$. The PD for time delay estimation contains all chirp signals corresponding to the parameter space sampled at a resolution of Δ, e.g.,

$$
\Psi = [\psi(0), \psi(\Delta), \ldots, \psi(\theta_{\max})],
\tag{32}
$$

with entries

$$
\psi(\theta)[n] = g(nT_s - \theta), n = 0, 1, \ldots, N - 1.
\tag{33}
$$

For frequency estimation, the parametric signals are the N-dimensional complex exponentials with entries

$$
\psi(\theta)[n] = \frac{\exp\left(j2\pi\theta\frac{n}{N}\right)}{\sqrt{N}}, \quad n = 0, 1, \ldots, N - 1.
\tag{34}
$$

The parameter space range goes from $\theta_{\min} = 0$ to $\theta_{\max} = N$ Hz. As before, the PD for frequency estimation is described in (32) with sampling interval Δ. In both time delay estimation and frequency estimation, the number of unknown parameters is set to $K = 4$.

---

[2]Additional experiments are available in an early version of this manuscript [42].

4.1. Theorem Validation

In the first experiment, we illustrate the relationship between the minimum separation distance and the maximum allowable error described in Theorem 2. We measure the performance for each minimum separation distance ζ by the maximum estimation error σ over 1000 signals with randomly chosen parameter values that are spaced by at least ζ. For time delay estimation, $f_a = 20$ MHz, $f_s = 50$ MHz, $N = 500$, and $\Delta = 0.001$ μs so that the PD contains observations for 10001 parameter samples, and we let the minimum separation $\zeta \in [0.07~\mu s, 0.1~\mu s]$. For frequency estimation, $N = 500$ and $\Delta = 0.05$ Hz, so that the PD contains observations for 10001 parameter samples, and we let the minimum separation $\zeta \in [35~\text{Hz}, 75~\text{Hz}]$.

Figure 2 shows the normalized cumulative correlation for the maximum error $\Lambda(\sigma)/\Lambda(0)$ as a function of the normalized cumulative correlation for the minimum separation distance $\Lambda(\zeta/2)/\Lambda(\infty)$ for both time delay estimation and frequency estimation. The figure also shows the relationship between the minimum separation ζ and the maximum error σ without the use of the cumulative correlation function for both example cases. The approximately linear relationship between $\Lambda(\zeta)/\Lambda(\infty)$ and $\Lambda(\sigma)/\Lambda(0)$ for both the time delay estimation and frequency estimation cases numerically verifies the result of Theorem 2. The difference between the performance results, as well as the relationship between ζ and σ, validates the conclusion that frequency estimation requires a significantly larger minimum separation than time delay estimation. We also see from Figure 2 that it is impossible to get an arbitrarily small estimation error even if the minimum separation keeps increasing, as the estimation error cannot be smaller than the parameter sampling step Δ (as observed in the figure). In fact, the figure shows that the relationship between $\Lambda(\zeta/2)/\Lambda(\infty)$ and $\Lambda(\sigma)/\Lambda(0)$ ends its linearity exactly when the error approach the parameter sampling resolution Δ for both application examples. To achieve precision above the resolution Δ, the use of additional methods such as interpolation is needed [43]. Nonetheless, though the function $\Lambda(\cdot)$ varies in different parameter estimation problems, we consistently obtain a linear relationship between the minimum separation distance and the parameter estimation error for several parameter estimation problems by employing the function $\Lambda(\cdot)$.

In the second experiment, we illustrate the application of Theorem 3 in the time delay estimation problem. While the values of all other parameters are kept the same as in the first experiment, we set $\Delta = 0.002$ μs and vary the chirp frequency sweep f_a between 2 Hz and 20 Hz to generate different rates of decay of the correlation function. For each chirp frequency sweep, we obtain the decay parameter a as the smallest value that enables the exponential function $\exp(-a|\omega|)$ to bound the correlation function $\lambda(\omega)$. We then determine the minimum separation ζ for which a maximum estimation error $\sigma = 0.04$ μs is obtained for Algorithm 1 over 1000 randomly drawn signals having the given minimum separation ζ. The first result
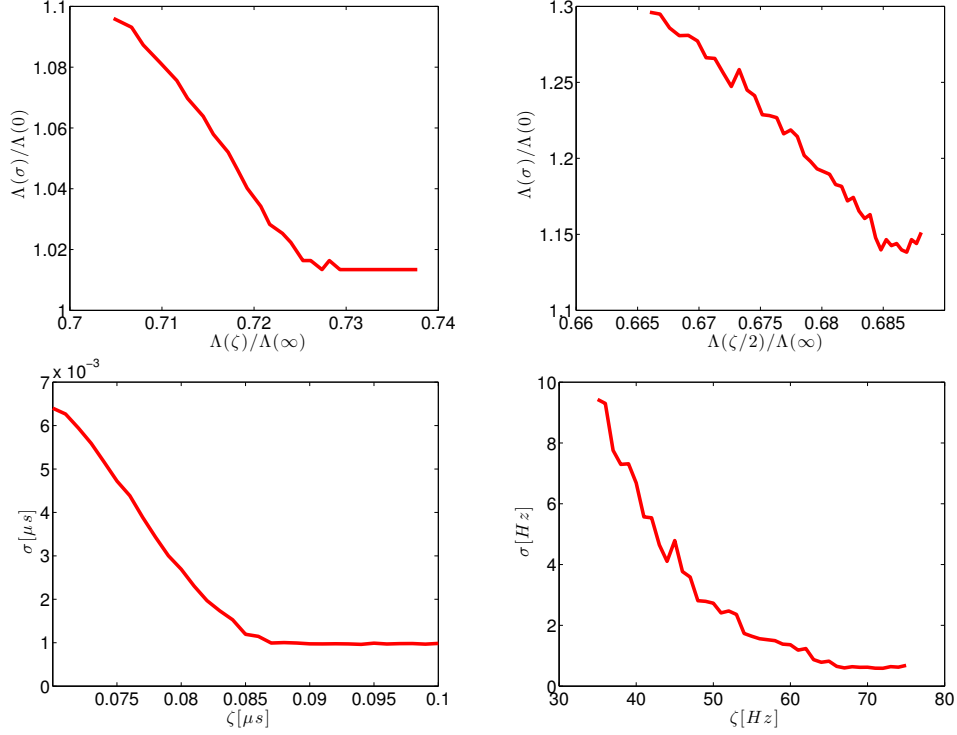
Figure 2: *Performance results for two parameter estimation problems as a function of the parameter separation. Left: TDE; right: FE. Top: Normalized cumulative correlation function for the maximum estimation error $\Lambda(\sigma)/\Lambda(0)$ as a function of the normalized cumulative correlation function for the minimum separation $\Lambda(\zeta/2)/\Lambda(\infty)$, showing a linear dependence. Bottom: Maximum error σ as a function of Minimum separation ζ.*

in Figure 3 shows the approximate reciprocal relationship between ζ and a, which is very close to the line $\zeta = \ln\left(\sqrt{8}/0.2 + 1\right)/a = 2.71/a$ predicted by Theorem 3, when the threshold level $t = 0.2$ and dynamic range $r = 1$. The second result shows the approximate logarithmic relationship between the minimum separation ζ and the dynamic range r, when we set the decay parameter $a = 8.28$ and the threshold $t = 0.9$. The minimum separation observed is smaller than the prediction $\zeta = \ln\left(\sqrt{8}r^2/0.9 + 1\right)/8.28 = \ln\left(3.1r^2 + 1\right)/8.28$ given by Theorem 3, which is likely explained by the difference between the correlation function and its exponential upper bound. The third result in Figure 3 shows that required threshold for accuracy estimation when $a = 12.67$ and $r = 1$ is slightly less than the value $t = \exp\left(-12.67 \cdot 0.04\right) = 0.60$ predicted by the theorem. The results presented above can be interpreted as a numerical corroboration of Theorem 3.

4.2. Performance Comparison

Our following experiments test the performance of clustering methods in compressive parameter estimation for different settings. We incorporate K-median clustering into subspace pursuit, a standard sparse recovery algorithm introduced in [44], by replacing each instance of the hard thresholding operator in subspace pursuit with an instance of Algorithm 1. The resulting clustering subspace pursuit (CSP) is shown
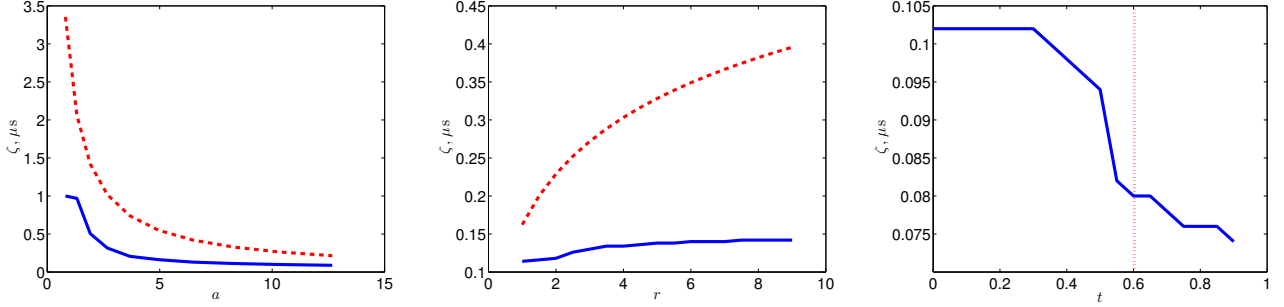
Figure 3: *Performance results for time delay estimation. The plots show the minimum separation necessary for accurate estimation as a function of (left) decay coefficient a, (middle) dynamic range r, and (right) threshold value t. Blue solid curves denote experimental results, while red dashed curves denote the bounds given by Theorem 3.*

Algorithm 2 *Clustering Subspace Pursuit (CSP)*

Input: measurement vector y, measurement matrix Φ, sparsity K, set of sampled parameters $\bar{\Theta}$, threshold t

Output: estimated signal $\widehat{x}$, estimated parameter values $\widehat{\theta}$

  1: **Initialize:** $\widehat{x} = 0$, $S = \emptyset$, generate PD Ψ from $\bar{\Theta}$.

  2: **repeat**

  3:     $y_r = y - \Phi\widehat{x}$ {Compute residual}

  4:     $v = (\Phi\Psi)^H y_r$ {Obtain proxy from residual}

  5:     $v(|v| \leq t) = 0$ {Threshold proxy}

  6:     $S = S \cup \mathbb{C}(v, \bar{\Theta}, K)$ {Augment parameter estimates from proxy}

  7:     $c = (\Phi\Psi_S)^+ y$ {Obtain proxy on parameter estimates}

  8:     $S = \mathbb{C}(c, \bar{\Theta}_S, K)$ {Refine parameter estimates}

  9:     $\widehat{x} = \Psi_S c_S$ {Assemble signal estimate}

10:     $\widehat{\theta} = \bar{\Theta}_S$ {Assemble parameter estimates}

11: **until** a convergence criterion is met

---

in Algorithm 2, where $\mathbb{C}$ refers to Algorithm 1 and M^+ denotes the Moore-Penrose pseudoinverse of M. Similarly to the band-exclusion subspace pursuit (BSP) used in [13, 16, 17], CSP repeatedly computes the proxy for the coefficient vectors from the measurement residual and then obtains the indices of the potential parameter estimates from the thresholded proxy using Algorithm 1. A second clustering refines the estimation after the potential estimates are merged with the previous estimates in order to maintain a final set of K estimates at the end of each iteration. CSP can also be armed with polar interpolation to significantly improve the estimation precision in a manner similar to band-excluded interpolating subspace pursuit (BISP) [1, 43]; we call the resulting algorithm clustering interpolating subspace pursuit (CISP).

Again, in Algorithm 2, we use the K-median clustering rather than K-means clustering because the criteria of K-median clustering to minimize Manhattan distance matches the definition of EMD and is more robust to the presence of outliers, which can be caused by a slowly decaying correlation function, compression, or noise. Figure 4 shows the average time delay estimation error of two versions of Algorithm 2: one that
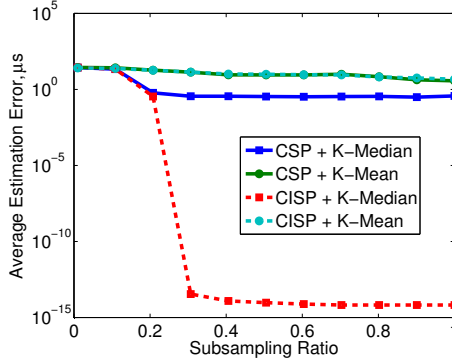
Figure 4: *Average time delay estimation error of Algorithm 2 with K-median clustering and K-means clustering.*

uses K-median clustering and another one that uses K-means clustering. It is clear that K-means clustering is not able to obtain the same estimation quality as K-median clustering.

Besides BSP and CSP, we also evaluate the performance on approximate message passing with MUSIC (AMP) [45, 46] and low-rank Hankel matrix reconstruction (HMR) [29] to estimate time delays or frequencies from the linear measurements. For estimation from subsampled observations, we also consider algorithms such as atomic soft thresholding (AST) for atomic norm minimization [25–27, 41] and enhanced matrix completion (EMC) [28]. All convex optimization methods (AST, EMC, HMR) are implemented using the CVX package [47], and the output time-domain signals are then input to the root MUSIC algorithm for line spectral estimation. For time delay estimation, these algorithms are modified according to (3) to perform indirect time delay estimation. In the following experiments, the signal length is set for both applications to $N = 101$ due to the computational burden of the optimization-based methods. Table 1 illustrates the difference between the algorithms considered in these experiments.

Time Delay Estimation: Our first experiment tests the estimation performance of the algorithms listed in this section on 1000 independent randomly-generated time delay estimation problems. The chirp signal has frequency sweep $f_a = 4$ MHz and sampling frequency $f_s = 10$ MHz. Thus, the duration of the whole chirp signal is $NT_s = 10.1$ μs. The $K = 4$ unknown time delays are generated randomly such that the minimum separation $\zeta = 0.05$ μs. The parametric dictionary used for CSP, CISP, BSP and BISP contains the signal observations for the sampling set with sampling step $\Delta = 0.01$ μs. The threshold value for CSP and CISP is set to $t = 0$ (i.e., no thresholding takes place). The maximum allowed coherence for BSP and BISP is chosen as $\mu = 0.01$, which was found to give the best estimation performance over a grid search. Figure 5 shows the average estimation error as a function of the compression rate $\kappa = M/N$ when noiseless linear measurements and subsampled signals are observed. The linear measurements are obtained

Table 1: *Comparison of algorithms for sparsity-based parameter estimation*

Algorithm	Estimation Problem	Observation Model	Performance Conditions	Guarantee
AST [25–27, 41]	line spectrum estimation	subsampled signals	minimum separation, fast decay in correlation function	exact parameter estimation
BSP/BISP [13, 16, 17]	arbitrary, with optional interpolation	arbitrary	maximum allowed coherence	exact signal recovery
EMC [28]	line spectrum estimation	subsampled signals	minimum separation	exact signal recovery
HMR [29]	line spectrum estimation, NMR spectroscopy estimation	linear measurements	Gaussian measurement matrix	exact signal recovery
AMP [45, 46]	arbitrary with statistical estimation from noisy observations	linear measurements	Gaussian measurement matrix, sufficient denoiser performance	recovery prediction via state evolution
CSP/CISP	arbitrary, with optional interpolation	arbitrary	minimum separation, fast decay in correlation function	approximated parameter estimation

via a measurement matrix with entries drawn i.i.d. from the standard Gaussian distribution. The results indicate that CSP can estimate the time delays with error below the sampling step δ when $\kappa \geq 0.3$ for both types of observations, which matches the performance of its band-exclusion counterparts. However, while BSP requires setting the maximum allowed coherence μ carefully via grid search, CSP can achieve similar performance without relying on thresholding. Due to the limitation of discretization, both CSP and BSP perform worse than HMR, which uses convex optimization to estimate off-grid time delays precisely. Armed with polar interpolation, both BISP and CISP can easily reach accurate estimation. We believe that the performance of HMR can increase with the use of more precise convex optimization solvers; however, this would also increase the estimation complexity. The poor performance of AMP is likely due to the fact that the measurement matrix derived from (3) does not allow for the signal estimate to be modeled statistically as the true signal with additive white Gaussian noise [45]. Additionally, AST and EMC cannot estimate the time delay from subsampled signals since the process of changing the time delay estimation to frequency estimation is possible only for linear measurements.

When the duration of the chirp signal is decreased, the Gram matrix G in (3) converges towards a scaled identity matrix. Thus, the measurement matrix for the resulting frequency estimation problem more closely matches that of the original delay estimation problem. Figure 6 presents the average estimation error for
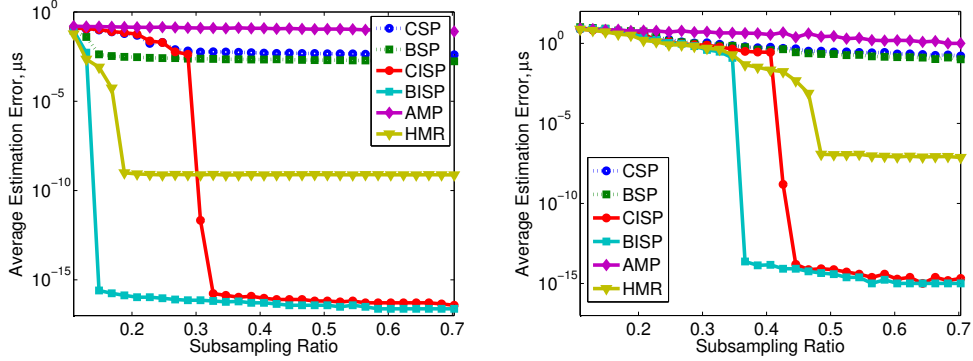
Figure 5: *Performance for the time delay estimation problem, as measured by the average parameter error, when (left) linear measurements and (right) subsampled signals are observed.*
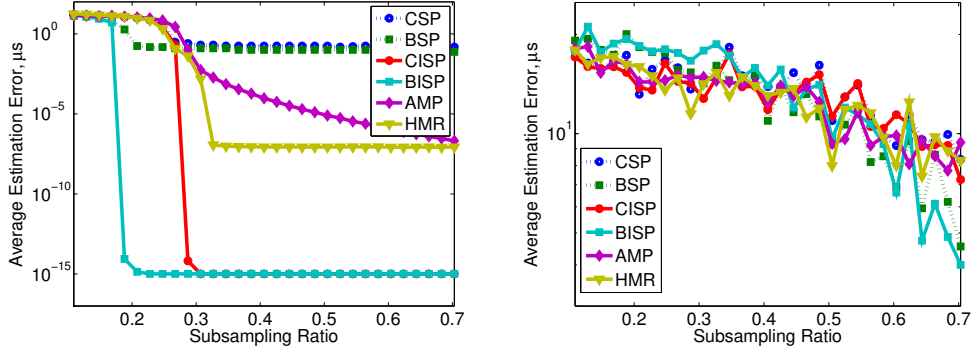

Figure 6: *Performance for the impulse delay estimation problem, as measured by the average parameter error, when (left) linear measurements and (right) subsampled signals are observed.*

an impulse delay estimation as the function of the compression rate κ. The good performance of AMP in estimation from linear measurements demonstrates that the necessary properties of measurement matrix remain. It is easy to see that when the observations correspond to a subsampled signal, it will be very difficult to estimate the impulse time delays.

Frequency Estimation: In our next experiment, we repeat the first experiment on frequency estimation with 1000 independent randomly drawn signals. The $K = 4$ unknown frequencies have minimum separation $\zeta = 0.1$Hz. The parametric dictionary used for CSP, CISP, BSP and BISP contains the signal observations for the sampling set with sampling step $\Delta = 0.1$Hz. The threshold value for CSP and CISP and the maximum allowed coherence for BSP and BISP are required to be $t = 0.4$ and $\mu = 0.1$ to have good performance, according to Theorem 3 and a grid search, respectively.

Similarly as in the first experiment, Figure 7 shows the average estimation error as a function of the compression rate and the SNR. The conclusion is similar: armed with polar interpolation, CISP achieves accurate estimation when compression is high enough and CISP shows similar noise robustness as other
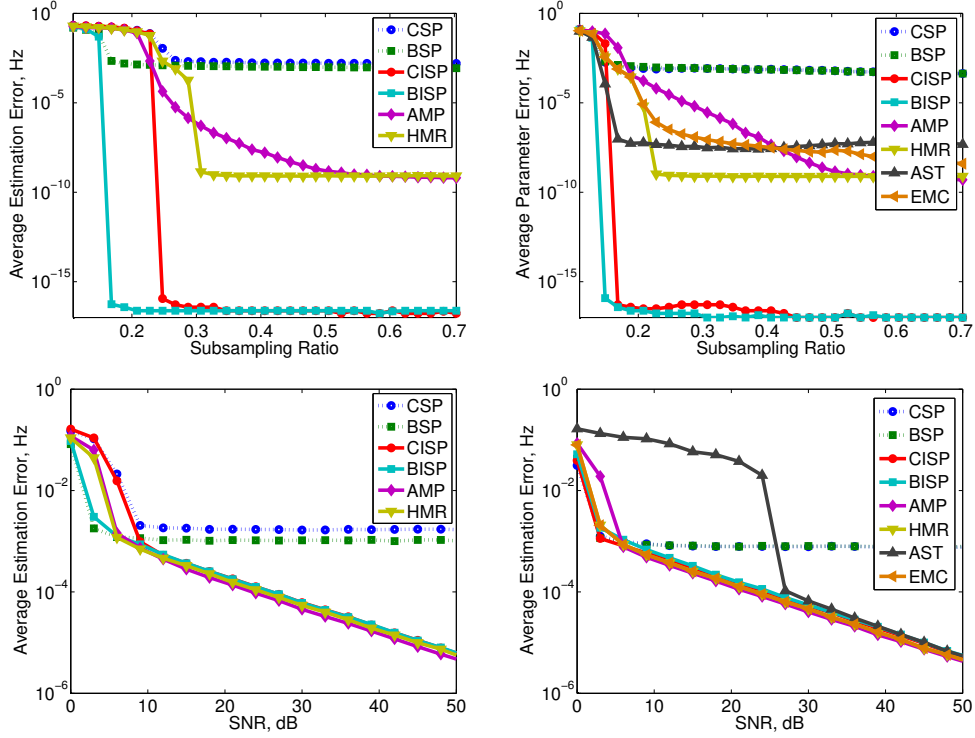
Figure 7: *Performance of compressive parameter estimation for the frequency estimation, as measured by the average parameter error, as a function of (top) the CS compression rate $\kappa = M/N$ with noiseless measurements and (bottom) the measurement SNR with $\kappa = 0.4$, under (left) linear measurements and (right) subsampled observations.*

algorithms; we found that the trends shown in the figure extend to SNRs of at least 90 dB. Though CSP needs the thresholding value to estimate the frequencies correctly, we can always choose a proper value according to Theorem 3.

In the last experiment, we apply our proposed clustering-based algorithm for compressive parameter estimation with a real-world signal. The lynx signal has $N = 114$ samples and is used to test the performance of line spectral estimation algorithms in [48]. It is well approximated by a sum of complex sinusoids with small minimum separation distance and large dynamic range among the component magnitudes. We increase the size of the signal by a factor of 10 using interpolation and obtain CS measurements of the resulting signal with a random matrix for various compression rates under several levels of measurement noise. The maximum allowed coherence in BISP is set to $\nu = 0.2$ after a grid search to optimize the algorithm's performance, while the threshold level in CISP is set to $t = 0$. Figure 8 shows the average relative estimation error between the estimates from CISP and BISP and those obtained from root MUSIC (a line spectral estimation algorithm with high accuracy [48]) when applied to the full length signal. The results show that CISP without thresholding has closer performance to root MUSIC than the best configuration of BISP.
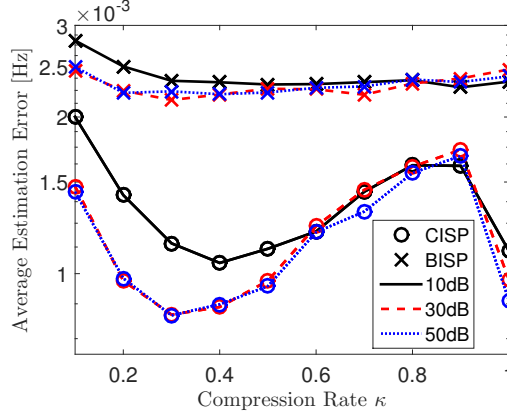
27

Figure 8: *Performance of compressive parameter estimation algorithm for the real FE problem. Dash-dot, dashed, dotted lines represent the average relative estimation errors when SNR = 10, 30, 50 dB, and lines with circle and cross are errors of CISP and BISP.*

5. Discussion and Conclusions

In this paper, we have introduced and analyzed the relationship between the EMD applied to PD coefficient vectors and the parameter estimation error obtained from sparse approximation methods applied to the PD representation. We also leveraged the relationship between EMD-based sparse approximation and K-median clustering algorithms in the design of new compressive parameter estimation algorithms. Based on the relationship between the EMD and the parameter estimation error, we have analytically shown that the EMD between PD coefficient vectors provides an upper bound of the parameter estimation error obtained from methods that use PDs and EMD. Furthermore, we leveraged the connection between EMD-sparse approximation and K-median clustering, to formulate new algorithms that employ sparse approximation in terms of the EMD; we then derived three theoretical results that provide performance guarantees for EMD-based parameter estimation under certain requirements for the signals observed, in contrast to existing work that does not provide such guarantees. Our experimental results show the validation of our analysis in several practical settings, and provides methods to control the effect of coherence, compression, and noise in the performance of compressive parameter estimation.

In our new compressive parameter estimation algorithms, we use K-median clustering rather than the more predominant K-means clustering to obtain the sparse approximation or the local maxima of the proxy. The main difference between K-median clustering and K-means clustering is that their criteria in (7) are different: the former uses the Manhattan distance and the latter uses the squared Euclidean distance. This difference makes K-median clustering more robust to noise and outliers and prevents K-means clustering from being able to return the EMD-optimal sparse approximation in general.

Our experiments have shown that our clustering-based algorithms for compressive parameter estima-

tion can achieve the same performance as those based on band exclusion. Though both methods use additional parameters to improve the performance, the clustering method is preferable as it does not need to rely on parameter tuning under a noiseless setting and for simpler problems, while the band exclusion is highly dependent on the allowed coherence in all cases. The threshold level is needed only in cases where the estimation problem is particularly ill-posed, CS has been heavily used, measurement noise is present, or in other cases where the correlation function decays slowly. As shown in Figure 5, the clustering method without thresholding has the same performance as the band-exclusion method with optimally tuned maximum coherence. Interested readers can refer to [1], where we have further studied the different sensitivities of these two methods on the additional parameters.

The results we obtained show the very different performance level of convex optimization-based methods on time delay estimation and frequency estimation, which is due to the fact that all convex methods can not be applied on time delay estimation directly. In contrast, the advantage of PD-based parameter estimation is that it can be utilized on almost all parameter estimation problems, as they only require that a PD can be obtained to yield sparse representation for the relevant signals. Additionally, the convex methods depend on convex optimization solver to recover signals, which have high computation cost. In the second part of numerical experiments, we have to set the length of the signal to be a small value otherwise the CVX package is not able to solve the proposed optimization problem. Instead, the essential part to the PD-based parameter estimation is the sparse approximation operator used in the algorithm, which usually can be much simpler. Furthermore, as shown in Table 1, the observation model of PD-based parameter estimation can be arbitrary: the observations can be the linear measured or subsampled signals, while existing approaches focus on a single observation model.

Acknowledgements

Appendix A. Proof of Theorem 1

Assume that the vectors c and $\widehat{c}$ have the same ℓ_1 norm, so that the standard definition of the EMD applies. Let $f^* = \left[f_{1,1}^*, f_{1,2}^*, \ldots, f_{1,K}^*, f_{2,1}^*, \ldots, f_{K,K}^* \right]^T$ be the vector that solves the optimization problem (11), g^* be the similarly-defined binary vector that solves the optimization problem (12), and $z = f^* - c_{\min} g^*$ is a flow residual. Similarly, let d and t be the similarly-defined vectors collecting all ground distances $d_{i,j}$

and $t_{i,j}$. Then from (12) and (11), we have

$$\text{EMD}\,(c, \widehat{c}) = d^T f^* = d^T\,(c_{\min} g^* + z) = \frac{c_{\min}}{\Delta} t^T g^* + d^T z = \frac{c_{\min}}{\Delta}\,\text{PEE}\left(\theta, \widehat{\theta}\right) + d^T z. \qquad (A.1)$$

Note that the first term in (A.1) is the value of the objective function in the optimization problem (11) when all entries of both c and $\widehat{c}$ have magnitude $c_{\min}$, i.e., when $z = 0$. The second term $d^T z$ corresponds to the contribution to the objective function due to magnitudes that are larger than $c_{\min}$. We show now that this second term is nonnegative.

When the magnitude of the entry of c corresponds to $\widehat{\theta}_i$ increases from its baseline value of $c_{\min}$ to c_i, at least one of the outgoing flows $f_{i,j}$ have to increase. This implies that the corresponding flow $f^*_{i,j} \geq c_{\min} g^*_{i,j}$, i.e., the residual $z_{i,j} = f^*_{i,j} - c_{\min} g^*_{i,j} \geq 0$. Thus, the entries r are all nonnegative, and $d^T r \geq 0$. Then one can rewrite (A.1) as $\text{EMD}(c, \widehat{c}) \geq \frac{c_{\min}}{\Delta}\text{PEE}\left(\theta, \widehat{\theta}\right)$.

To end, we consider the case when $\|c\|_1 \neq \|\widehat{c}\|_1$, where we have to add additional entries to the vectors: $c_{N+1} = c_{\min} + \max\{0, \|\widehat{c}\|_1 - \|c\|_1\}$ and $\widehat{c}_{N+1} = c_{\min} + \max\{0, \|c\|_1 - \|\widehat{c}\|_1\}$. The additional estimated parameters due to those additional entries are both θ_m since the indices of the additional entries are the same. The resulting $K + 1$-sparse vectors with additional entries are $c^* = [c_1, c_2, \ldots, c_N, c_{N+1}]$ and $\widehat{c}^* = [\widehat{c}_1, \widehat{c}_2, \ldots, \widehat{c}_N, \widehat{c}_{N+1}]$; these two new vectors have matching ℓ_1 norms, $\|c^*\|_1 = \|\widehat{c}^*\|_1$, and the resulting estimated parameters are $\theta^* = [\theta_1, \theta_2, \ldots, \theta_K, \theta_m]$ and $\widehat{\theta}^* = \left[\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_K, \theta_m\right]$, where θ_m is an (arbitrary) parameter value assigned to the newly added $N + 1$-index entry. Therefore, we get $\text{PEE}\left(\theta, \widehat{\theta}\right) = \text{PEE}\left(\theta^*, \widehat{\theta}^*\right)$, due to a match between the added parameter values θ_m in θ^* and in $\widehat{\theta}^*$.

Next, we consider the effect of the mismatch of the ℓ_1 norms of the vectors c and $\widehat{c}$ in the computation of $\text{EMD}\,(c, \widehat{c})$ and $\text{EMD}\,(c^*, \widehat{c}^*)$. We can see that $\text{EMD}\,(c, \widehat{c}) \geq \text{EMD}\,(c^*, \widehat{c}^*)$: when we compute $\text{EMD}\,(c, \widehat{c})$, the ℓ_1 norm mismatch penalty in the objective function of (11) – as defined in Footnote 1 and involving the flows $f_{i,N+1}$ and $f_{N+1,j}$ from/to the entries $c_{N+1}, \widehat{c}_{N+1}$ to/from any $c_j, \widehat{c}_j$ – is no less than the actual contribution of the flows from/to the new entries $c^*_{N+1}, \widehat{c}^*_{N+1}$ to/from the existing entries $c^*_j, \widehat{c}^*_j$ in computing $\text{EMD}\,(c^*, \widehat{c}^*)$, as those distances $d_{i,N+1}$ and $d_{N+1,j}$ are all less than or equal to N. Putting this together with the equivalence in PEE and Theorem 1 when applied to vectors with matching ℓ_1 norms, we obtain that $\frac{c_{\min}}{\Delta}\text{PEE}\left(\theta, \widehat{\theta}\right) = \frac{c_{\min}}{\Delta}\text{PEE}\left(\theta^*, \widehat{\theta}^*\right) \leq \text{EMD}\,(c^*, \widehat{c}^*) \leq \text{EMD}\,(c, \widehat{c})$.

Appendix B. Proof of Theorem 2

As the sampling step of the parameter $\Delta \to 0$, the proxy defined as (19) becomes a continuous function such that

$$v(\omega) = \sum_{i=1}^{K} c_i \lambda(\omega - \theta_i), \tag{B.1}$$

for all $\omega \in \bar{\Theta}$. Additionally, the balanced weight properties around the cluster centroid $\widehat{\theta}_i$, as defined in (10), reduces in the asymptotic case to

$$\int_{p \in C_j, p \le \widehat{\theta}_i} w(p)dp = \int_{p \in C_j, p \ge \widehat{\theta}_j} w(p)dp, \tag{B.2}$$

where p is the position function and w is the weight function. Furthermore, the cumulative correlation function converges to

$$\Lambda(\theta) = \int_{-\infty}^{\theta} \lambda(\omega)d\omega. \tag{B.3}$$

Without loss of generality, if $\theta_{\min} = \min\left(\bar{\Theta}\right)$ and $\theta_{\max} = \max\left(\bar{\Theta}\right)$, assume that parameter values are sorted so that

$$\theta_{\min} + \epsilon \le \theta_1 < \theta_2 < \cdots < \theta_K \le \theta_{\max} - \epsilon. \tag{B.4}$$

When the entries of the proxy v are clustered into K groups according to the centroids $\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_K$, as shown in Algorithm 1, the point $\left(\widehat{\theta}_i + \widehat{\theta}_{i+1}\right)/2$ is the upper bound for i^{th} cluster and the lower bound for $(i+1)^{\text{th}}$ cluster, since that value/location has the same distance to the two centroids around it. We will show how large the minimum separation ζ and minimum off-bound distance ϵ need to be such that the maximum estimation error is σ, i.e. $\max_k \left|\widehat{\theta}_k - \theta_k\right| \le \sigma$.

We first consider the kth cluster with centroid $\widehat{\theta}_k$, where $k = 2, 3, \ldots, K - 1$. The kth cluster includes the parameter range $\left[\left(\widehat{\theta}_{k-1} + \widehat{\theta}_k\right)/2, \left(\widehat{\theta}_k + \widehat{\theta}_{k+1}\right)/2\right]$. According to the weight balance property (B.2), we need for the proxy function in (B.1) to have the same sum over the range $\left[\left(\widehat{\theta}_{k-1} + \widehat{\theta}_k\right)/2, \widehat{\theta}_k\right]$ and

$\left[\widehat{\theta}_k, \left(\widehat{\theta}_k + \widehat{\theta}_{k+1}\right)/2\right]$, i.e.,

$$\int_{\frac{\widehat{\theta}_{k-1}+\widehat{\theta}_k}{2}}^{\widehat{\theta}_k} v(\omega)\, d\omega = \int_{\widehat{\theta}_k}^{\frac{\widehat{\theta}_k+\widehat{\theta}_{k+1}}{2}} v(\omega)\, d\omega,$$

$$2\int_{-\infty}^{\widehat{\theta}_k} v(\omega)\, d\omega = \int_{-\infty}^{\frac{\widehat{\theta}_{k-1}+\widehat{\theta}_k}{2}} v(\omega)\, d\omega + \int_{-\infty}^{\frac{\widehat{\theta}_k+\widehat{\theta}_{k+1}}{2}} v(\omega)\, d\omega,$$

$$2\sum_{i=1}^{K} c_i \Lambda\left(\widehat{\theta}_k - \theta_i\right) = \sum_{i=1}^{K} c_i \Lambda\left(\frac{\widehat{\theta}_{k-1}+\widehat{\theta}_k}{2} - \theta_i\right) + \sum_{i=1}^{K} c_i \Lambda\left(\frac{\widehat{\theta}_k+\widehat{\theta}_{k+1}}{2} - \theta_i\right),$$

$$2c_k \Lambda\left(\widehat{\theta}_k - \theta_k\right) = \sum_{i=1}^{K} c_i \Lambda\left(\frac{\widehat{\theta}_{k-1}+\widehat{\theta}_k}{2} - \theta_i\right) + \sum_{i=1}^{K} c_i \Lambda\left(\frac{\widehat{\theta}_k+\widehat{\theta}_{k+1}}{2} - \theta_i\right) - 2\sum_{i\neq k} c_i \Lambda\left(\widehat{\theta}_k - \theta_i\right). \quad \text{(B.5)}$$

Since $\widehat{\theta}_k - \theta_k \in \{-\sigma, \sigma\}$ and $\theta_{k+1} - \theta_k \geq \zeta$, for $k = 2, 3, \cdots, K-1$, we obtain a lower bound of the term on the left hand side of (B.5) by repeatedly using the fact that $\Lambda(\omega)$ is nondecreasing:

$$\sum_{i=1}^{K} c_i \Lambda\left(\frac{\widehat{\theta}_{k-1}+\widehat{\theta}_k}{2} - \theta_i\right) + \sum_{i=1}^{K} c_i \Lambda\left(\frac{\widehat{\theta}_k+\widehat{\theta}_{k+1}}{2} - \theta_i\right) - 2\sum_{i\neq k} c_i \Lambda\left(\widehat{\theta}_k - \theta_i\right)$$

$$= \sum_{i=1}^{k-1} c_i \Lambda\left(\frac{\widehat{\theta}_{k-1}+\widehat{\theta}_k}{2} - \theta_{k-1}\right) + \sum_{i=k}^{K} c_i \Lambda\left(\frac{\widehat{\theta}_{k-1}+\widehat{\theta}_k}{2} - \theta_K\right) + \sum_{i=1}^{k} c_i \Lambda\left(\frac{\widehat{\theta}_k+\widehat{\theta}_{k+1}}{2} - \theta_k\right)$$

$$+ \sum_{i=k+1}^{K} c_i \Lambda\left(\frac{\widehat{\theta}_k+\widehat{\theta}_{k+1}}{2} - \theta_K\right) - 2\sum_{i=1}^{k-1} c_i \Lambda\left(\widehat{\theta}_k - \theta_1\right) - 2\sum_{i=k+1}^{K} c_i \Lambda\left(\widehat{\theta}_k - \theta_{k+1}\right)$$

$$\geq \sum_{i=1}^{k-1} c_i \Lambda\left(\frac{\widehat{\theta}_{k-1} - \theta_{k-1} + \widehat{\theta}_k - \theta_k + \theta_k - \theta_{k-1}}{2}\right) + \sum_{i=k}^{K} c_i \Lambda(-\infty)$$

$$+ \sum_{i=1}^{k} c_i \Lambda\left(\frac{\widehat{\theta}_k - \theta_k + \widehat{\theta}_{k+1} - \theta_{k+1} + \theta_{k+1} - \theta_k}{2}\right) + \sum_{i=k+1}^{K} c_i \Lambda(-\infty) - 2\sum_{i=1}^{k-1} c_i \Lambda(\infty)$$

$$- 2\sum_{i=k+1}^{K} c_i \Lambda\left(\widehat{\theta}_k - \theta_k + \theta_k - \theta_{k+1}\right)$$

$$\geq \sum_{i=1}^{k-1} c_i \Lambda\left(\frac{\zeta}{2} - \sigma\right) + \sum_{i=1}^{k} c_i \Lambda\left(\frac{\zeta}{2} - \sigma\right) - 2\sum_{i=1}^{k-1} c_i \Lambda(\infty) - 2\sum_{i=k+1}^{K} c_i \Lambda\left(-\frac{\zeta}{2} + \sigma\right)$$

$$\geq \sum_{i=1}^{k-1} c_i \Lambda\left(\frac{\zeta}{2} - \sigma\right) + \sum_{i=1}^{k} c_i \Lambda\left(\frac{\zeta}{2} - \sigma\right) - 2\sum_{i=1}^{k-1} c_i \Lambda(\infty) - 2\sum_{i=k+1}^{K} c_i \left(\Lambda(\infty) - \Lambda\left(\frac{\zeta}{2} - \sigma\right)\right)$$

$$\geq -2\sum_{i\neq k} \frac{c_i}{c_k} c_k \left(\Lambda(\infty) - \Lambda\left(\frac{\zeta}{2} - \sigma\right)\right) + c_k \Lambda\left(\frac{\zeta}{2} - \sigma\right)$$

$$\geq -2(K-1)r c_k \left(\Lambda(\infty) - \Lambda\left(\frac{\zeta}{2} - \sigma\right)\right) + c_k \Lambda\left(\frac{\zeta}{2} - \sigma\right)$$

$$\geq (2(K-1)r + 1)\, c_k \Lambda\left(\frac{\zeta}{2} - \sigma\right) - 2(K-1)r c_k \Lambda(\infty). \quad \text{(B.6)}$$

32

So the lower bound of the term on the left hand side of (B.5) is

$$2\Lambda\left(\widehat{\theta}_k - \theta_k\right) \geq (2(K-1)r + 1)\,\Lambda\left(\frac{\zeta}{2} - \sigma\right) - 2(K-1)r\Lambda(\infty). \tag{B.7}$$

Similarly, one can obtain the following upper bound of term on the left hand side of (B.5):

$$2\Lambda\left(\widehat{\theta}_k - \theta_k\right) \leq (2(K-1)r + 2)\,\Lambda(\infty) - (2(K-1)r + 1)\,\Lambda\left(\frac{\zeta}{2} - \sigma\right). \tag{B.8}$$

Thus, if ζ satisfies

$$\Lambda\left(\frac{\zeta}{2} - \sigma\right) \geq \Lambda(\infty)\left(1 - \frac{\Lambda(\sigma)/\Lambda(0) - 1}{2(K-1)r + 1}\right), \tag{B.9}$$

then the estimation error satisfies

$$
\begin{aligned}
2\Lambda\left(\widehat{\theta}_k - \theta_k\right) &\geq (2(K-1)r + 1)\,\Lambda\left(\frac{\zeta}{2} - \sigma\right) - 2(K-1)r\Lambda(\infty) \\
&\geq (2(K-1)r + 1)\,\Lambda(\infty) - 2\Lambda(\sigma) + \Lambda(\infty) - 2(K-1)r\Lambda(\infty) \\
&\geq 2\Lambda(\infty) - 2\Lambda(\sigma) \geq 2\Lambda(-\sigma),
\end{aligned}
\tag{B.10}
$$

and

$$
\begin{aligned}
2\Lambda\left(\widehat{\theta}_k - \theta_k\right) &\leq (2(K-1)r + 2)\,\Lambda(\infty) - (2(K-1)r + 1)\,\Lambda\left(\frac{\zeta}{2} - \sigma\right) \\
&\leq (2(K-1)r + 2)\,\Lambda(\infty) - (2(K-1)r + 1)\,\Lambda(\infty) + 2\Lambda(\sigma) - \Lambda(\infty) \\
&\geq 2\Lambda(\sigma),
\end{aligned}
\tag{B.11}
$$

which implies $-\sigma \leq \widehat{\theta}_k - \theta_k \leq \sigma$ for $k = 2, 3, \cdots, K-1$.

Next, we consider the first cluster with centroid $\widehat{\theta}_1$, which includes the parameter range $\left[\theta_{\min}, \left(\widehat{\theta}_1 + \widehat{\theta}_2\right)/2\right]$.

From the weight balance property, we have

$$\int_{\theta_{\min}}^{\widehat{\theta}_1} v(\omega)\, d\omega = \int_{\widehat{\theta}_1}^{\frac{\widehat{\theta}_1 + \widehat{\theta}_2}{2}} v(\omega)\, d\omega,$$

$$2\int_{-\infty}^{\widehat{\theta}_1} v(\omega)\, d\omega = \int_{-\infty}^{\theta_{\min}} v(\omega)\, d\omega + \int_{-\infty}^{\frac{\widehat{\theta}_1 + \widehat{\theta}_2}{2}} v(\omega)\, d\omega,$$

$$2\sum_{i=1}^{K} c_i \Lambda\left(\widehat{\theta}_1 - \theta_i\right) = \sum_{i=1}^{K} c_i \Lambda\left(\theta_{\min} - \theta_i\right) + \sum_{i=1}^{K} c_i \Lambda\left(\frac{\widehat{\theta}_1 + \widehat{\theta}_2}{2} - \theta_i\right),$$

$$2c_1 \Lambda\left(\widehat{\theta}_1 - \theta_1\right) = \sum_{i=1}^{K} c_i \Lambda\left(\theta_{\min} - \theta_i\right) + \sum_{i=1}^{K} c_i \Lambda\left(\frac{\widehat{\theta}_1 + \widehat{\theta}_2}{2} - \theta_i\right) - 2\sum_{i=2}^{K} c_i \Lambda\left(\widehat{\theta}_1 - \theta_i\right). \tag{B.12}$$

If ϵ satisfies

$$\Lambda(\epsilon) \geq \Lambda(\infty)\left(1 - \frac{\Lambda(\sigma)/\Lambda(0) - 1}{2Kr}\right), \tag{B.13}$$

then we have the following result from (B.12) by using the relationship (B.9):

$$
\begin{aligned}
2\Lambda\left(\widehat{\theta}_1 - \theta_1\right) &= \sum_{i=1}^{K} \frac{c_i}{c_1}\Lambda\left(\theta_{\min} - \theta_i\right) + \sum_{i=1}^{K} \frac{c_i}{c_1}\Lambda\left(\frac{\widehat{\theta}_1 + \widehat{\theta}_2}{2} - \theta_i\right) - 2\sum_{i=2}^{K} \frac{c_i}{c_1}\Lambda\left(\widehat{\theta}_1 - \theta_i\right) \\
&\leq \sum_{i=1}^{K} \frac{c_i}{c_1}\Lambda\left(\theta_{\min} - \theta_1\right) + \Lambda\left(\frac{\widehat{\theta}_1 + \widehat{\theta}_2}{2} - \theta_1\right) + \sum_{i=2}^{K} \frac{c_i}{c_1}\Lambda\left(\frac{\widehat{\theta}_1 + \widehat{\theta}_2}{2} - \theta_2\right) - 2\sum_{i=2}^{K} \frac{c_i}{c_1}\Lambda\left(\widehat{\theta}_1 - \theta_K\right) \\
&\leq Kr\Lambda(-\epsilon) + \Lambda(\infty) + (K-1)r\Lambda\left(-\frac{\zeta}{2} + \sigma\right) \\
&\leq Kr\left(\Lambda(\infty) - \Lambda(\epsilon)\right) + \Lambda(\infty) + (K-1)r\left(\Lambda(\infty) - \Lambda\left(\frac{\zeta}{2} - \sigma\right)\right) \\
&\leq Kr\frac{2\Lambda(\sigma) - \Lambda(\infty)}{2Kr} + \Lambda(\infty) + (K-1)r\frac{2\Lambda(\sigma) - \Lambda(\infty)}{2(K-1)r + 1} \\
&\leq 2\Lambda(\sigma) \tag{B.14}
\end{aligned}
$$

It can be similarly shown for the last cluster centroid that $2\Lambda\left(\widehat{\theta}_K - \theta_K\right) \geq 2\Lambda(-\sigma)$.

In summary, when the minimum separation ζ and off-bound distance ϵ satisfy equation (B.9) and (B.13), respectively, all estimates satisfy $\left|\widehat{\theta}_k - \theta_k\right| \leq \sigma, k = 1, 2, \ldots, K$, and hence PEE $\left(\widehat{\theta}, \theta\right) \leq K\sigma$

Appendix C. Proof of Theorem 3

When the redefined correlation function is $\lambda_\Phi(\omega) = \exp(-a|\omega|)$, the proxy function given in (26) is

$$v(\omega) = \sum_{i=1}^{K} c_i \exp(-a|\omega - \theta_i|). \tag{C.1}$$

Without loss of generality, we assume that parameter values are sorted so that $\theta_1 < \theta_2 < \cdots < \theta_K$ and all component magnitudes are no smaller than 1. We also assume that the minimum off-bound distance ϵ is sufficiently large and can be ignored during the proof.

We denote by $[l_j, u_j]$ the parameter range around each θ_j will be preserved after thresholding with level t. We have $l_1 < \theta_1 < u_1 < \cdots < \theta_{j-1} < u_{j-1} < l_j < \theta_j < u_j < \cdots < \theta_K$. Due to the important fact that the preserved parameter range is small when the threshold level is large, as t increase, u_j decrease (i.e., the upper bound decrease) and l_j increase (i.e., the lower bound increase). So the proxy at $\omega = u_j$ has value equal to the threshold, i.e.,

$$\frac{t}{c_j} = \frac{\nu(u_j)}{c_j} = \sum_{i=1}^{K} \frac{c_i}{c_j} \exp(-a|u_j - \theta_i|),$$

$$\frac{t}{c_j} = \sum_{i=1}^{j-1} \frac{c_i}{c_j} \exp(-a(u_j - \theta_i)) + \exp(-a(u_j - \theta_j)) + \sum_{i=j+1}^{K} \frac{c_i}{c_j} \exp(-a(\theta_i - u_j)),$$

$$\frac{t}{c_j} = \exp(-a(u_j - \theta_j)) \sum_{i=1}^{j-1} \frac{c_i}{c_j} \exp(-a(\theta_j - \theta_i)) + \exp(-a(u_j - \theta_j)),$$

$$+ \exp(-a(\theta_j - u_j)) \sum_{i=j+1}^{K} \frac{c_i}{c_j} \exp(-a(\theta_i - \theta_j)),$$

$$T_j = A_j \frac{1}{U_j} + \frac{1}{U_j} + B_j U_j, \tag{C.2}$$

where $U_j = \exp(a(u_j - \theta_j)) > 1$,

$$0 \leq A_j = \sum_{i=1}^{j-1} \frac{c_i}{c_j} \exp(-a(\theta_j - \theta_i)) \leq r \sum_{i=1}^{j-1} \exp(-a(j-i)\zeta) \leq r\frac{1 - \exp(-a\zeta j)}{1 - \exp(-a\zeta)} \leq \frac{r}{\exp(a\zeta) - 1}, \tag{C.3}$$

$$0 \leq B_j = \sum_{i=j+1}^{K} \frac{c_i}{c_j} \exp(-a(\theta_i - \theta_j)) \leq r \sum_{i=1}^{K-j} \exp(-a(j-i)\zeta) \leq r\frac{1 - \exp(-a\zeta(K-j+1))}{1 - \exp(-a\zeta)}$$

$$\leq \frac{r}{\exp(a\zeta) - 1}, \tag{C.4}$$

and $T_j = t/c_j$. One solution of the quadratic equation (C.2) is

$$U_j = \frac{T_j - \sqrt{T_j^2 - 4(1 + A_j)B_j}}{2B_j}. \tag{C.5}$$

In solution (C.5), U_j decreases as T_j increases. The alternative solution is omitted due to the fact that U_j will increase as T_j increases, which implies that u_j will increase as t decreases. Similarly,

$$\frac{t}{c_j} = \frac{\nu(l_j)}{c_j} = \sum_{i=1}^{K} \exp\left(-a\left|l_j - \theta_i\right|\right),$$

$$\frac{t}{c_j} = \sum_{i=1}^{j-1} \frac{c_i}{c_j} \exp\left(-a\left(l_j - \theta_i\right)\right) + \exp\left(-a\left(\theta_j - l_j\right)\right) + \sum_{i=j+1}^{K} \frac{c_i}{c_j} \exp\left(-a\left(\theta_i - l_j\right)\right),$$

$$\frac{t}{c_j} = \exp\left(-a\left(l_j - \theta_j\right)\right) \sum_{i=1}^{j-1} \frac{c_i}{c_j} \exp\left(-a\left(\theta_j - \theta_i\right)\right) + \exp\left(-a\left(\theta_j - l_j\right)\right),$$

$$+ \exp\left(-a\left(\theta_j - l_j\right)\right) \sum_{i=j+1}^{K} \frac{c_i}{c_j} \exp\left(-a\left(\theta_i - \theta_j\right)\right),$$

$$T_j = A_j L_j + \frac{1}{L_j} + B_j \frac{1}{L_j}, \tag{C.6}$$

where $L_j = \exp\left(a\left(\theta_j - l_j\right)\right) > 1$. The corresponding solution is

$$L_j = \frac{T_j - \sqrt{T_j^2 - 4A_j(1 + B_j)}}{2A_j}. \tag{C.7}$$

In order to have real values for solutions (C.5) and (C.7), T_j (and therefore t) should be sufficiently large. Since

$$\max\left((1 + A_j)B_j, A_j(1 + B_j)\right) \leq \left(1 + \frac{r}{\exp(a\zeta) - 1}\right) \frac{r}{\exp(a\zeta) - 1} \leq \frac{2r^2}{(\exp(a\zeta) - 1)^2}, \tag{C.8}$$

if

$$T_j^2 = \left(\frac{t}{c_j}\right)^2 \geq \left(\frac{t}{c_{\max}}\right)^2 \geq \frac{8r^2}{(\exp(a\zeta) - 1)^2}, \tag{C.9}$$

then both (C.5) and (C.7) are real valued.

Let $\widehat{\theta}_j \in [l_j, u_j]$ be the estimated parameter for θ_j. Asymptotically, when the sampling step of the

36

parameter space Δ goes to zero, the balance weight properties (B.2) implies

$$\int_{l_j}^{\widehat{\theta}_j} \sum_{i=1}^{K} c_i \exp\left(-a\left|\omega - \theta_i\right|\right) d\omega = \int_{\widehat{\theta}_j}^{u_j} \sum_{i=1}^{K} c_i \exp\left(-a\left|\omega - \theta_i\right|\right) d\omega. \tag{C.10}$$

When $\widehat{\theta}_j \leq \theta_j$, the left hand side of (C.10) is

$$\frac{a}{c_j} \int_{l_j}^{\widehat{\theta}_j} \sum_{i=1}^{K} c_i \exp\left(-a\left|\omega - \theta_i\right|\right) d\omega$$

$$= a \sum_{i=1}^{j-1} \frac{c_i}{c_j} \int_{l_j}^{\widehat{\theta}_j} \exp\left(-a\left(\omega - \theta_i\right)\right) d\omega + a \int_{l_j}^{\widehat{\theta}_j} \exp\left(-a\left(\theta_j - \omega\right)\right) d\omega + a \sum_{i=j+1}^{K} \frac{c_i}{c_j} \int_{l_j}^{\widehat{\theta}_j} \exp\left(-a\left(\theta_i - \omega\right)\right) d\omega$$

$$= a A_j \int_{l_j}^{\widehat{\theta}_j} \exp\left(-a\left(\omega - \theta_j\right)\right) d\omega + a \int_{l_j}^{\widehat{\theta}_j} \exp\left(-a\left(\theta_j - \omega\right)\right) d\omega + a B_j \int_{l_j}^{\widehat{\theta}_j} \exp\left(-a\left(\theta_j - \omega\right)\right) d\omega$$

$$= A_j \left(L_j - E_j\right) + \left(B_j + 1\right) \left(\frac{1}{E_j} - \frac{1}{L_j}\right)$$

$$= -A_j E_j + B_j \frac{1}{E_j} + \frac{1}{E_j} + A_j L_j - \frac{1}{L_j} - B_j \frac{1}{L_j}, \tag{C.11}$$

where $E_j = 1/\lambda(\widehat{\theta}_j - \theta_j) = \exp\left(a\left|\widehat{\theta}_j - \theta_j\right|\right) = \exp\left(a\left(\theta_j - \widehat{\theta}_j\right)\right) \geq 1$, and the right hand side of (C.10) is

$$\frac{a}{c_j} \int_{\widehat{\theta}_j}^{u_j} \sum_{i=1}^{K} c_i \exp\left(-a\left|\omega - \theta_i\right|\right) d\omega$$

$$= a \sum_{i=1}^{j-1} \frac{c_i}{c_j} \int_{\widehat{\theta}_j}^{u_j} \exp\left(-a\left(\omega - \theta_i\right)\right) d\omega + a \int_{\widehat{\theta}_j}^{u_j} \exp\left(-a\left|\omega - \theta_j\right|\right) d\omega + a \sum_{i=j+1}^{K} \frac{c_i}{c_j} \int_{\widehat{\theta}_j}^{u_j} \exp\left(-a\left(\theta_i - \omega\right)\right) d\omega$$

$$= a \sum_{i=1}^{j-1} \frac{c_i}{c_j} \int_{\widehat{\theta}_j}^{u_j} \exp\left(-a\left(\omega - \theta_i\right)\right) d\omega + a \int_{\widehat{\theta}_j}^{\theta_j} \exp\left(-a\left(\theta_j - \omega\right)\right) d\omega$$

$$\qquad + a \int_{\theta_j}^{u_j} \exp\left(-a\left(\omega - \theta_j\right)\right) d\omega + a \sum_{i=j+1}^{K} \frac{c_i}{c_j} \int_{\widehat{\theta}_j}^{u_j} \exp\left(-a\left(\theta_i - \omega\right)\right) d\omega$$

$$= A_j \left(E_j - \frac{1}{U_j}\right) + 1 - \frac{1}{E_j} + 1 - \frac{1}{U_j} + B_j \left(U_j - \frac{1}{E_j}\right)$$

$$= A_j E_j - B_j \frac{1}{E_j} + 2 - \frac{1}{E_j} - A \frac{1}{U_j} - \frac{1}{U_j} + B_j U_j. \tag{C.12}$$

After plugging (C.11) and (C.12) into (C.10) and moving all terms with L_j or U_j to the right side and

moving other terms to the left side, we obtain

$$2A_j E_j - 2B_j \frac{1}{E_j} + 2 - \frac{2}{E_j}$$

$$=A_j L_j - \frac{1}{L_j} - B_j \frac{1}{L_j} + A_j \frac{1}{U_j} + \frac{1}{U_j} - B_j U_j$$

$$=A_j L_j + A_j L_j - T_j + T_j - B_j U_j - B_j U_j$$

$$=2A_j L_j - 2B_j U_j.$$

$$= \left(T_j - \sqrt{T_j^2 - 4A_j(1+B_j)} \right) - \left(T_j - \sqrt{T_j^2 - 4(1+A_j)B_j} \right), \tag{C.13}$$

$$=\sqrt{T_j^2 - 4(1+A_j)B_j} - \sqrt{T_j^2 - 4A_j(1+B_j)}$$

$$=\frac{\left(T_j^2 - 4(1+A_j)B_j \right) - \left(T_j^2 - 4A_j(1+B_j) \right)}{\sqrt{T_j^2 - 4A_j(1+B_j)} + \sqrt{T_j^2 - 4(1+A_j)B_j}}$$

$$=\frac{4(A_j - B_j)}{\sqrt{T_j^2 - 4A_j(1+B_j)} + \sqrt{T_j^2 - 4(1+A_j)B_j}}$$

where the second equality results from (C.2) and (C.6).

One can show a similar result when $\theta_j > \widehat{\theta}_j$ and $E_j = \exp\left(a \left| \theta_j - \widehat{\theta}_j \right| \right) = \exp\left(a \left(\theta_j - \widehat{\theta}_j \right) \right)$, so (C.10) can be written as

$$\frac{A_j - B_j}{S_j} = \begin{cases} A_j E_j - B_j \dfrac{1}{E_j} + 1 - \dfrac{1}{E_j} & \text{if} \quad \theta_j \leq \widehat{\theta}_j \\[2mm] A_j \dfrac{1}{E_j} - B_j E_j - 1 + \dfrac{1}{E_j} & \text{if} \quad \theta_j > \widehat{\theta}_j \end{cases}, \tag{C.14}$$

where

$$S_j = \frac{1}{2} \left(\sqrt{T_j^2 - 4A_j(1+B_j)} + \sqrt{T_j^2 - 4(1+A_j)B_j} \right) \leq T_j \leq \frac{t}{c_{\min}} \leq 1 \leq E_j, \tag{C.15}$$

and

$$S_j = \frac{1}{2} \left(\sqrt{T_j^2 - 4A_j(1+B_j)} + \sqrt{T_j^2 - 4(1+A_j)B_j} \right) \geq \sqrt{\left(\frac{t}{rc_{\min}} \right) - 8 \left(\frac{r}{\exp(a\zeta) - 1} \right)^2} \tag{C.16}$$

using (C.8).

When $\theta_j \leq \widehat{\theta}_j$,

$$(A_j - B_j)/S_j = A_j E_j - B_j/E_j + 1 - 1/E_j \geq (A_j - B_j)/E_j, \tag{C.17}$$

38

So $A_j \geq B_j$ due to the fact that $S_j \leq E_j$. If

$$S_j \geq \sqrt{\left(\frac{t}{rc_{\min}}\right) - 8\left(\frac{r}{\exp(a\zeta) - 1}\right)^2} \geq \exp(-a\sigma) \geq 0, \tag{C.18}$$

which can be rewritten as

$$\zeta \geq \frac{1}{a}\ln\left(\sqrt{\frac{8r^2}{t^2/(rc_{\min})^2 - \exp(-2a\sigma)}} + 1\right), \tag{C.19}$$

the condition (C.9) satisfies, so the solutions (C.5) and (C.7) are both real valued. Additionally, we have

$$A_j E_j - B_j \frac{1}{E_j} + 1 - \frac{1}{E_j} = \frac{A_j - B_j}{S_j} \leq (A_j - B_j)\exp(a\sigma) \leq A_j\exp(a\sigma) - B_j\exp(-a\sigma) + 1 - \exp(-a\sigma). \tag{C.20}$$

This implies that $E_j \leq \exp(a\sigma)$ and hence $0 \leq \widehat{\theta}_j - \theta_j \leq \sigma$. Similarly, when $\theta_j > \widehat{\theta}_j$, $A_j \leq B_j$ and $-\sigma \leq \widehat{\theta}_j - \theta_j \leq 0$.

Finally, the condition for ζ to be real valued is $t \geq rc_{\min}\exp(-a\sigma)$, which can be obtained from (C.19).

References

References

[1] D. Mo, M. F. Duarte, Compressive parameter estimation with earth mover's distance via k-median clustering, in: Proc. SPIE Wavelets and Sparsity XV, Vol. 8858, San Diego, CA, 2013.

[2] D. Mo, M. F. Duarte, Compressive line spectrum estimation with clustering and interpolation, in: 2016 Annual Conference on Information Science and Systems (CISS), 2016, pp. 572 – 577.

[3] R. G. Baraniuk, Compressive sensing, IEEE Signal Proc. Mag. 24 (4) (2007) 118–121.

[4] E. J. Candés, Compressive sampling, in: Proc. Int. Congr. Math. (ICM), Vol. 3, Madrid, Spain, 2006, pp. 1433–1452.

[5] D. L. Donoho, Compressed sensing, IEEE Trans. Inf. Theory 52 (4) (2006) 1289–1306.

[6] V. Cevher, A. C. Gurbuz, J. H. McClellan, R. Chellappa, Compressive wireless arrays for bearing estimation, in: IEEE Int. Conf. Acoustics, Speech and Signal Proc. (ICASSP), Las Vegas, NV, 2008, pp. 2497–2500.

[7] M. F. Duarte, Localization and bearing estimation via structured sparsity models, in: IEEE Stat. Signal Proc. Workshop (SSP), Ann Arbor, MI, 2012, pp. 333–336.

[8] M. A. Herman, T. Strohmer, High resolution radar via compressed sensing, IEEE Trans. Inf. Theory 57 (6) (2009) 2275–2284.

[9] H. J. Rad, G. Leus, Sparsity-aware tdoa localization of multiple sources, IEEE Trans. Signal Proc. 61 (19) (2013) 4021–4025.

[10] S. Sen, G. Tang, A. Nehorai, Multiobjective optimization of ofdm radar waveform for target detection, IEEE Trans. Inf. Theory 59 (2) (2011) 639–652.

[11] I. Stojanovic, M. Çetin, W. C. Karl, Compressed sensing of monostatic and multistatic sar, IEEE Geosci. Remote Sens. Lett 10 (6) (2013) 1444–1448.

[12] A. Eftekhari, J. Romberg, M. B. Wakin, Matched filtering from limited frequency samples, IEEE Trans. Inf. Theory 59 (6) (2013) 3475–3496.

[13] K. Fyhn, M. F. Duarte, S. H. Jensen, Compressive time delay estimation using interpolation, in: IEEE Global Conf. Signal and Info. Processing (GlobalSIP), Austin, TX, 2013, pp. 624–624.

[14] C. D. Austin, R. L. Moses, J. N. Ash, E. Ertin, On the relation between sparse reconstruction and parameter estimation with model order selection, IEEE J. Sel. Topics in Signal Proc. 4 (3) (2010) 560–570.

[15] S. Bourguignon, H. Carfantan, J. Idier, A sparsity-based method for the estimation of spectral lines from irregularly sampled data, IEEE J. Sel. Topics in Signal Proc. 1 (4) (2007) 575–585.

[16] M. F. Duarte, R. G. Baraniuk, Spectral compressive sensing, Appl. and Comput. Harmonic Anal. 35 (1) (2013) 111–129.

[17] K. Fyhn, H. Dadkhahi, M. F. Duarte, Spectral compressive sensing with polar interpolation, in: IEEE Int. Conf. Acoustics, Speech and Signal Proc. (ICASSP), Vancouver, Canada, 2013, pp. 6225–6229.

[18] A. Fannjiang, W. Liao, Coherence pattern-guided compressive sensing with unresolved grids, SIAM J. Imaging Sci. 5 (1) (2012) 179–202.

[19] Y. Chi, L. L. Scharf, A. Pezeshki, A. R. Calderbank, Sensitivity to basis mismatch in compressive sensing, IEEE Trans. Signal Proc. 59 (5) (2011) 2182 – 2195.

[20] Y. Rubner, C. Tomasi, L. J. Guibas, A metric for distributions with applications to image databases, in: Int. Conf. Comput. Vision (ICCV), Bombay, India, 1998, pp. 59–66.

[21] R. Gupta, P. Indyk, E. Price, Sparse recovery for earth mover distance, in: 48th Ann. Allerton Conf. Commun, Control, and Computing, Monticello, IL, 2010, pp. 1742–1744.

[22] P. Indyk, E. Price, k-median clustering, model-based compressive sensing, and sparse recovery for earth mover distance, in: ACM Symp. Theory of Computing (STOC), San Jose, CA, 2011, pp. 627–636.

[23] E. J. Candés, C. Fernandez-Granda, Towards a mathematical theory of super-resolution, Comm. Pure Appl. Math. 67 (6) (2014) 906 – 956.

[24] E. J. Candés, C. Fernandez-Granda, Super-resolution from noisy data, J. Fourier Analysis and Applicat. 19 (6) (2013) 1229–1254.

[25] B. N. Bhaskar, G. Tang, B. Recht, Atomic norm denoising with applications to line spectral estimatin, IEEE Trans. Signal Proc. 61 (23) (2013) 5987–5999.

[26] G. Tang, B. N. Bhaskar, P. Shah, B. Recht, Compressed sensing off the grid, IEEE Trans. Inf. Theory 59 (11) (2013) 7465–7490.

[27] Y. Chi, Y. Chen, Compressive recovery of 2-D off-grid frequencies, in: 2013 Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, 2013, pp. 687 – 691.

[28] Y. Chen, Y. Chi, Robust spectral compressed sensing via structured matrix completion, IEEE Trans. Inf. Theory 60 (10) (2014) 6576 – 6601.

[29] J.-F. Cai, X. Qu, W. Xu, G.-B. Ye, Robust recovery of complex exponential signals from random gaussian projections via low rank hankel matrix reconstruction, Appl. and Comput. Harmonic Anal. 41 (2) (2016) 470–490.

[30] H. Saariiisaari, TLS-ESPRIT in a time delay estimation, in: IEEE 47th Vehicular Technology Conference, Vol. 3, 1997, pp. 1619 – 1623.

[31] K. Fyhn, M. F. Duarte, S. H. Jensen, Compressive parameter estimation for sparse translation-invariant signals using polar interpolation, IEEE Trans. Signal Proc. 63 (4) (2015) 870 – 881.

[32] D. L. Donoho, M. Elad, Optimally sparse representation in general (non-orthogonal) dictionaries via ℓ_1 minimization, Proc. Natl. Acad. Sci. U.S.A. 100 (5) (2003) 2197–2202.

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

[34] H. Rauhut, K. Schnass, P. Vandergheynst, Compressed sensing and redundant dictionaries, IEEE Trans. Inf. Theory 54 (2008) 2210–2219.

[35] P. S. Bradley, O. L. Mangasarian, W. N. Street, Clustering via concave minimization, in: Advances in Neural Inf. Proc. Systems (NIPS), Vol. 9, Denver, CO, 1996, pp. 368–374.

[36] P. N. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, 2nd Edition, Pearson Education, Limited, 2014.

[37] O. Pele, M. Werman, Fast and robust earth mover's distance, in: Int. Conf. Comput. Vision (ICCV), Kyoto, Japan, 2009, pp. 460–467.

[38] A. Schrijver, Theory of Linear and Integer Programming, John Wiley & Sons, New York, NY, 1998.

[39] S. S. Vempala, The Random Projection Method, American Math. Soc., 2005.

[40] A. Eftekhari, M. B. Wakin, New analysis of manifold embeddings and signal recovery from compressive measurements, Appl. and Comput. Harmonic Anal. 39 (1) (2014) 67–109.

[41] G. Tang, B. Recht, Atomic decomposition of mixtures of translation-invariant signals, in: Comput. Advances in Multi-Sensor Adaptive Proc. (CAMSAP), Saint Martin, France, 2013.

[42] D. Mo, M. F. Duarte, Performance of compressive parameter estimation with earth mover's distance via k-median clustering, Tech. rep., University of Massachusetts, Amherst, MA, available at `http://arxiv.org/pdf/1412.6724` (Dec. 2014).

[43] C. Ekanadham, D. Tranchina, E. P. Simoncelli, Recovery of sparse translation-invariant signals with continuous basis pursuit, IEEE Trans. Inf. Theory 59 (10) (2011) 4735–4744.

[44] W. Dai, O. Milenkovic, Subspace pursuit for compressive sensing signal reconstruction, IEEE Trans. Inf. Theory 55 (5) (2009) 2230–2249.

[45] S. Hamzehei, M. F. Duarte, Compressive parameter estimation via approximate message passing, in: IEEE Int. Conf. Acoustics, Speech and Signal Proc. (ICASSP), Brisbane, Australia, 2015, pp. 3327 – 3331.

[46] S. Hamzehei, M. F. Duarte, Compressive direction-of-arrival estimation compressive direction-of-arrival estimation off the grid, in: 50th Asilomar Conf. Signals, Systems and Computers, Pacific Grove, CA, 2016, pp. 1081 – 1085.

[47] M. Grant, S. Boyd, CVX: Matlab software for disciplined convex programming, version 2.1, `http://cvxr.com/cvx` (Mar. 2014).

[48] P. Stoica, R. L. Moses, Spectral Analysis of Signals, Prentice Hall, 2005.