2014

Compressive Parameter Estimation with Emd

Dian Mo
University of Massachusetts Amherst

COMPRESSIVE PARAMETER ESTIMATION WITH EMD

A Thesis Presented

by

DIAN MO

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

MASTER OF SCIENCE IN ELECTRICAL AND COMPUTER ENGINEERING

February 2014

Electrical and Computer Engineering

COMPRESSIVE PARAMETER ESTIMATION WITH EMD

A Thesis Presented

by

DIAN MO

Approved as to style and content by:

Marco F. Duarte, Chair

Patrick A. Kelly, Member

Mario Parente, Member

Christopher V. Hollot, Department Chair
Electrical and Computer Engineering

ABSTRACT

COMPRESSIVE PARAMETER ESTIMATION WITH EMD

FEBRUARY 2014

DIAN MO

B.Sc., BEIHANG UNIVERSITY

M.S.E.C.E., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Marco F. Duarte

In recent years, sparsity and compressive sensing have attracted significant attention in parameter estimation tasks, including frequency estimation, delay estimation, and localization. Parametric dictionaries collect signals for a sampling of the parameter space and can yield sparse representations for the signals of interest when the sampling is sufficiently dense. While this dense sampling can lead to high coherence in the dictionary, it is possible to leverage structured sparsity models to prevent highly coherent dictionary elements from appearing simultaneously in a signal representation, alleviating these coherence issues. However, the resulting approaches depend heavily on a careful setting of the maximum allowable coherence; furthermore, their guarantees apply to the coefficient vector recovery and do not translate in general to the parameter estimation task. We propose a new algorithm based on optimal sparse approximation measured by earth mover's distance (EMD). Theoretically, we show that EMD provides a better metric for the performance of parametric dictionary-based parameter estimation and K-median clustering algorithms has the potential

to solve the EMD-optimal sparse approximation problems. Simulations show that the resulting compressive parameter estimation algorithm is better at addressing the coherence issuers without a careful setting of additional parameters.

TABLE OF CONTENTS

APPENDICES

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INTRODUCTION

Compressive sensing (CS) has emerged as a framework for integrated sensing and compression of signals that are known to be sparse or compressible in some transformations [1, 2, 3]. CS has attracted significant attention in recent years when its applications have been extended from signal recovery to parameter estimation through the design of parametric dictionaries (PDs), which yield sparse representations for the signals of interest. A PD consists of a set of signals corresponding to a discrete set of parameter values sampled from a continuous parameter space, such as the possible values of frequencies in frequency estimation, delays in time delay estimation, and locations in localization. Intuitively, a PD collects a set of samples of the signal space. Making this connection between parameter estimation and sparsity recovery allows for compressive parameter estimation algorithms that rely on the rich sparsity-based CS framework. The resulting coefficient vectors obtained from CS signal recovery are interpreted by matching each nonzero entry of coefficient vectors to a parameter value. This PD-based approach has been previously formulated for a bunch of landmark parameter estimation problems, including localization, bearing estimation [4, 5, 6, 7, 8, 9], time delay estimation (TDE) [10, 11, 12], and frequency estimation (FE) [13, 14, 15, 16].

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 compressive parameter estimation be perfect. Fortunately, it may give low estimation error if the unknown parameters are very close to some sampled parameters [5]. Thus, dense

sampling of the parameter space may be able to improve the parameter estimation error. However, the resulting PDs will have significantly high coherence, i.e., the largest normalized inner product of any pair of PD elements will become closer to one, which is known to hamper the performance of standard CS recovery algorithms [17]. Previous approaches address this coherence problem by leveraging structured sparsity models [18] to inhibit the highly coherent PD elements from appearing simultaneously in the recovered signals' representation [5, 15, 19, 16, 11]. However, the performance of the resulting algorithms are highly dependent on the careful setting of an allowable value of the maximum coherence between the chosen elements and have to be compensated by the spacing distance of the parameters that can be observed simultaneously.

Another issue that arises in PD-based compressive parameter estimation is that almost all proposed CS recovery algorithms guarantee stable recovery of coefficient vectors with error measured by the ℓ_2 norm, i.e., the estimated coefficient vector is very close to the true coefficient vector in Euclidean distance. Most of these algorithms link the proof of the guarantee to a thresholding operation, which sets all entries of a input vector to zero except for those with largest magnitudes and returns the optimal sparse approximation to the input vector, again, in terms of the ℓ_2 norm. However, such a guarantee has a very limited impact on PD-based compressive parameter estimation, since only in the most demanding case of perfect recovery can the guarantee be linked to the accurate estimation of the indices of nonzero entries of coefficient vectors, which can be translated into accurate parameter estimate. This motivates the need for a new metric for coefficient vectors that is able to capture the difference between two coefficient vectors in terms of similarity between their nonzero entries.

Several metrics are available to measure the error of coefficient vectors in respect of similarity between their nonzero entries rather than Euclidean distance and are

applied to compressive parameter estimation. The Hamming distance measures the number of the entries that are either both zero or both nonzero in the true coefficient vectors and its estimated coefficient vector, and certain CS recovery approaches consider this criterion [20, 21, 22, 23]. Unfortunately, the Hamming distance only controls the number of errors committed in parameter estimation, not the magnitude of the errors that occur. As an alternative, the earth mover's distance (EMD) [24, 25] quantifies the magnitudes of the errors by minimizing the amount and distance of flow among the entries of the estimated coefficient vector that make the estimated coefficient vector become equal to the true coefficient vector. Using this distance in compressive parameter estimation leverages the fact that if the entries of the coefficient vector are sorted by the corresponding parameter values, the distance of flow between any pair of entries is linear with the distance of corresponding parameters, so the EMD between the true coefficient vector and the estimated coefficient vector is indicative of the parameter estimation error. Very recently, the EMD has been integrated within CS to provide recovery algorithms for sparse and compressible signals, where the accuracy is measured in terms of the EMD [26, 27].

In this project, our goal is to derive a new method of PD-based compressive parameter estimation that can address the coherence issue and leverage the EMD to measure the estimation error of coefficient vectors. We will replace the thresholding operation in proposed CS recovery algorithms by a new sparse approximation to search the optimal K-sparse approximations to input vectors in the sense of EMD, which is believed to be well solved by the K-median clustering [26, 27]. Additionally, we will present theorems showing that the parameter estimate error is bounded by EMD of coefficient vectors and that the estimate error resulting from K-median clustering, under some certain condition, can be very small. With these theorems, we are able to incorporate the K-median clustering into standard CS recovery algorithms.

CHAPTER 2

BACKGROUND

2.1 Compressive Sensing

2.1.1 Sparsity

For a long time, sparsity has been exploited in signal processing and approximation, including applications such as image compression and denoising. Usually, a discrete signal $x \in \mathbb{C}^N$ is K-sparse when it has at most K nonzero entries.

Let $I = \{1, 2, \ldots, N\}$ denotes the index set of any N-dimensional vector x. $S \subset I$ represents a subset with only K elements, i.e., $|S| = K$, and S^c represents the relative complement of S in I:

$$S^c = \{i \in I : i \notin S\}. \tag{2.1}$$

$x_S \in \mathbb{C}^K$ represents a K-dimensional vector with K entries of x corresponding to the indices S. Then that x is a K-sparse vector at S means x has nonzero entries only at indices S, i.e., $x_S \in \mathbb{C}^K$ and $x_{S^c} = 0$, where those indices S is called the support of x.

It is easy to verify that both the sum of two sparse vectors with the same support and the scalar multiplication of a sparse vector are also sparse vectors with the same support. So all the sparse signals x with support S form a subspace, called K-dimensional canonical subspace $\mathcal{X}_S$:

$$\mathcal{X}_S = \{x \in \mathbb{C}^N : x_S \in \mathbb{C}^K, x_{S^c} = 0\}. \tag{2.2}$$

There are a total of $\binom{N}{K}$ canonical subspaces and all sparse vectors with at most K nonzero entries lie in the the union of all canonical K-dimension subspaces, which is called the K-sparse model:

$$\Sigma_K = \bigcup_{S \subset I} \mathcal{X}_S. \tag{2.3}$$

Usually, signals are not sparse themselves, but have sparse representation in some bases or frames. In this case, a signal x is K-sparse in a basis or a frame Ψ when there exists a coefficient vector c that has at most K nonzero entries such that

$$x = \Psi c, \quad c \in \Sigma_K. \tag{2.4}$$

A sparse signal is always sparse in the canonical basis, whose matrix representation is the identity matrix. In this paper the concept of sparse refers to the concept of sparse in a basis or a frame unless otherwise specifically stated.

While the elements of a basis are linearly independent, a frame is a generalized concept of basis that the elements in the frame can be linear depend. Due to the redundancy, a frame provides more flexibility than an orthonormal basis due to its redundancy, which leads to improved sparsity properties. Therefore, frames are more often employed than orthonormal bases.

2.1.2 Compressive Sensing

Compressive sensing (CS) acquires and compresses the sparse signal in a random fashion [1, 2, 3]. In CS, a discrete signal $x \in \mathbb{C}^N$ is compressed using a dimension-reducing measurement matrix $\Phi \in \mathbb{R}^{M \times N}$ to obtain linear measurements $y \in \mathbb{C}^M$ as

$$y = \Phi x = \Phi \Psi c. \tag{2.5}$$

In general, it is ill-posed to recover the signal x from its measurements y when $M < N$ and so Φ has a nontrivial null space. But it is possible to recover an accurate sparse

estimation $\widehat{x}$ to the signal x from the measurements y when x is known to be sparse in some basis or frame.

In order to recover a good estimation $\widehat{x}$, the matrix $\Upsilon = \Phi\Psi$ must satisfy the restricted isometry property (RIP) [2, 28]. A matrix Υ has RIP with constant η if, for all $c \in \Sigma_K$,

$$(1-\eta)\|c\|_2^2 \leq \|\Upsilon c\|_2^2 \leq (1+\eta)\|c\|_2^2. \tag{2.6}$$

When $\eta \ll 1$, the matrix Υ approximately preserves the Euclidean distance between any pair of K-sparse signals, so it is possible to invert the sampling process stably. While checking whether the matrix Υ satisfies the RIP is an NP-hard problem [2], fortunately, random matrices whose entries are independently and identically drawn from Gaussian, Bernoulli, or more generally sub-Gaussian distributions satisfy the RIP with high probability providing that $M = \mathcal{O}(K\log(N/K))$ [29, 30].

Given measurements y and the knowledge that x is sparse, it is natural to attempt to recover the coefficients c by solving an optimization problem:

$$\widehat{c} = \arg\min_c \|c\|_0 \quad \text{s.t.} \quad \Upsilon c = y. \tag{2.7}$$

The ℓ_0 norm $\|\cdot\|_0$ counts the number of all nonzero entries of a vector:

$$\|c\|_0 = |\operatorname{supp}(c)|. \tag{2.8}$$

Since the ℓ_0 norm, which does not satisfy the absolute scalability property required by a norm, i.e., $\|\alpha c\|_0 \neq |\alpha|\|c\|_0$, is not a norm and is not a convex function, solving (2.7) is NP-hard.

Recent research shows the possibility to replace the ℓ_0 norm by the ℓ_1 norm to formulate a convex optimization problem as

$$\widehat{c} = \arg\min_c \|c\|_1 \quad \text{s.t.} \quad \Upsilon c = y, \tag{2.9}$$

and the equivalence between (2.7) and (2.9) [31, 32], where the ℓ_1 norm $\|\cdot\|_1$ sums the absolute magnitudes of all nonzero entries of a vector:

$$\|c\|_1 = \sum_i |c_i|. \tag{2.10}$$

The ℓ_1 norm is a convex function and the resulting problem (2.9) can be well posed as a linear programming problem [33].

While convex optimization approaches to recover sparse signals, including interior-point methods [31] and projected gradient method [34], are powerful methods for computing sparse representations and obtaining accurate estimation, there are a variety of greedy methods for solving CS recovery problems, which are often much faster than the convex optimization methods but have similar performance. The greedy algorithms rely on iterative approximation, either by iteratively obtaining an improved estimation of the coefficient vector such as Iterative Hard Thresholding (IHT) [35, 36], or by iteratively identifying the support of the coefficient vector such as Orthogonal Matching Pursuit (OMP) [37, 38], Compressed Sampling Matching Pursuit (CoSaMP) [39], Subspace Pursuit (SP) [40].

The core of the mentioned greedy algorithms is the thresholding operator $\mathbb{H}(\cdot)$ that sets all but the K entries of the input vectors with largest magnitudes to zero and returns the nearest K sparse approximations in terms of the ℓ_2 norm:

$$\mathbb{H}(c, K) = \arg\min_{c^* \in \Sigma_K} \|c - c^*\|_2 \tag{2.11}$$

The sparse approximation resulting from the thresholding operator finds optimal K sparse approximations for input vectors in the sense that the output vectors are K-sparse and the Euclidean distance between the inputs vectors and the output vectors are small.

Algorithm 1 Iterative Hard Thresholding

Input: measurement matrix Φ, basis or frame matrix Ψ, measurements y, sparsity K

Output: estimated signal $\widehat{x}$

1: Initialize: $\widehat{c} = 0$, $\widehat{x} = 0$, $\Upsilon = \Phi\Psi$
2: **repeat**
3: $\quad \widehat{c} = \mathbb{H}(\widehat{c} + \Upsilon^T(y - \Upsilon\widehat{c}), K)$
4: $\quad \widehat{x} = \Psi\widehat{c}$
5: **until** stop criterion is met

For example, as defined in Algorithm 1, IHT iterates a gradient descent step followed by thresholding until a convergence criterion is met. Other algorithms including OMP, CoSaMP and SP also implement the hard thresholding to search the optimal sparse approximation.

2.1.3 Model-Based Compressive Sensing

While classical CS processes signals by exploiting the fact that the signals can be described as sparse in some basis or frame, the locations of the nonzero entries of coefficient vectors often have underlying structure. Such structure can be captured by model-based CS, which reduces the freedom degree of sparse signals by permitting only certain entries to be nonzero.

A K-sparse coefficient vector x lies in the K-sparse model Σ_K, which is the union of all $\binom{N}{K}$ canonical subspaces $\mathcal{X}_S$. In contrast to sparsity, where there are no constrain on the support of coefficient vectors, structured sparsity endows sparse signals with an additional structure that allows only certain canonical subspaces and disallows others. If $\{S_1, S_2, \ldots, S_J\}$ is the set containing all allowed supports with $|S_j| = K$ for each $j = 1, 2, \ldots, J$, then a K-structured sparse signal x lies in the K-structured sparse model $\Xi_K \subset \mathbb{C}^N$, the union of the S canonical subspaces $\mathcal{X}_{S_1}, \mathcal{X}_{S_2}, \ldots, \mathcal{X}_{S_J}$:

$$\Xi_K = \bigcup_{s=1}^{S} \mathcal{X}_{\Lambda_s} \tag{2.12}$$

If a signal is known to be structured sparse, then the RIP constraint on the CS measurement matrix can be relaxed to a model-based RIP that (2.6) are required to hold only for structured sparse signals rather than all sparse signals [41, 42]. This prior knowledge reduces the required number of linear measurements to accurately recover the sparse signals to $M = \mathcal{O}(K + \log S)$, which can be a significant reduction from $M = \mathcal{O}(K \log(N/K))$ [18, 41].

To take advantage of the theory of model-based CS, several model-based CS recovery algorithms are derived by replacing the standard sparse approximation with a structured sparse approximation algorithm. In place of standard sparse approximation $\mathbb{H}(\cdot)$ that results from thresholding, which returns the best sparse approximation in the sparse signal set, model-based CS uses a structured sparse approximation $\mathbb{M}(\cdot)$ that returns the nearest sparse approximation in the structured sparse signal set with allowed support in terms of ℓ_2 norm:

$$\mathbb{M}(c, K) = \arg \min_{c^* \in \Xi_k} \|c - c^*\|_2. \tag{2.13}$$

In a similar way to sparse approximation, structured sparse approximation finds the optimal structured sparse approximations for input vectors in the sense that the output vectors are sparse with additional structures and the Euclidean distance between input vectors and output vectors are small.

Structured sparsity has recently been incorporated into IHT [41, 15], OMP [19], and CoSaMP [18]. As in Algorithm 2, a model-based IHT can be easily formulated by integrating a structured sparse approximation into the classical IHT algorithm in Algorithm 1.

Algorithm 2 Model-Based IHT

Input: measurement matrix Φ, basis or frame matrix Ψ, measurements y, sparsity K

Output: estimated signal $\widehat{x}$

1: Initialize: $\widehat{c} = 0$, $\widehat{x} = 0$, $\Upsilon = \Phi\Psi$.
2: **repeat**
3: $\quad \widehat{c} = \mathbb{M}(\widehat{c} + \Upsilon^T(y - \Upsilon\widehat{c}), K)$
4: $\quad \widehat{x} = \Psi\widehat{c}$
5: **until** stop criterion is met

2.2 Compressive Parameter Estimation

2.2.1 Introduction to compressive parameter estimation

The parametric models form another and a more general class of low dimensional signal model, where a K-dimensional continuous parameter $\theta \in \mathbb{R}^K$ can be identified that carries the relevant information about a signal $x \in \mathbb{C}^N$, which changes as a continuous (typically nonlinear) function of this parameter. Typically, parametric signals are defined via a mapping $\psi : \Theta \to X$ from a parameter space $\Theta \subseteq \mathbb{R}$ to a signal space $X \subseteq \mathbb{C}^N$ that connects a parameter $\theta \in \Theta$ and its parametric signal $x = \psi(\theta) \in X$. $\psi(\Theta) \subseteq X$ represents all parametric signals corresponding to all possible parameters.

Parameter estimation problems deal with the estimation of the underlying parameters from the observed signals, especially when the signals are contaminated by noise. Parameter estimation from a noisy signal $y = x + n$, where n denotes Additive White Gaussian Noise (AWGN), aims to find the nearest signal $\widehat{x} \in \psi(\Theta)$ to y and then invert the mapping to estimate the parameter values $\widehat{\theta}$. CS theory suggests that the distance between any two parametric signals can be approximately preserved by a random projection operator $\Phi \in \mathbb{R}^{M \times N}$ [43]. In other words, parameter estimation can be performed directly on noisy CS measurements $y = \Phi x + n$ without having to recover the full signal $\widehat{x}$ from y and then estimate the parameter $\widehat{\theta}$, when the mapping is known and available. However, this mapping often takes the form of a nonlinear manifold and therefore is complex to leverage.

Parametric dictionaries (PDs) have been used to perform parameter estimation directly on noisy CS measurements without recovering the full observed signals using sparse signal models. PDs are motivated by the fact that, in many practical application, the observed signals can be expressed or approximated by a linear combination of parametric signals with distinct unknown parameters, i.e.,

$$x = \sum_{k=1}^{K} c_k \psi(\theta_k). \tag{2.14}$$

By introducing a PD $\Psi \subseteq \psi(\Theta)$ as a collection of a samples from the set of parametric signals

$$\Psi = [\psi(\omega_1), \psi(\omega_2), \ldots, \psi(\omega_L)], \tag{2.15}$$

which correspond to a set of samples from parameter space $\Omega = \{\omega_1, \omega_2, \ldots, \omega_L\} \subseteq \Theta$, the signal can be written as the product of the PD and a sparse coefficient vector $x = \Psi c$, in the case that the sampling is large and dense enough so that unknown parameters are all contained in the sampling set, i.e., $\theta = \{\theta_1, \theta_2, \ldots, \theta_K\} \subseteq \Omega$. Therefore, finding the unknown parameters is reduced to finding at most K PD elements from Ψ whose linear combination corresponds to CS measurements that are close to the observed CS measurements, or, in other words, to obtaining an estimation of the support of the coefficient vector.

2.2.2 Issues in compressive parameter estimation

The PD-based compressive parameter estimation can be perfect only if the oversampling set of parameter space is dense and large enough to contain all of the unknown parameters. If this tough case is not met for some unknown parameter θ_k, a denser and larger sampling of the parameter space increases the chance that a observation $\psi(\omega_l)$ for some sampled parameter ω_l is sufficiently close to the observation $\psi(\theta_k)$ for the unknown parameter θ_k, i.e., $\|\psi(\omega_l) - \psi(\theta_k)\|_2$ is very small, such that we can estimate the unknown parameter θ_k by the sampled parameter $\omega_l \in \Omega$.

When the sampling step of the parameter space is Δ, it is easy to verify that the distance between the the nearest sampled parameter ω_l and the unknown parameter θ_k is bounded by half of the sampling step:

$$\min_{\omega_l \in \Omega} |\omega_l - \theta_k| \leq \frac{\Delta}{2}. \tag{2.16}$$

This shows that the small error of estimating the unknown parameter by the nearest sampled parameter requires the small sampling step or the dense sampling of the parameter space.

However, highly dense sampling increases the similarity between the PD elements for adjacent parameters and the coherence in the PD, which is measured by the maximum normalized inner product of PD elements:

$$\mu(\Psi) = \max_{1 \leq i \neq j \leq L} \frac{|\langle \psi(\omega_i), \psi(\omega_j) \rangle|}{\|\psi(\omega_i)\|_2 \|\psi(\omega_i)\|_2}. \tag{2.17}$$

The denser that the sampling is, the higher similarity that the dictionary elements for adjacent parameter have, and the closer that coherence $\mu(\Psi)$ is to one. This increases the difficulty of distinguishing between elements and severely hampers the performance of compressive parameter estimation [44, 17].

Alternatively, one can take advantage of structured sparsity to address the coherence issue [18]. In contrast with classical sparsity that searches for the sparse approximation among all sparse vectors, structured sparsity only searches for the sparse approximation among the sparse vectors that exhibit particular additional structure. One can use a coherence-inhibiting structured sparse approximation in which the resulting K nonzero entries of coefficient vectors correspond to PD elements that have low coherence, in order to inhibit the highly coherent PD elements from appearing in signal representation simultaneously [5, 15, 16, 11]. Such coherence-inhibiting framework has derived a variety of parameter estimation algorithms, such as structured

Algorithm 3 Coherence-Inhibiting Structured Sparse Approximation

Input: input vector c, parameter dictionary Ψ, maximum coherence ν, sparsity K
Output: estimated vector $\widehat{c}$

Initialize: $\widehat{c} = 0$, $D = \Psi^*\Psi$, $S = \emptyset$
for $k = 1$ to K **do**
    $j = \arg\max_i c_i$
    $\widehat{c}_j = c_j$
    $S = \{i : D_{i,j}\} \geq \mu$
    $c_S = 0$
end for

iterative hard thresholding (SIHT) [5, 15], band-exclusion orthogonal matching pursuit (BOMP) [19], and band-exclusion interpolating subspace pursuit (BISP) [16, 11].

Algorithm 3 shows the coherence-inhibiting structured sparse approximation, in which the coefficient with largest magnitude is kept and the coefficients corresponding to the elements that are highly coherent with the element of largest coefficient are vanished. The resulting estimate coefficient vector $\widehat{c}$ is K-sparse with the support $\widehat{S}$ corresponding to low coherent PD elements

$$\mu(\Psi_{\widehat{S}}) \leq \nu. \tag{2.18}$$

The maximum coherence ν in the algorithm is defined as a band width within the dictionary elements can not appear simultaneously in the signal representation. Although it is clear that an appropriate choice of maximum coherence can improve the performance of band-exclusion algorithms, little has been studied about the choice of maximum coherence and the sensitivity of the aforementioned algorithms to the maximum coherence. Intuitively, setting the parameter too low results in performance that is similar to that of standard algorithms, which is poor. Alternatively, setting the parameter too high results in wide band exclusion and thus strict requirements on the minimum separation of the parameters in signals, resulting in suboptimal performance for the signals that do not meet this requirement. Figure 2.1 shows the average estimation error of time delay estimation as a function of CS subsampling

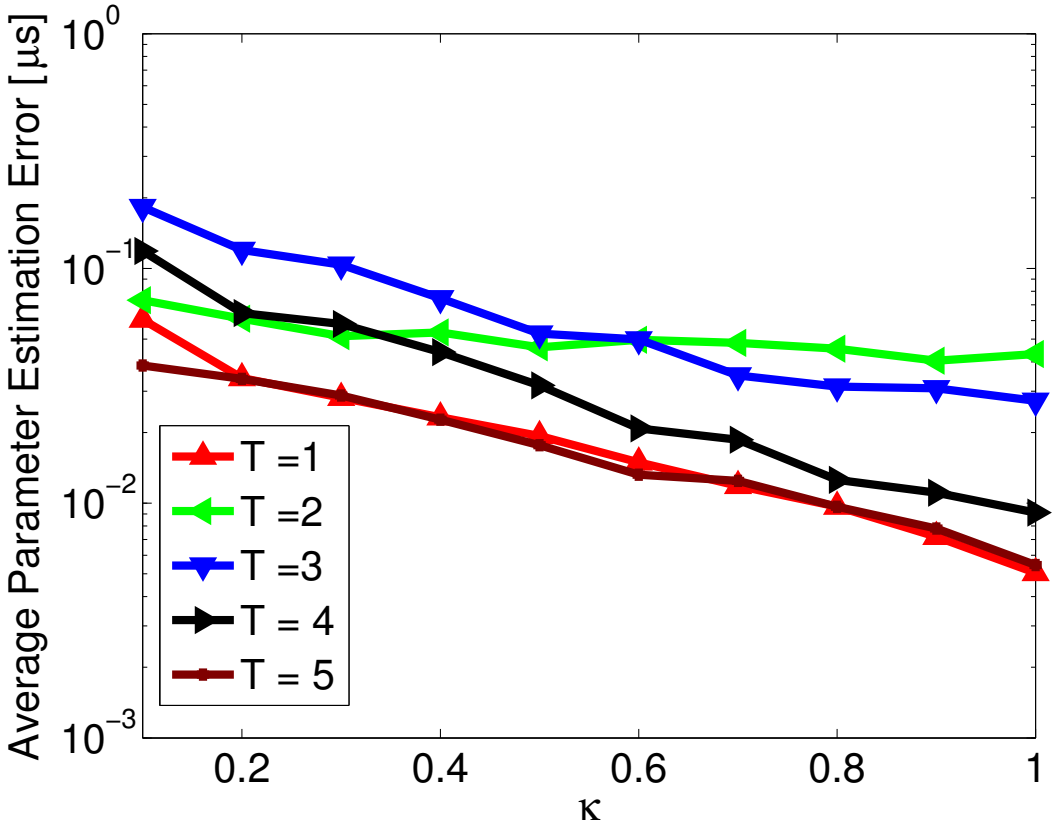


Figure 2.1: Average time delay estimation error as a function of subsampling rate with different chirp duration

rate κ when the maximum coherence is fixed at $\nu = 0.001$, which is the optimal value for the signals with chirp duration $T = 1$ μs, but the chirp duration T of the signal varies from 1 μs to 5 μs. Since the different chirp durations lead to different coherences of corresponding PDs, the band-exclusion algorithm with fixed maximum coherence have different performance on the signals with different chirp duration. The result confirms the conclusion that the performances of parameter estimation based on band-exclusion algorithms is very sensitive to the choice of maximum coherence.

The second issue in the PD-based compressive parameter estimation is that all proposed CS recovery algorithms guarantee stable recovery of coefficient vectors with the error being measured by the ℓ_2 norm, i.e., the estimated coefficient vector are close to the true coefficient vector in Euclidean distance. The guarantee is linked with the

core thresholding operation, which thresholds all entries of input vector to zero except for those with the largest magnitudes and returns the optimal sparse approximation to the input vector in terms of the ℓ_2 norm. However, the guarantee provides perfect parameter estimation that accurately estimates the support of the coefficient vectors only in the most demanding case of exact recovery, i.e., the estimated coefficient vector exactly match the true coefficient vector. Otherwise if the exact recovery can not be met, such a guarantee is meaningless since the ℓ_2 norm can not precisely measure the difference between the supports of two coefficient vectors. Consider a simple frequency estimation as in Figure 2.2, where the true coefficient vector is a canonical basis vector $c = e_i$, which corresponds to a signal with a single component at frequency f_i. Two candidate estimated coefficient vectors $\widehat{c}_1 = e_{i+1}$ and $\widehat{c}_2 = e_{i+7}$, which respectively correspond to signals with single components at frequencies f_{i+1} and f_{i+7}, share the same Euclidean distance to c, when the exact estimation is impossible. Nonetheless, the estimated frequncy $\widehat{f}_1 = f_{i+1}$ from $\widehat{c} = e_{i+1}$ has smaller error than $\widehat{f}_2 = f_{i+7}$ from $\widehat{c} = e_{i+7}$, when the Fourier transform is sorted so that $f_i < f_{i+1} < f_{i+7}$. This motivates the need for new metrics that can capture the difference between the two candidates and prefer the former over latter in the simple example.

There are several metrics that measure the distance between two vectors in terms of similarity between their supports. The Hamming distance measures the number of either both zeros or both nonzeros at the same entries of two vectors and has been considered in some certain algorithms [20, 21, 22, 23]. Unfortunately, The Hamming distance only controls the number of errors committed in parameter estimation, but not the magnitude of the errors that occur. As an alternative, the earth mover's distance quantifies the magnitudes of the errors by minimizing the amount and distance of flow among the entries of one vector to match another one [24, 25].

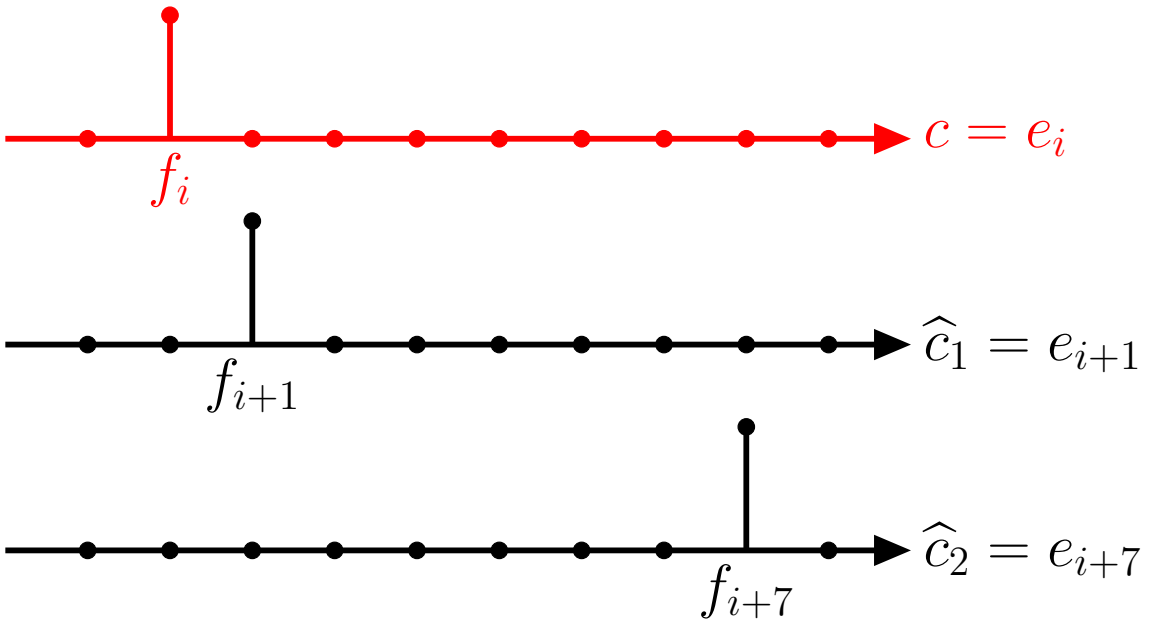


Figure 2.2: Example of frequency estimation

2.3 Earth Mover's Distance

The earth mover's distance (EMD) was first proposed by Rubner et al. as a crossbin distance to measure the similarity between signatures [24, 25]. Intuitively, if two signatures can be seen as earth and holes, the EMD measures the least amount of work needed to fill a collection of holes properly spread in space with a collection of earth spread in the same space, where a unit of work corresponds to transporting a unit of earth by a unit of ground distance.

The EMD between two vectors relies on the notion of mass being assigned to each entry of the two vectors involved, with the goal being to transfer mass among the entries of the first vector in order to match the mass of the entries of the second vector. The EMD captures the difference between two vectors by finding the flow with the smallest work (measured as the product of the amount of the flow to be moved and the distance of the flow) among all possible flow that is applied to the first vector to yield the second one.

More precisely, EMD between two vectors with the same ℓ_1 norm can be defined as the minimum flow work moving among the nonzero entries of the first vectors c to match the second vector $\widehat{c}$. If $S = \{s_1, s_2, \ldots, s_K\}$ and $\widehat{S} = \{\widehat{s}_1, \widehat{s}_2, \ldots, \widehat{s}_K\}$ are the supports of c and $\widehat{c}$ respectively, f_{ij} denotes the amount of flow moved from the entry s_i to the entry $\widehat{s}_j$, and d_{ij} measures the ground distance or the distance of flow f_{ij} to be moved:

$$d_{ij} = |s_i - \widehat{s}_j|, \tag{2.19}$$

then the EMD results from an optimization problem:

$$\begin{aligned} \mathrm{EMD}(c, \widehat{c}) &= \min_f \sum_{i,j} f_{ij} d_{ij} \\ \text{s.t.} \sum_j f_{ij} &= c_{s_i}, \quad i = 1, 2, \ldots, K; \\ \sum_i f_{ij} &= \widehat{c}_{\widehat{s}_j}, \quad j = 1, 2, \ldots, K; \\ f_{ij} &\geq 0, \quad i, j = 1, 2, \ldots K. \end{aligned} \tag{2.20}$$

In the case that the two vectors have different ℓ_1, one can always easily add an extra sink or an extra source to receive or provide the difference associated with a large ground distance to other entries.

That the EMD can be a metric for coefficient vectors in compressive parameter estimation leverages the fact that if the entries of the coefficient vectors are sorted by the corresponding parameter values, the distance of flow between any pair of entries is linear with the distance of corresponding parameters, so the EMD between the true coefficient vector and the estimated coefficient vector is indicative of the parameter estimation error.

Recently, there has been a significant interest in developing methods for geometric representations of EMD to provide approximated and simple computations of EMD. The goal is to find a mapping g so that the EMD between two vectors c and $\widehat{c}$ can

be easily approximated by the distance of their mapping values $g(c)$ and $g(\widehat{c})$. The mapping provided in [45, 46] guarantees that, for some constant $C > 0$,

$$\|g(c) - g(\widehat{c})\|_1 \leq \text{EMD}(c, \widehat{c}) \leq C\|g(c) - g(\widehat{c})\|_1. \tag{2.21}$$

Based on these results, it appears that the EMD between coefficient vectors can provide a approximate bound of the error between corresponding parameter values if the parametric model can be expressed as such mappings. Though the EMD has been recently integrated within CS to provide recovery algorithms for sparse and compressible signals, where the accuracy is measured in terms of the EMD [26, 27], no such guarantee has been proposed to show the relationship between the EMD of coefficient vectors and the error of corresponding parameters.

2.4 K-Median Clustering

Cluster analysis partitions data points based on the the information that describes the points and their similarity [47, 48]. Clustering is a task of partitioning a set of points into different groups in such 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 that the similarity within a group is or the greater that the difference among groups is, the better or more distinct that the clustering is.

The goal of clustering L points $\{p_1, p_2, \ldots, p_L\}$ associated with weights $w_1, w_2, \ldots, w_L$ and mutual similarity $d(p_i, p_j)$ into K clusters is to find K points $\{q_1, q_2, \ldots, q_K\}$, which are called centroids of the clusters, and then K clusters $\{C_1, C_2, \ldots, C_K\}$, each of which contains all points that are more similar to its centroid than other centroids:

$$C_i = \{p_l : d(p_l, q_i) \leq d(p_j, q_j)\} \tag{2.22}$$

To qualify how good the clustering is, the measure function or the objective function is defined as the total sum of the similarity between each point and its closet centroid:

$$J = \sum_{i=1}^{K} \sum_{p_j \in C_i} w_j d(q_i, p_j). \tag{2.23}$$

Different choices of similarity can result in different meaning for the centroids [48]. If the mutual similarity is measured as the squared Euclidean distance, the centroids will be the means of the clusters, and so the clustering is called K-mean clustering. If the similarity is represented by the Manhattan distance:

$$d(p_i, p_j) = |p_i - p_j|, \tag{2.24}$$

then the measure function defined in (2.23) becomes the sum of the ℓ_1 norm of each point to its nearest centroid:

$$J = \sum_{i=1}^{K} \sum_{p_j \in C_i} w_j |q_i - p_j|. \tag{2.25}$$

One can solve for the centroids by differentiating the measure function and setting it to zero:

$$\begin{aligned}
&\frac{\partial}{\partial q_i} J = 0 \\
\Rightarrow &\frac{\partial}{\partial q_i} \sum_{i=1}^{K} \sum_{p_j \in C_i} w_j |q_i - p_j| = 0 \\
\Rightarrow &\sum_{i=1}^{K} \sum_{p_j \in C_i} w_j \frac{\partial}{\partial q_i} |q_i - p_j| = 0 \\
\Rightarrow &\sum_{p_j \in C_i} w_j \frac{\partial}{\partial q_i} |q_i - p_j| = 0 \\
\Rightarrow &\sum_{p_j \in C_i} w_j \text{sign}(q_i - p_j) = 0.
\end{aligned} \tag{2.26}$$

Algorithm 4 K-Median Clustering $(Q, S) = \mathbb{C}(W, P, K)$

Input: weights W, points P, number of cluster K
Output: centroids Q, cluster indices S
1: Initialize: choose Q as random points.
2: **repeat**
3: $s_i = \arg\min_{j=1,\ldots,K} |p_i - q_j|$ for each $i = 1, \ldots, L$
4: $q_j = \arg_p \sum_{i:s_i=j} w_i \text{sign}(p - p_i) = 0$ for each $j = 1, \ldots, K$
5: **until** S does not change

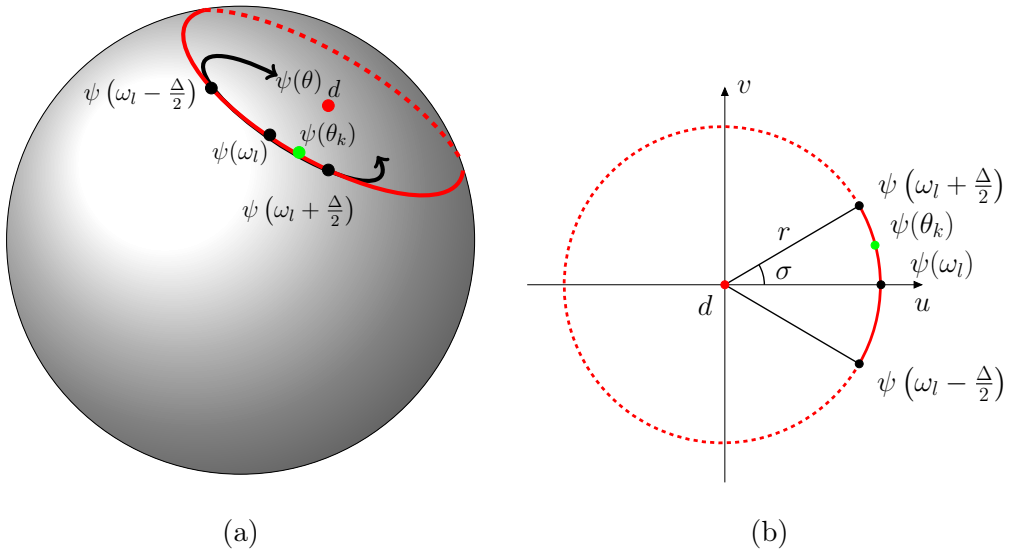
Where the sign$(\cdot)$ gets the sign of a number. Equation (2.26) illustrates that the resulting centroids are the medians of the clusters and the points on the different sides of the centroids have balanced weight:

$$\sum_{j:p_j \in C_i, p_j < q_i} w_j = \sum_{j:p_j \in C_i, p_j > q_i} w_j. \tag{2.27}$$

A well-known K-median clustering is formally described as Algorithm 4 [47], in which initial centroids are generated randomly. Then each point is assigned to the cluster with nearest centroid and each centroid is updated by the median of all points in the cluster repeatedly until the centroids do not change.

2.5 Polar Interpolation

A recently proposed alternative to improve the estimation performance of PD-based parameter estimation when the unknown parameters are not all contained in the sampling set of the parameter space is to use interpolation in the parameter space [49, 16, 11]. The motivation behind such approaches is that the low-dimensional parametric model that expresses the relationship between parameters and signals in a small neighborhood can be well approximated by a closed-form expression that integrates as much knowledge of the parametric model characteristics as possible while remaining computationally feasible. Therefore, the signal of a parameter that is outside of the sampling set can be accurately estimated from the signals of its surrounding



(a) (b)

Figure 2.3: Illustration of polar interpolation

sampled parameters via interpolation. Although Taylor series interpolation is a popular model, certain applications that feature parametric invariance of the norm and distances between signals are better suited to a polar interpolation scheme [49, 16, 11].

For many parameter estimation problems, including frequency estimation and time delay estimation, all parametric signals share the same magnitude and therefore are characterized by a curve contained in the surface of a high-dimensional hypersphere in $\mathbb{C}^N$. A small curve of this manifold can therefore be approximated by an arc of a circle on the high-dimensional hypersphere, which is uniquely determined by a triplet of PD elements corresponding to three sampled parameters contained in the segment. It is possible to find a basis for the span of the triplet of elements that provides a trigonometric map from the angle between the middle element and the observed signal and the differential of the parameter values for the two signals.

More specifically, assume that the parameter space is sampled with a step size Δ. As shown in Figure 2.3, The unknown parameter θ_k is linked to the closest value ω_l within the sampled set Ω; therefore, the unknown signal $\psi(\theta_k)$ lies on the

segment of the manifold $\{\psi(\theta) : \omega_l - \Delta/2 \leq \theta \leq \omega_l + \Delta/2\}$. This segment is approximated by the unique circular arc that contains the triplet of PD elements $\{\psi(\omega_l - \Delta/2), \psi(\omega_l), \psi(\omega + \Delta/2)\}$. The polar approximation is obtained as

$$\psi(\theta_k) \approx d(\omega_l) + r\cos\left(\frac{2(\theta_k - \omega_l)}{\Delta}\sigma\right) u(\omega_l) + r\sin\left(\frac{2(\theta_k - \omega_l)}{\Delta}\sigma\right) v(\omega_l), \tag{2.28}$$

where $d(\omega_l)$, $u(\omega_l)$, and $v(\omega_l)$ are a basis for the circle corresponding to its center and trigonometric coordinates and the constants r and σ represent the radius and the half-angle of the relevant circular arc. The approximation basis elements can be computed in closed form using the formula

$$\begin{aligned}&[d(\omega_l), u(\omega_l), v(\omega_l)] = \\ &\left[\psi\left(\omega_l - \frac{\Delta}{2}\right), \psi(\omega_l), \psi\left(\omega_l + \frac{\Delta}{2}\right)\right] \begin{bmatrix} 1 & 1 & 1 \\ r\cos(\sigma) & r & r\cos(\sigma) \\ -r\sin(\sigma) & 0 & r\sin(\sigma) \end{bmatrix}^{-1},\end{aligned} \tag{2.29}$$

which intuitively provides the mapping between the angles $\{-\sigma, 0, \sigma\}$ and the PD element triplet. When multiple parameters are observed simultaneously, we collect the estimation basis elements $d(\omega_l)$, $u(\omega_l)$ and $v(\omega_l)$ into the matrices D, U, and V so that the observed signal x can be expressed as as

$$x = \sum_{k=1}^{K} c_k \psi(\theta_k) \approx Dc + U\alpha + V\beta, \tag{2.30}$$

where c, α and β collect the trigonometric coefficients from the individual approximations (2.28). The solution to this equation can be obtained by posing a constrained convex optimization problem [49, 11]; the resulting coefficients α and β yield an estimate of the parameter via the bijective relation

$$\widehat{\theta}_k = \omega_l + \frac{\Delta}{2\sigma}\arctan\left(\frac{\beta_k}{\alpha_k}\right). \tag{2.31}$$

CHAPTER 3

CLUSTERING PARAMETER ESTIMATION

3.1 Estimation Error

The task of parameter estimation is to make the error between the true parameters and estimated parameters is as small as possible. Solving for the estimation error between a set of unknown parameter values $\theta = \{\theta_1, \theta_2, \ldots, \theta_K\}$ and a set of estimated parameter values $\widehat{\theta} = \{\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_K\}$ is an assignment problem that minimizes the cost of assigning each true parameter value to an estimated parameter value. If the cost of assigning the true parameter θ_i to the estimated parameter $\widehat{\theta}_j$ is the ground distance between the two parameters:

$$t_{ij} = |\theta_i - \widehat{\theta}_j|, \tag{3.1}$$

and binary value $g_{ij} \in \{0, 1\}$ denotes the status of the assignment:

$$g_{ij} = \begin{cases} 1 & \theta_i \text{ is assigned to } \widehat{\theta}_j \\ 0 & \theta_i \text{ is not assigned to } \widehat{\theta}_j \end{cases}, \tag{3.2}$$

then the parameter estimation error (PEE) results from the integer programming problem:

$$
\begin{aligned}
\mathrm{PEE}(\theta, \widehat{\theta}) &= \min_{g} \sum_{i,j} g_{ij} t_{ij} \\
\text{s.t.} \sum_{j} g_{ij} &= 1, \quad i = 1, 2, \ldots, K; \\
\sum_{i} g_{ij} &= 1, \quad j = 1, 2, \ldots, K; \\
g_{ij} &\in \{0, 1\}, \quad i, j = 1, 2, \ldots, K.
\end{aligned} \tag{3.3}
$$

Collecting all g_{ij} and t_{ij}, we have the vectors:

$$
g = [g_{11}, g_{12}, \ldots, g_{1K}, g_{21}, g_{22}, \ldots, g_{KK}]^T \tag{3.4}
$$

and

$$
t = [t_{11}, t_{12}, \ldots, t_{1K}, t_{21}, t_{22}, \ldots, t_{KK}]^T. \tag{3.5}
$$

Providing that $\mathbf{1}_K^T$ and I_K denotes the K-dimensional vector whose entries are all 1 and the $K \times K$ identity matrix, respectively, and

$$
A = \begin{bmatrix} I_K \otimes \mathbf{1}_K^T \\ \mathbf{1}_K^T \otimes I_K \end{bmatrix}, \tag{3.6}
$$

where $\otimes$ is the Kronecker tensor product, (3.3) can be written in the canonical form of integer programming:

$$
\begin{aligned}
\mathrm{PEE}(\theta, \widehat{\theta}) &= \min_{g} t^T g \\
\text{s.t.} Ag &= \mathbf{1}_{2K}^T \quad . \\
g &\in \{0, 1\}^{K^2}
\end{aligned} \tag{3.7}
$$

It is easy to show that A is a total unimodular matrix, whose square non-singular submatrices are all integer matrices with determinant 1 or -1. So the integer programming problem (3.7) has the equivalent answer as its linear programming [50]:

$$\begin{aligned} \mathrm{PEE}(\theta, \widehat{\theta}) &= \min_{g} t^T g \\ \text{s.t.} Ag &= \mathbf{1}_{2K}^T \\ g &\geq 0 \end{aligned} \quad . \tag{3.8}$$

In PD-based compressive parameter estimation, based on the fact that there exists a bijective relationship between parameters and entries of the coefficient vectors, it will be possible to control the parameter estimation error by controlling the estimation error of the supports of the coefficient vectors. The main reason to discard the ℓ_2 norm used to measure the recovery error in previously proposed methods is that ℓ_2 norm of the coefficient vectors can not really measure the difference of their nonzero entries, which is the PD-based compressive parameter estimation to minimized.

The EMD can be a potential metric for the estimation error of the coefficient vectors due to the fact that the distance of flow moving from one entry to another is proportional to the distance between the corresponding parameters. When the sampling step of the parameter space is denoted by Δ, the distance between the entries s_i and $\widehat{s}_j$ of the coefficient vectors and the distance between the parameters θ_i and $\widehat{\theta}_j$ have such relationship:

$$t_{ij} = |\theta_i - \widehat{\theta}_j| = \Delta |s_i - \widehat{s}_j| = \Delta d_{ij}. \tag{3.9}$$

If vectors f and d collect f_{ij} and d_{ij} in the same way as (3.4) and (3.5), and b collects all nonzero entries of c and $\widehat{c}$ as

$$b = [c_{s_1}, c_{s_2}, \ldots, c_{s_K}, \widehat{c}_{\widehat{s}_1}, \widehat{c}_{\widehat{s}_2}, \ldots, \widehat{c}_{\widehat{s}_K}]^T, \tag{3.10}$$

the equation (2.20) for solving EMD can also be simplified as the canonical form:

$$\begin{aligned} \mathrm{EMD}(c, \widehat{c}) = \min_{f} \; & d^T f \\ \text{s.t.} \; & Af = b \\ & f \geq 0 \end{aligned} \quad . \tag{3.11}$$

Based on the similarity of (3.8) and (3.11), It is straightforward to formulate the following theorem, which is proved in Appendix A.

Theorem 3.1.1. *Assume that Δ is the sampling step of the parameter space. If c and $\widehat{c}$ are the coefficient vectors corresponding to two sets of parameters θ and $\widehat{\theta}$, then the EMD between the two coefficient vectors provide an upper bound of the parameter estimation error between the two sets of parameters:*

$$PEE(\theta, \widehat{\theta}) \leq \frac{\Delta}{c_m} EMD(c, \widehat{c}), \tag{3.12}$$

where c_m is the smallest magnitude among the entries of c and $\widehat{c}$.

To illustrate the theorem, 100 pairs of sets of parameters are randomly sampled from the parameter space with the sampling step $\Delta = 1$ and the corresponding coefficient vectors are generated with minimum magnitude $c_m = 0.5$. Each point has the coordinates of the EMD of the coefficient vectors and the PEE of the parameters. As shown in Figure 3.1, all points are below the line, which represents $\mathrm{PEE}(\theta, \widehat{\theta}) = (\Delta/c_m)\mathrm{EMD}(c, \widehat{c})$.

Following the aforementioned analysis, if the EMD-optimal sparse approximation is available to provide the guarantee on the stable recovery of coefficient vectors in terms of EMD, one can can potentially provide a simple extension of the guarantee to PD-based compressive parameter estimation.

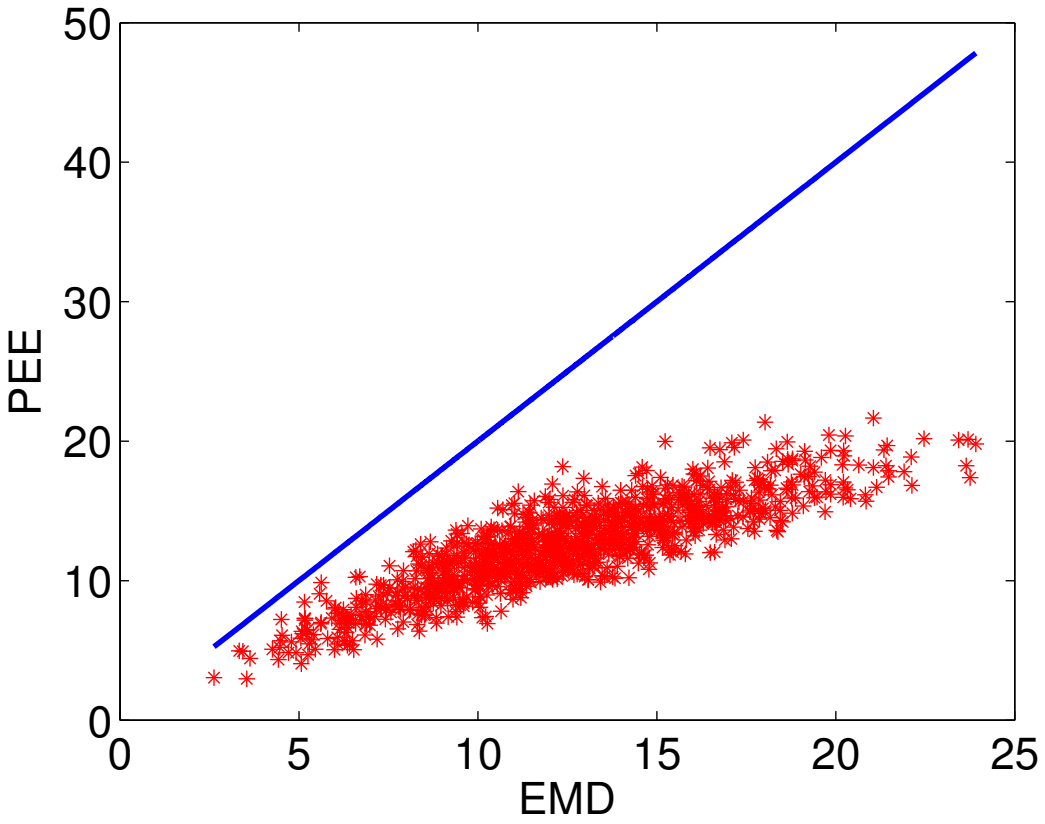


Figure 3.1: PEE as a function of EMD

3.2 EMD-Optimal Sparse Approximation

An EMD-optimal sparse approximation, which stably finds the optimal sparse approximation in the sense that the EMD from the input vector to the output vector is small, plays the crucial role in our proposed compressive parameter estimation. Motivated by the aforementioned algorithms such as SIHT or BOMP, in order to integrate EMD into the CS framework, an EMD-optimal sparse approximation should be formulated to provide a best sparse approximation to the input vector with error being measured in terms of EMD.

In the case that one want to find a K-sparse vector $\widehat{c} \in \mathbb{C}^L$ with fixed support $\widehat{S} \subset I = \{1, 2, \ldots, L\}$ that has smallest EMD to an arbitrary vector $v \in \mathbb{C}^L$, the minimum flow work defined in (2.20) is achieved when the flow is only active between

each entry of the vector v and its nearest nonzero entry $\widehat{s}_i$ of the vector $\widehat{c}$, i.e., the entry $\widehat{s}_i$ has a connection to all the entries of v that have smaller distance to $\widehat{s}_i$ than to other nonzero entries of $\widehat{c}$. In other words, K nonzero entries of $\widehat{c}$ partition the entries of v into K different groups:

$$V_i = \{l \in I : |l - \widehat{s}_i| \leq |l - \widehat{s}_j|\}, \tag{3.13}$$

and the EMD defined in (2.20) yields

$$\text{EMD}(v, \widehat{c}) = \sum_{i=1}^{K} \sum_{j \in V_i} |v_i||j - \widehat{s}_i|. \tag{3.14}$$

It is important to note that objective function (3.14) matches the objective function (2.25). If entries of v represent L points associated with weight $|v_1|, |v_2|, \ldots, |v_L|$, and support $\widehat{S}$ represent the centroids resulting from performing K-median clustering on the entries, then the EMD between v and $\widehat{c}$ equals to the objective function of K-median clustering, which is minimized. So K-median clustering is able to returns the nearest sparse vector to an arbitrary vector in terms of EMD, which is the task of EMD-optimal sparse approximation.

As we mentioned before, the goal of the compressive parameter estimation is to find a number of PD elements, whose linear combination has the linear measurements that are close to the observed linear measurements. To find those PD elements, greedy algorithms project the observed measurements into the space spanned by all PD elements and obtain the correlation values between the observed measurements and all PD elements, which are the inner product of the observed measurements and PD elements. Obviously, the PD elements with large correlation values are the elements that we expect.

In the limiting case that the sampling step of the parameter space $\Delta \to 0$, both the sampling of signal space and parameter space are extremely dense and therefore

signals are continuous functions of a continuous parameter. The PD element corresponding to parameter $\theta \in \mathbb{R}$ can be expressed as $\psi(\omega - \theta)$, which is a shifted version of a smooth and unit-energy function $\psi(\omega)$. A parametric signal $x(\omega)$ involving the unknown parameters $\theta_1, \theta_2, \ldots, \theta_K$ has the form as

$$x(\omega) = \sum_{i=1}^{K} c_i \psi(\omega - \theta_i), \tag{3.15}$$

where $c_i > 0$ is the magnitude. Let the auto-correlation function be defined as

$$\lambda(\theta) = \langle \psi(\omega - \theta), \psi(\omega) \rangle = \left| \int_{\mathbb{R}} \psi^*(\omega - \theta) \psi(\omega) \mathrm{d}\omega \right|. \tag{3.16}$$

When there is no compressive sensing and noise, where the parametric signal is the observed linear measurements, the correlation function between measurements and PD elements can be expressed as a linear combination of the auto-correlation functions:

$$\begin{aligned} v(\theta) &= \langle \psi(\omega - \theta), x(\omega) \rangle \\ &= \langle \psi(\omega - \theta), \sum_{i=1}^{K} c_i \psi(\omega - \theta_i) \\ &= \sum_{i=1}^{K} c_i \langle \psi(\omega - \theta), \psi(\omega - \theta_i) \rangle \\ &= \sum_{i=1}^{K} c_i \langle \psi\left(\omega - (\theta - \theta_i)\right), \psi(\omega) \rangle \\ &= \sum_{i=1}^{K} c_i \lambda(\theta - \theta_i). \end{aligned} \tag{3.17}$$

In most parameter estimation problems, the auto-correlation function $\lambda(\theta)$ has bounded variation such that the cumulative auto-correlation function, defined as

$$\Lambda(\theta) = \int_{-\infty}^{\theta} \lambda(\omega) \mathrm{d}\omega, \tag{3.18}$$

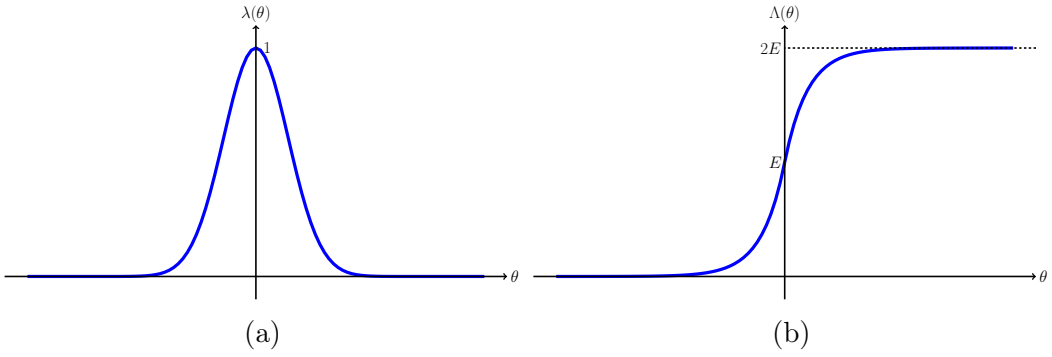


Figure 3.2: (a) auto-correlation function and (b) cumulative auto-correlation function

is bounded and has the supremum at infinity, i.e., $\Lambda(\infty) = 2E$. As shown in Figure 3.2b, $\Lambda(\theta)$ is a monotonic increasing function with some other important properties as $\Lambda(0) = E$ and $\Lambda(-\theta) + \Lambda(\theta) = 2E$.

Due to the smoothness and unit energy of $\psi(\omega)$, $\lambda(\theta)$ is a continuous function and reaches the maximum value 1 at $\theta = 0$, as shown in Figure 3.2a. So that the correlation function has the local maximum $c_1, c_2, \ldots, c_K$ at the parameters $\theta_1, \theta_2, \ldots, \theta_K$, which exactly match the unknown parameters. So the greedy algorithms estimated the parameters by finding the parameters corresponding to the local maximum of correlation function.

Another property of the auto-correlation function is that the function decreases as the distance of the parameter to zero increases until finally vanishing, since the coherence between PD elements decreases as the distance of their parameters increases. When the auto-correlation function decays very fast, where the PD elements are almost incoherent and the coherence in PD is very small, it is possible to find the local maximum of correlation function by the thresholding operator. When the auto-correlation function decays slowly, where the PD elements are highly coherent, the thresholding operator will unavoidably find the values around the global maximum that are larger than other local minimums if additional approaches like the band

exclusion are not implemented. Instead, the K-median clustering tries to locate the local minimum directly.

There are some conditions that the auto-correlation function should satisfy to ensure small estimation error performing K-median clustering on the correlation function. First, any pair of parameters must be well separated. If two parameters θ_i and θ_j are too close to each other, the similarity of $\psi(\omega - \theta_i)$ and $\psi(\omega - \theta_j)$ makes it difficult to distinguish them. So the conditions should contain the minimum separation distance:

$$\zeta = \min_{i \neq j} |\theta_i - \theta_j|. \tag{3.19}$$

Second, any parameter should be far away from the bound of the sampling range. Often it is enough and convenient to estimate parameter in a range $[\theta_{\min}, \theta_{\max}] \subset \mathbb{R}$ rather than the entire real number. According to (2.27), which implies that a centroid balances the weights of points on different sides of the cluster, there will be a bias in the estimation to compensate the missed weight of the correlation function on one side when one parameter is too close to the bound of parameter range. Therefore the conditions should also contain the minimum off-bound distance ϵ so that

$$\theta_{\min} + \epsilon \leq \theta_i \leq \theta_{\max} - \epsilon. \tag{3.20}$$

Third, if the magnitudes of some parameters are too small, K-median clustering may regard them as the noisy and ignore them. So another condition should be the dynamic range of component magnitudes:

$$r = \max_{i,j} \frac{c_i}{c_j}. \tag{3.21}$$

Combining all these conditions, we can state the following theorem to guarantee the performance of K-median clustering, which is proved in Appendix B.

Theorem 3.2.1. *Assume that the signal x in a parameter estimation problem has K parameters $\theta_1, \theta_2, \ldots, \theta_K$ with magnitude dynamic range r. For any error tolerance $\delta > 0$, if the minimum separation distance satisfies*

$$\zeta \geq 2\Lambda^{-1}\left(2E\left(1 - \frac{\Lambda(\delta)/E - 1}{(2K-2)r+1}\right)\right) + 2\delta, \tag{3.22}$$

and the minimum off-bound distance satisfies

$$\epsilon \geq \Lambda^{-1}\left(2E\left(1 - \frac{\Lambda(\delta)/E - 1}{(2K-2)r+1}\right)\right), \tag{3.23}$$

then the estimation error using K-median clustering is bounded by the error tolerance:

$$|\theta_k - \widehat{\theta}_k| \leq \delta, \quad k = 1, 2, \ldots, K \tag{3.24}$$

Theorem 3.2.1 provides a very important insight that the smaller that the target error tolerance δ is, the larger that the minimum separation ζ and the minimum distance to bound ϵ will need to be. For a simple example, assume that auto-correlation function has the form of Gaussian function as

$$\lambda(\theta) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\theta^2}{2\sigma^2}\right) \tag{3.25}$$

and $K = 2$, $r = 1.5$. When $\delta = 0.1\sigma$, $\zeta = 3.50\epsilon$ and $\epsilon = 1.65\sigma$; when $\delta = 0.01\sigma$, $\zeta = 4.39\sigma$ and $\epsilon = 2.19\sigma$.

It is worth noting that the theorem focuses on the worst case in practice. Figure 3.3 gives both maximum estimation error from the theorem and the experiments with randomly generated data as a function of minimum separation. It is remarkable that the required minimum separation distance in practice is much less than the theoretic value. Nevertheless, the theorem still help a lot in the theoretical analysis of parameter estimation problems.

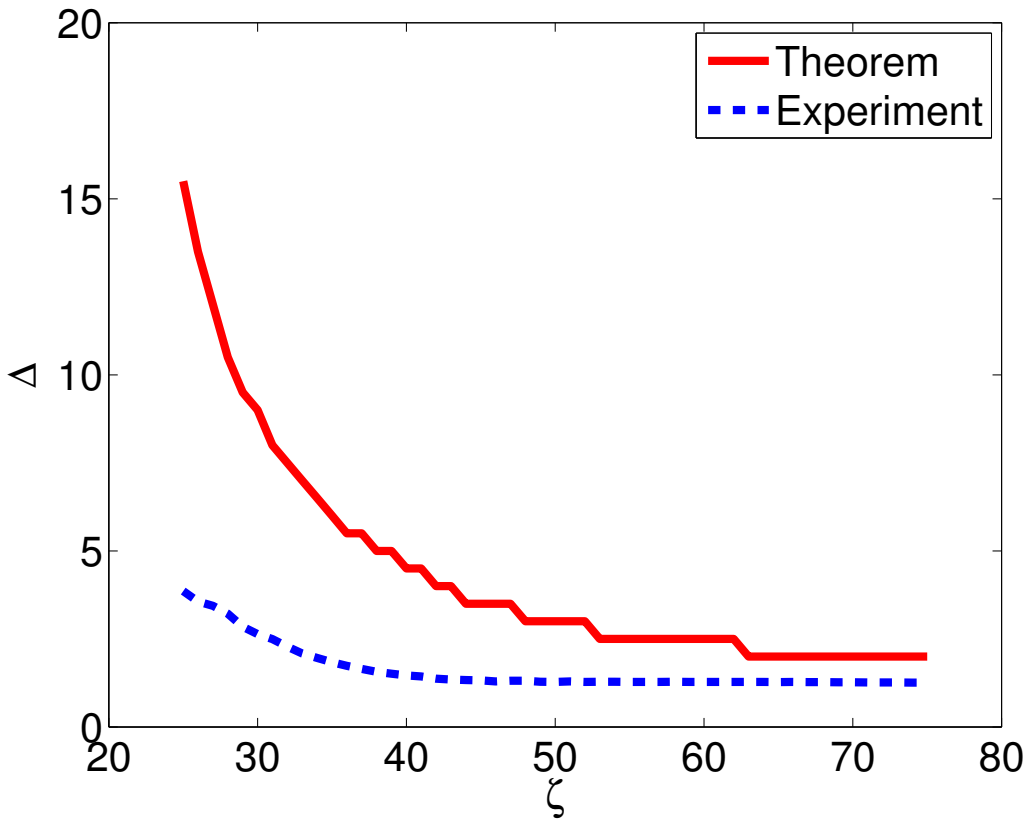


Figure 3.3: Theoretical and experimental maximum error as a function of minimum separation

3.3 Parameter Estimation Algorithm

Since most existing sparse recovery algorithms rely on a thresholding operation to obtain optimal sparse approximations in the ℓ_2 metric, it is particularly easy to modify the existing recovery algorithms to achieve sparse recovery with EMD. For example, we propose a new PD-based parameter estimation algorithm, called Clustering Subspace Pursuit (CSP), as shown in Algorithm 5. CSP merges the Subspace Pursuit algorithm [40] with EMD-based sparse approximation and replaces the thresholding steps from subspace pursuit by K-median clustering to find estimated support S from a residual (or proxy) coefficient vector v.

As mentioned, CSP will have the potential to provide the EMD-optimal error guarantee for coefficient vectors, which inherits from K-median clustering. Additionally,

Algorithm 5 Clustering Subspace Pursuit (CSP)

Input: measurement vector y, measurement matrix Φ, parametric dictionary Ψ, sparsity K, parameter sampling set Ω

Output: estimated signal x, estimated parameter values θ

Initialize $x = 0$, $\Sigma = \emptyset$.
repeat
  $v = (\Phi\Psi)^T(y - \Phi x)$
  $S = S \cup \mathbb{C}(v, K)$
  $c = (\Phi\Psi_\Sigma)^+ y$
  $S = \mathbb{C}(c, K)$
  $x = \Psi_S c$
  $\theta = \Omega_S$
until a convergence criterion is met

CSP has the ability to avoid highly coherent PD elements from appearing simultaneously in the signal representation without the requirement for a coherence-inhibiting parameter ν that is required when structured sparsity is used. The reason is that the entries of the coefficient vector corresponding to the highly coherent PD elements are close to each other and will be assigned into the same cluster. All those PD elements are represented by the elements of the centroid and cannot appear simultaneously.

CHAPTER 4

RESULTS

To evaluate the performance of our proposed approach, we consider the time delay estimation problem where the signals are measured using a CS measurement matrix. The continuous signal model is a chirp waveform with time delay s defined as

$$g(t,s) := p(t-s)\exp\left(j2\pi\left(f_0 + f_\Delta\frac{t-s}{2T}\right)(t-s)\right)$$

where $f_0 = 1$ MHz is the chirp center frequency, $f_\Delta = 5$ MHz is the sweep frequency, and $p(t)$ is a raised cosine pulse that windows the chirp signal in time:

$$p(t) = \begin{cases} 1+\cos(2\pi t/T), & t \in (0,T), \\ 0 & \text{otherwise.} \end{cases}$$

Here, $T = 1$ μs is the duration of the chirp signals. We sample the chirp signals with a frequency $f_s = 50$ MHz and collect $N = 500$ samples of each continuous signal to generate the discrete signal g_s, whose samples can be written as

$$g_s[n] = \frac{1}{\sqrt{1.5Tf_s}}g\left(\frac{n-1}{f_s}, s\right), \quad n = 0, 1, \ldots, N-1,$$

where the coefficient $1/\sqrt{1.5Tf_s}$ normalizes the discrete signals. The observed signals can then be written as $x = \sum_{i=1}^{K} a_k g_{s_k}$, where the parameters s_k are selected at arbitrary resolution from the range $[0, 10\mu\text{s}]$, with a minimum separation distance $\zeta = 1$ μs and a minimum off-bound distance $\epsilon = 0.5$ μs, and a_k is the magnitude

for the k^{th} component. We generate a PD for this problem by sampling the the time delay (i.e., parameter) space with a spacing of $T_s = 0.02$ μs (matching the sampling period), which can be written as

$$\Psi = [g_0, g_{T_s}, g_{2T_s}, \ldots, g_{(N-1)T_s}]. \tag{4.1}$$

The signal x is then sensed using a random demodulator [51] simulated by an $M \times N$ CS matrix Φ for a variety of values of M.

Our experiments compare the time delay estimation performance of CSP to that of two existing baseline algorithms designed for high coherence PDs: band-excluding subspace pursuit (BSP) and band-excluded orthogonal matching pursuit (BOMP) [19]. Furthermore, we integrate polar interpolation within these algorithms to accommodate arbitrary values for the delay outside of the sampled set [49, 16]; note in particular that the BSP+Polar algorithm is equivalent to BISP [16, 11]. We set the maximum allowed coherence to $\nu = 0.001$ for all structured sparsity (band-excluded) algorithms.

Our first experiment considers compressive time delay estimation from noiseless measurements as a function of the CS subsampling rate $\kappa = M/N$ (Figure 4.1a), as well as for noisy measurements under AWGN with fixed subsampling rate $\kappa = 0.4$ as the function of the SNR level (Figure 4.1b). Both figures shows the performance of the algorithms in these setups averaged over 1000 randomized realizations. While the performance of CSP does match that of the band-excluding algorithms when no interpolation is used, there is a significant improvement in estimation performance when polar interpolation is added to the algorithms. In this case, BOMP estimates only one time delay at a time; the interference from the remaining copies of the delayed signal can cause noticeable errors in the interpolation stage.

Our next experiment evaluates the role that the maximum allowed coherence ν has on the performance of the algorithms. To verify this parameter, we vary the minimum

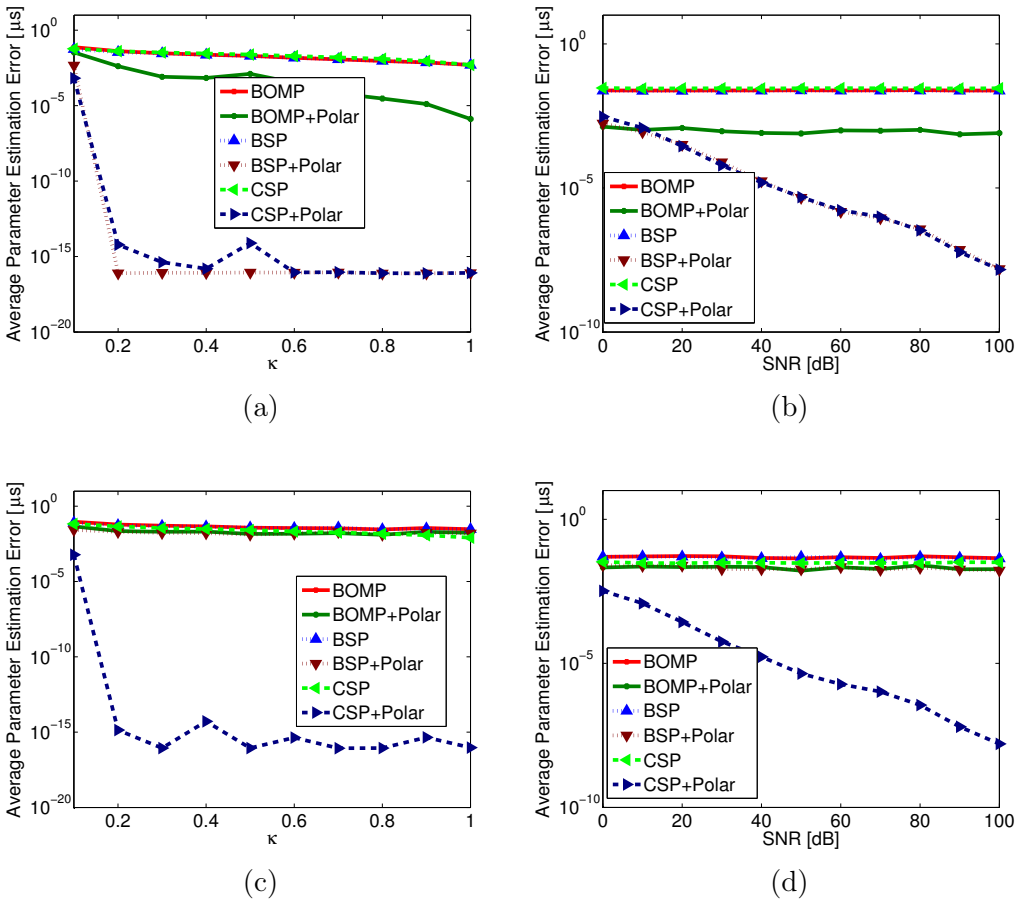


Figure 4.1: Average delay estimation error as a function of the CS sampling rate κ for noiseless measurements, where the minimum separation (a) $\zeta = 1$ μs and (c) $\zeta = 0.5$ μs, and of the SNR level with $\kappa = 0.4$, where the minimum separation (b) $\zeta = 1$ μs and (d) $\zeta = 0.5$ μs

separation distance $\zeta = 0.5\mu$s. We expect that the band-exclusion algorithms will be sensitive to the fixed choice of the maximum allowed coherence ν. Figure 4.1c and Figure 4.1d replicate the setups in Figure 4.1a and Figure 4.1b respectively except for the minimum separation distance and shows decreased performance for all algorithms except for CSP and CSP+Polar. Clearly, the drop in performance in the band-exclusion algorithms is due to a suboptimal choice of the parameter ν for the time delay problems that feature the decreased minimum separation. This gap

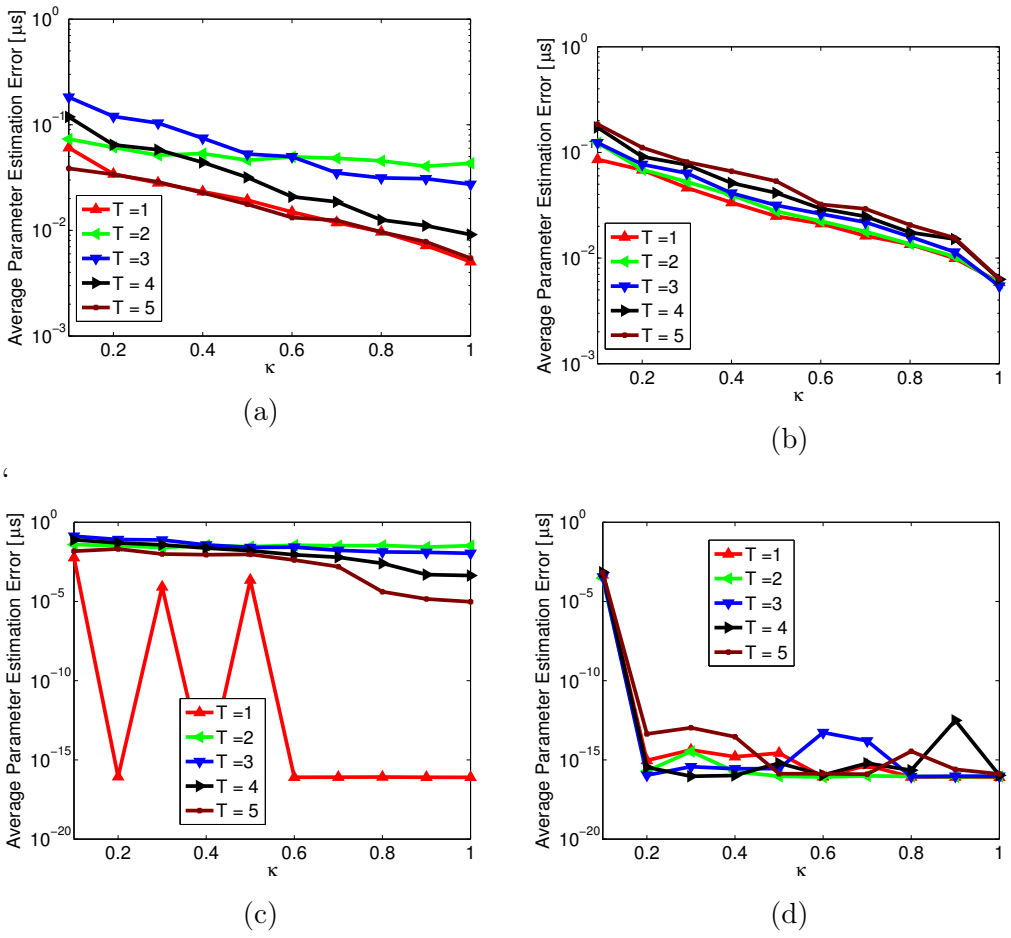


Figure 4.2: Average time delay estimation error of (a) BSP, (b) CSP, (c) BSP+Polar, and (b) CSP+Polar as a function of the CS subsampling rate κ for several chirp durations T

in performance is expected to become more significant as the minimum separation between the delays (or parameters in general) becomes smaller.

Our last experiment further focuses on this sensitivity on the choice of the maximum allowed coherence parameter ν for BSP in contrast with CSP. We test the performance of these algorithms as a function of the CS subsampling rate κ, with and without polar interpolation, for a variety of chirp duration lengths. Figure **??** shows the average performance over 100 randomized realizations per setup, and shows both a range of degradation levels in BSP performance and a sharp degradation in

BSP+Polar performance as the chirp duration varies. This is in contrast to CSP and CSP+Polar in Figure 4.2 , whose performances are essentially stable over the choice of PD.

CHAPTER 5

CONCLUSION

In this project, based on compressive sensing and parameter estimation, we prove that the earth mover's distance (EMD) is a potential metric to measure the difference of the coefficient vectors, and the EMD of the coefficient vectors can approximate the parameter estimation error of the corresponding parameters. Also we prove that the K-median clustering can directly locate the local maximum points of the correlation function of the observed signals and the parametric dictionary elements, which are exactly the unknown parameters, and therefore K-median clustering can serves as the EMD-optimal sparse approximation to output the nearest sparse vector to a input vector in terms of EMD. The proposed algorithms for compressive parameter estimation resulting from incorporating K-median clustering into the standard recovery algorithms shows its advantage to prevent the highly coherent dictionary elements from appealing simultaneously without additional parameters, such as the maximum coherence required by the band-exclusion algorithms, and its ability to accurate estimate the unknown parameters in different cases.

In future, it is expected to apply the new method to other parameter estimation problems, where more sophisticated analysis is needed. The theorems derived for time delay estimations is base on the fact that all unknown parameters have dimension 1 and all coefficients in the signal representation are non-negative. For other parameter estimation problems with 1-dimensional parameters and both negative and positive coefficients, such as the frequency estimation, the difference between the resulting correlation function and previous correlation function required a modified theorem to

stated the precise conditions for accurate estimations. For the parameter estimation with parameters in more than 1 dimension space, such as the localization and bearing estimation, new methods are needed to locate the local maximum of the resulting function, which will be a function on a multi-dimension parameter space and can not be efficiently handled by K-median clustering. Besides, approaches to deal with the noisy in correlation function, which comes from the noisy in observed signals and the subsampling in compressive sensing, can improve the robust of the algorithms to noisy and subsampling.

APPENDIX A

PROOF OF THEOREM 3.1.1

Assume that f^* and g^* respectively solve the optimization problem in (3.11) and (3.8) and $r = f^* - c_m g^*$. Then from (3.11), we have

$$\begin{aligned} \mathrm{EMD}(c, \widehat{c}) &= d^T f^\star \\ &= d^T (c_m g^\star + r) \\ &= c_m d^T g^\star + d^T r \\ &= \frac{c_m}{\Delta} t^T g^\star + d^T r \\ &= \frac{c_m}{\Delta} \mathrm{PEE}(\theta, \widehat{\theta}) + d^T r. \end{aligned} \tag{A.1}$$

Obviously, the first term in (A.1) answers the optimization problem (3.11) when b is the vector with all entries having magnitude c_m. The second term compensates the distance when the magnitudes increase from c_m, which should be non-negative.

Note that $f_{ij}^\star \geq 0$, and $g_{ij}^\star \in \{0, 1\}$ for any $i, j = 1, 2, \ldots, K$. When $g_{ij}^* = 0$, $r_{ij} = f_{ij}^* - c_m g_{ij}^* \geq 0$. When $g_{ij} = 1$, since the amount of flow f_{ij} from the entry s_i to the entry $\widehat{s}_j$ increases from c_m as the magnitudes on both entries increase from c_m to c_i and $\widehat{c}_j$,respectively, we have that $r_{ij} = f_{ij}^\star - c_m g_{ij}^\star \geq 0$. So $d^T r \geq 0$ due to the fact that d has non-negative entries.

Then it is possible to rewrite (A.1) as

$$\mathrm{EMD}(c, \widehat{c}) \geq \frac{c_m}{\Delta} \mathrm{PEE}(\theta, \widehat{\theta}), \tag{A.2}$$

and prove the theorem. □

APPENDIX B

PROOF OF THEOREM 3.2.1

Without loss of generality, assume that parameter values are sorted so that

$$\theta_{\min} + \epsilon \leq \theta_1 < \theta_2 < \cdots < \theta_K \leq \theta_{\max} - \epsilon, \tag{B.1}$$

For any $\sigma > 0$, there exists $D > 0$ such that

$$\Lambda(D) = 2E(1 - \sigma). \tag{B.2}$$

Since $\Lambda(\theta)$ is monotonically increasing and $\Lambda(-\theta) + \Lambda(\theta) = 2E$, then

$$\Lambda(\theta) \in \begin{cases} [0, 2E\sigma] & \theta \in (-\infty, -D] \\ [2E(1-\sigma), 2E] & \theta \in [D, \infty) \end{cases} \tag{B.3}$$

Assume that the estimate error is bounded by $e > 0$ such that,

$$-e \leq \theta_k - \widehat{\theta}_k \leq e. \tag{B.4}$$

When K-median clustering partitions the correlation function $v(\theta)$ into K groups according to $\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_K$, the point $(\widehat{\theta}_i + \widehat{\theta}_{i+1})/2$ is the bound for both cluster i and

cluster $i+1$, since the point has the same distance to both centroids. Using (B.4), we have

$$-e \leq \frac{\theta_k + \theta_{k+1}}{2} - \frac{\widehat{\theta}_k + \widehat{\theta}_{k+1}}{2} \leq e. \tag{B.5}$$

Assume that the minimum off-bound distance

$$\epsilon \geq D, \tag{B.6}$$

with the relationship in (B.1), it is obvious that $\theta_{\min} - \theta_k \leq \theta_{\min} - \theta_1 \leq -\epsilon \leq -D$ which further leads to

$$0 \leq \Lambda(\theta_{\min} - \theta_k) \leq 2E\sigma, \tag{B.7}$$

and $\theta_{\max} - \theta_k \geq \theta_{\max} - \theta_K \geq \epsilon \geq D$, which also leads to

$$2E(1-\sigma) \leq \Lambda(\theta_{\max} - \theta_k) \leq 2E. \tag{B.8}$$

Assume again that the minimum separation distance

$$\zeta \geq 2D + 2e. \tag{B.9}$$

When $i < k$, the results that

$$\begin{aligned} \widehat{\theta}_k - \theta_i &= \widehat{\theta}_k - \theta_k + \theta_k - \theta_i \\ &\geq \widehat{\theta}_k - \theta_k + \theta_k - \theta_{k-1} \\ &\geq -e + \zeta \\ &\geq 2D + e \\ &> D \end{aligned} \tag{B.10}$$

and

$$\begin{aligned}\frac{\widehat{\theta}_k + \widehat{\theta}_{k+1}}{2} - \theta_i &= \frac{\widehat{\theta}_k - \theta_i}{2} + \frac{\widehat{\theta}_{k+1} - \theta_i}{2} \\ &\geq D + \frac{e}{2} + D + \frac{e}{2} \\ &\geq 2D + e \\ &> D\end{aligned} \tag{B.11}$$

from (B.1) and (B.4) lead to

$$2E(1-\sigma) \leq \Lambda(\widehat{\theta}_k - \theta_i) \leq 2E, \tag{B.12}$$

and

$$2E(1-\sigma) \leq \Lambda\left(\frac{\widehat{\theta}_k + \widehat{\theta}_{k+1}}{2} - \theta_i\right) \leq 2E. \tag{B.13}$$

When $i = k$, the result that

$$\begin{aligned}\frac{\widehat{\theta}_k + \widehat{\theta}_{k+1}}{2} - \theta_k &= \frac{\widehat{\theta}_k - \theta_k}{2} + \frac{\widehat{\theta}_{k+1} - \theta_k}{2} \\ &\geq -\frac{e}{2} + D + \frac{e}{2} \\ &\geq D\end{aligned} \tag{B.14}$$

can lead to

$$2E(1-\sigma) \leq \Lambda\left(\frac{\widehat{\theta}_k + \widehat{\theta}_{k+1}}{2} - \theta_k\right) \leq 2E. \tag{B.15}$$

When $i = k+1$, the results that

$$\begin{aligned}\widehat{\theta}_k - \theta_{k+1} &= \widehat{\theta}_k - \theta_k + \theta_k - \theta_{k+1} \\ &\leq e - 2D - 2e \\ &\leq -2D - e \\ &< -D\end{aligned} \tag{B.16}$$

and

$$\begin{aligned}\frac{\widehat{\theta}_k+\widehat{\theta}_{k+1}}{2}-\theta_{k+1} &= \frac{\widehat{\theta}_k-\theta_{k+1}}{2}+\frac{\widehat{\theta}_{k+1}-\theta_{k+1}}{2}\\ &\leq -D-\frac{e}{2}+\frac{e}{2}\\ &\leq -D\end{aligned} \tag{B.17}$$

can also lead to

$$0 \leq \Lambda(\widehat{\theta}_k-\theta_{k+1}) \leq 2E\sigma, \tag{B.18}$$

and

$$0 \leq \Lambda\left(\frac{\widehat{\theta}_k+\widehat{\theta}_{k+1}}{2}-\theta_{k+1}\right) \leq 2E\sigma. \tag{B.19}$$

When $i > k+1$, the results that

$$\begin{aligned}\widehat{\theta}_k-\theta_i &= \widehat{\theta}_k-\theta_k+\theta_k-\theta_i\\ &\leq \widehat{\theta}_k-\theta_k+\theta_k-\theta_{k+1}\\ &\leq e-2D-2e\\ &\leq -2D-e\\ &< -D\end{aligned} \tag{B.20}$$

and

$$\begin{aligned}\frac{\widehat{\theta}_k+\widehat{\theta}_{k+1}}{2}-\theta_i &= \frac{\widehat{\theta}_k-\theta_i}{2}+\frac{\widehat{\theta}_{k+1}-\theta_i}{2}\\ &\leq -D-\frac{e}{2}-D-\frac{e}{2}\\ &\leq -2D-e\\ &< -D\end{aligned} \tag{B.21}$$

can also lead to

$$0 \leq \Lambda(\widehat{\theta}_k-\theta_i) \leq 2E\sigma, \tag{B.22}$$

and

$$0 \leq \Lambda\left(\frac{\widehat{\theta}_k+\widehat{\theta}_{k+1}}{2}-\theta_i\right) \leq 2E\sigma. \tag{B.23}$$

In summary, we have that

$$\Lambda(\widehat{\theta}_k - \theta_i) \in \begin{cases} [2E(1-\sigma), 2E] & i = 1, 2, \ldots, k-1 \\ [0, 2E\sigma] & i = k+1, k+2, \ldots, K \end{cases} \tag{B.24}$$

and

$$\Lambda\left(\frac{\widehat{\theta}_k + \widehat{\theta}_{k+1}}{2} - \theta_i\right) \in \begin{cases} [2E(1-\sigma), 2E] & i = 1, 2, \ldots, k \\ [0, 2E\sigma] & i = k+1, k+2, \ldots, K \end{cases} \tag{B.25}$$

The cluster 1 with centroid $\widehat{\theta}_1$ includes the parameter range in $[\theta_{\min}, (\widehat{\theta}_1 + \widehat{\theta}_2)/2]$. Then the weight balance property (2.27) gives that

$$\begin{aligned}
&\int_{\theta_{\min}}^{\widehat{\theta}_1} v(\theta)\mathrm{d}\theta = \int_{hp_1}^{\frac{\widehat{\theta}_1+\widehat{\theta}_2}{2}} v(\theta)\mathrm{d}\theta \\
\Rightarrow &\int_{-\infty}^{\widehat{\theta}_1} v(\theta)\mathrm{d}\theta - \int_{-\infty}^{\theta_{\min}} v(\theta)\mathrm{d}\theta = \int_{-\infty}^{\frac{\widehat{\theta}_1+\widehat{\theta}_2}{2}} v(\theta)\mathrm{d}\theta - \int_{-\infty}^{\widehat{\theta}_1} v(\theta)\mathrm{d}\theta \\
\Rightarrow &2\int_{-\infty}^{\widehat{\theta}_1} v(\theta)\mathrm{d}\theta = \int_{-\infty}^{\theta_{\min}} v(\theta)\mathrm{d}\theta + \int_{-\infty}^{\frac{\widehat{\theta}_1+\widehat{\theta}_2}{2}} v(\theta)\mathrm{d}\theta \\
\Rightarrow &2\int_{-\infty}^{\widehat{\theta}_1} \sum_{i=1}^{K} c_i\lambda(\theta-\theta_i)\mathrm{d}\theta = \int_{-\infty}^{\theta_{\min}} \sum_{i=1}^{K} c_i\lambda(\theta-\theta_i)\mathrm{d}\theta + \int_{-\infty}^{\frac{\widehat{\theta}_1+\widehat{\theta}_2}{2}} \sum_{i=1}^{K} c_i\lambda(\theta-\theta_i)\mathrm{d}\theta \\
\Rightarrow &2\sum_{i=1}^{K} c_i \int_{-\infty}^{\widehat{\theta}_1} \lambda(\theta-\theta_i)\mathrm{d}\theta = \sum_{i=1}^{K} c_i \int_{-\infty}^{\theta_{\min}} \lambda(\theta-\theta_i)\mathrm{d}\theta + \sum_{i=1}^{K} c_i \int_{-\infty}^{\frac{\widehat{\theta}_1+\widehat{\theta}_2}{2}} \lambda(\theta-\theta_i)\mathrm{d}\theta \\
\Rightarrow &2\sum_{i=1}^{K} c_i \int_{-\infty}^{\widehat{\theta}_1-\theta_i} \lambda(\theta)\mathrm{d}\theta = \sum_{i=1}^{K} c_i \int_{-\infty}^{\theta_{\min}-\theta_i} \lambda(\theta)\mathrm{d}\theta + \sum_{i=1}^{K} c_i \int_{-\infty}^{\frac{\widehat{\theta}_1+\widehat{\theta}_2}{2}-\theta_i} \lambda(\theta)\mathrm{d}\theta \\
\Rightarrow &2\sum_{i=1}^{K} c_i\Lambda(\widehat{\theta}_1-\theta_i) = \sum_{i=1}^{K} c_i\Lambda(\theta_{\min}-\theta_i) + \sum_{i=1}^{K} c_i\Lambda\left(\frac{\widehat{\theta}_1+\widehat{\theta}_2}{2}-\theta_i\right) \\
\Rightarrow &2c_1\Lambda(\widehat{\theta}_1-\theta_1) = \\
&\sum_{i=1}^{K} c_i\Lambda(\theta_{\min}-\theta_i) + \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(\widehat{\theta}_1-\theta_i).
\end{aligned} \tag{B.26}$$

Combining the results

$$0 \leq \sum_{i=1}^{K} c_i \Lambda(\theta_{\min} - \theta_i) \leq \sum_{i=1}^{K} c_i 2E\sigma \tag{B.27}$$

from (B.7),

$$\begin{aligned} &\sum_{i=1}^{K} c_i \Lambda\left(\frac{\widehat{\theta}_1 + \widehat{\theta}_2}{2} - \theta_i\right) = c_1 \Lambda\left(\frac{\widehat{\theta}_1 + \widehat{\theta}_2}{2} - \theta_1\right) + \sum_{i=2}^{K} c_i \Lambda\left(\frac{\widehat{\theta}_1 + \widehat{\theta}_2}{2} - \theta_i\right) \\ \Rightarrow &2E(1-\sigma)c_1 \leq \sum_{i=1}^{K} c_i \Lambda\left(\frac{\widehat{\theta}_1 + \widehat{\theta}_2}{2} - \theta_i\right) \leq 2Ec_1 + \sum_{i=2}^{K} c_i 2E\sigma \end{aligned} \tag{B.28}$$

from (B.25), and

$$-2\sum_{i=2}^{K} c_i 2E\sigma \leq -2\sum_{i=2}^{K} c_i \Lambda(\widehat{\theta}_1 - \theta_i) \leq 0 \tag{B.29}$$

from (B.24), (B.26) lead to the bound for estimate error for θ_1:

$$\begin{aligned} &2E(1-\sigma)c_1 - 2E\sigma \sum_{i=2}^{K} 2c_i \leq 2c_1 \Lambda(\widehat{\theta}_1 - \theta_1) \leq 2E(1+\sigma)c_1 + 2E\sigma \sum_{i=2}^{K} 2c_i \\ \Rightarrow &E(1 - (1 + 2(K-1)r)\sigma) \leq \Lambda(\widehat{\theta}_1 - \theta_1) \leq E(1 + (1 + 2(K-1)r)\sigma). \end{aligned} \tag{B.30}$$

For any $2 \leq k \leq K-1$, cluster k with centroid hp_k includes the parameter range $[(\widehat{\theta}_{k-1} + \widehat{\theta}_k)/2, (\widehat{\theta}_k + \widehat{\theta}_{k+1})/2]$. Following the same weight balance result as before, we get

$$\begin{aligned} &\int_{\frac{\widehat{\theta}_{k-1}+\widehat{\theta}_k}{2}}^{\widehat{\theta}_k} v(\theta)\mathrm{d}\theta = \int_{\widehat{\theta}_k}^{\frac{\widehat{\theta}_k+\widehat{\theta}_{k+1}}{2}} v(\theta)\mathrm{d}\theta \\ \Rightarrow &2\sum_{i=1}^{K} c_i \Lambda(\widehat{\theta}_k - \theta_i) = \sum_{i=1}^{K} c_i \Lambda\left(\frac{\widehat{\theta}_{k-1} + \widehat{\theta}_k}{2} - \theta_i\right) + \sum_{i=1}^{K} c_i \Lambda\left(\frac{\widehat{\theta}_k + \widehat{\theta}_{k+1}}{2} - \theta_i\right) \\ \Rightarrow &2c_k \Lambda(\widehat{\theta}_k - \theta_k) = \\ &\sum_{i=1}^{K} c_i \Lambda\left(\frac{\widehat{\theta}_{k-1} + \widehat{\theta}_k}{2} - \theta_i\right) + \sum_{i=1}^{K} c_i \Lambda\left(\frac{\widehat{\theta}_k + \widehat{\theta}_{k+1}}{2} - \theta_i\right) - 2\sum_{i \neq k} c_i \Lambda(\widehat{\theta}_k - \theta_i) \end{aligned} \tag{B.31}$$

Similarly, combining the results that

$$\begin{aligned}
&\sum_{i=1}^{K} c_i \Lambda\left(\frac{\widehat{\theta}_{k-1}+\widehat{\theta}_k}{2}-\theta_i\right) = \sum_{i=1}^{k-1} c_i \Lambda\left(\frac{\widehat{\theta}_{k-1}+\widehat{\theta}_k}{2}-\theta_i\right) + \sum_{i=k}^{K} c_i \Lambda\left(\frac{\widehat{\theta}_{k-1}+\widehat{\theta}_k}{2}-\theta_i\right) \\
\Rightarrow & 2E(1-\sigma)\sum_{i=1}^{k-1} c_i \leq \sum_{i=1}^{K} c_i \Lambda\left(\frac{\widehat{\theta}_{k-1}+\widehat{\theta}_k}{2}-\theta_i\right) \leq 2E\sum_{i=1}^{k-1} c_i + 2E\sigma\sum_{i=k}^{K} c_i
\end{aligned} \tag{B.32}$$

from (B.25),

$$\begin{aligned}
&\sum_{i=1}^{K} c_i \Lambda\left(\frac{\widehat{\theta}_{k}+\widehat{\theta}_{k+1}}{2}-\theta_i\right) = \sum_{i=1}^{k} c_i \Lambda\left(\frac{\widehat{\theta}_{k}+\widehat{\theta}_{k+1}}{2}-\theta_i\right) + \sum_{i=k+1}^{K} c_i \Lambda\left(\frac{\widehat{\theta}_{k}+\widehat{\theta}_{k+1}}{2}-\theta_i\right) \\
\Rightarrow & 2E(1-\sigma)\sum_{i=1}^{k} c_i \leq \sum_{i=1}^{K} a_i \Lambda\left(\frac{\widehat{\theta}_{k}+\widehat{\theta}_{k+1}}{2}-\theta_i\right) \leq 2E\sum_{i=1}^{k} c_i + 2E\sigma\sum_{i=k+1}^{K} c_i
\end{aligned} \tag{B.33}$$

from (B.25), and

$$\begin{aligned}
&-2\sum_{i\neq k} c_i \Lambda(\widehat{\theta}_k-\theta_i) = -2\sum_{i=1}^{k-1} c_i \Lambda(\widehat{\theta}_k-\theta_i) - 2\sum_{i=k+1}^{K} a_i \Lambda(\widehat{\theta}_k-\theta_i) \\
\Rightarrow & -2E\sum_{i=1}^{k-1} 2a_i - 2E\sigma\sum_{i=k+1}^{K} 2c_i \leq -2\sum_{i\neq k} 2c_i \Lambda(\widehat{\theta}_k-\theta_i) \leq -2E(1-\sigma)\sum_{i=1}^{k-1} 2c_i
\end{aligned} \tag{B.34}$$

from (B.24), (B.31) lead to the same bound of estimation error for θ_k

$$\begin{aligned}
&2E(1-\sigma)c_k - 2E\sigma\sum_{i\neq k} 2c_i \leq 2c_k \Lambda(\widehat{\theta}_k-\theta_k) \leq 2E(1+\sigma)c_k + 2E\sigma\sum_{i\neq k} 2c_i \\
\Rightarrow & E(1-(1+2(k-1)r)\sigma) \leq \Lambda(\widehat{\theta}_k-\theta_k) \leq E(1+(1+2(k-1)r)\sigma).
\end{aligned} \tag{B.35}$$

The cluster K with centroid $\widehat{\theta}_K$ includes the parameter ranged $[(\widehat{\theta}_{K-1}+\widehat{\theta}_K)/2, \theta_{\max}]$ and has centroid $\widehat{\theta}_K$. The balance weight property gives that

$$\int_{\frac{\widehat{\theta}_{K-1}+\widehat{\theta}_K}{2}}^{\widehat{\theta}_K} v(\theta)\mathrm{d}\theta = \int_{\widehat{\theta}_K}^{\theta_{\max}} v(\theta)\mathrm{d}\theta$$
$$\Rightarrow 2\sum_{i=1}^{K} c_i\Lambda(\widehat{\theta}_K - \theta_i) = \sum_{i=1}^{K} c_i\Lambda\left(\frac{\widehat{\theta}_{K-1}+\widehat{\theta}_K}{2} - \theta_i\right) + \sum_{i=1}^{K} c_i\Lambda(\theta_{\max} - \theta_i)$$
$$\Rightarrow 2c_K\Lambda(\widehat{\theta}_K - \theta_K) = \sum_{i=1}^{K} c_i\Lambda\left(\frac{\widehat{\theta}_{K-1}+\widehat{\theta}_K}{2} - \theta_i\right) + \sum_{i=1}^{K} c_i\Lambda(\theta_{\max} - \theta_i) - 2\sum_{i=1}^{K-1} c_i\Lambda(\widehat{\theta}_K - \theta_i), \tag{B.36}$$

Combine

$$2E(1-\sigma)\sum_{i=1}^{K-1} c_i \le \sum_{i=1}^{K} c_i\Lambda\left(\frac{\widehat{\theta}_{K-1}+\widehat{\theta}_K}{2} - \theta_i\right) \le 2E\sum_{i=1}^{K-1} c_i + 2E\sigma c_K \tag{B.37}$$

from (B.25),

$$2E(1-\sigma)\sum_{i=1}^{K} c_i \le \sum_{i=1}^{K} c_i\Lambda(\theta_{\max} - \theta_i) \le 2E\sum_{i=1}^{K} c_i, \tag{B.38}$$

from (B.8), and

$$-2E\sum_{i=1}^{K-1} c_i \le -2\sum_{i=1}^{K-1} c_i\Lambda(\widehat{\theta}_K - \theta_i) \le -2E(1-\sigma)\sum_{i=1}^{K-1} 2c_i \tag{B.39}$$

from (B.24), we have the same bound of estimate error for θ_K

$$E(1-(1+2(K-1)r)\sigma) \le \Lambda(\widehat{\theta}_K - \theta_K) \le E(1+(1+2(K-1)r)\sigma) \tag{B.40}$$

In summary, for any $k = 1, 2, \ldots, K$, we have the same result

$$E(1-(1+2(K-1)r)\sigma) \le \Lambda(\widehat{\theta}_k - \theta_k) \le E(1+(1+2(K-1)r)\sigma). \tag{B.41}$$

If we choose $\delta = e > 0$ such that

$$\Lambda(\delta) = E(1+(1+2(K-1)a)\sigma), \tag{B.42}$$

then all estimate error is bound by δ:

$$-\delta \leq \widehat{\theta}_k - \theta_k \leq \delta \tag{B.43}$$

Combine (B.2) and (B.42), we can calculate

$$D = \Lambda^{-1}\left(2E\left(1 - \frac{\Lambda(\delta)/E - 1}{(2K-2)r+1}\right)\right). \tag{B.44}$$

With this, the final conlusion can be derived using (B.6) and (B.9). □

BIBLIOGRAPHY

[1] Richard G. Baraniuk, "Compressive sensing," *IEEE Signal Processing Magazine*, vol. 24, no. 4, pp. 118–120, 124, Jul. 2007.

[2] Emamnuel J. Candes, "Compressive sampling," in *International Congress Proceedings of the International Congressof Mathematicians*, Madrid, Spain, Aug. 2006, vol. 3, pp. 1433–1452.

[3] David L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[4] Volkan Cevher, Ali C. Gurbuz, James H. McClellan, and Rama Chellappa, "Compressive wireless arrays for bearing estimation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, USA, 2008, pp. 2497–2500.

[5] Marco F. Duarte, "Localization and bearing estimation via structured sparsity models," in *Statistical Signal Processing Workshop*, Ann Arbor, MI, Aug. 2012, IEEE, pp. 333–336.

[6] Matthew A. Herman and Thomas Strohmer, "High resolution radar via compressed sensing," *IEEE Transactions on Information Theory*, vol. 57, no. 6, pp. 2275–2284, Jun. 2009.

[7] Ioannis Kyriakides, "Adaptive compressive sensing and processing of delay-doppler radar waveforms," *IEEE Transactions on Information Theory*, vol. 60, no. 2, pp. 730–739, Feb. 2012.

[8] Satyabrata Sen, Gongguo Tang, and Arye Nehorai, "Multiobjective optimization of ofdm radar waveform for target detection," *IEEE Transactions on Information Theory*, vol. 59, no. 2, pp. 639–652, Feb. 2011.

[9] Ivana Stojanovic, Müjdat Çetin, and W. Clem Karl, "Compressed sensing of monostatic and multistatic sar," in *Algorithms for Synthetic Aperture Radar Imagery XVI*, 2009, pp. 7337–7342.

[10] Armin Eftekhari, Justin Romberg, and Michael B. Wakin, "Matched filtering from limited frequency samples," *IEEE Transactions on Information Theory*, vol. 59, no. 6, pp. 3475–3496, Jun. 2013.

[11] Karsten Fyhn, Marco F. Duarte, and Søren Holdt Jensen, "Compressive parameter estimation for sparse translation-invariant signals using polar interpolation," Available at `http://arxiv.org/abs/1306.2434`, 2013.

[12] Hadi Jamali-Rad and Geert Leus, "Sparsity-aware tdoa localization of multiple sources," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 4021–4025.

[13] Christian D. Austin, Randolph L. Moses, Joshua N. Ash, and Emre Ertin, "On the relation between sparse reconstruction and parameter estimation with model order selection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 3, pp. 560–570, Jun. 2010.

[14] Sebastien Bourguignon, Herve Carfantan, and Jerome Idier, "A sparsity-based method for the estimation of spectral lines from irregularly sampled data," *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 575–585, Dec. 2007.

[15] Marco F. Duarte and Richard G. Baraniuk, "Spectral compressive sensing," *Applied and Computational Harmonic Analysis*, vol. 35, no. 1, pp. 111–129, Jan. 2013.

[16] Karsten Fyhn, Hamid Dadkhahi, and Marco F. Duarte, "Spectral compressive sensing with polar interpolation," Available at `http://arxiv.org/abs/1303.2799`, 2013.

[17] Joel A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.

[18] Richard G. Baraniuk, Volkan Cevher, Marco F. Duarte, and Chinmay Hegde, "Model-based compressive sensing," *IEEE Transactions on Information Theory*, vol. 56, no. 4, pp. 1982–2001, Apr. 2010.

[19] Albert Fannjiang and Wenjing Liao, "Coherence pattern-guided compressive sensing with unresolved grids," *SIAM Journal on Imaging Sciences*, vol. 5, no. 1, pp. 179–202, Feb. 2012.

[20] Waheed U. Bajwa, Robert Calderbank, and Sina Jafarpour, "Why gabor frames? two fundamental measures of coherence and their role in model selection," *Journal of Communication and Network*, vol. 12, no. 4, pp. 289–307, Aug. 2012.

[21] Alyson K. Fletcher, Sundeep Rangan, and Vivek K Goyal, "Necessary and sufficient conditions for sparsity pattern recovery," *IEEE Transactions on Information Theory*, vol. 55, no. 12, pp. 5758–5772, Dec. 2009.

[22] Amin Karbasi, Ali Hormati, Soheil Mohajer, and Martin Vetterli, "Support recovery in compressed sensing: An estimation theoretic approach," in *IEEE International Symposium on Information Theory*, Seoul, South Korea, 2009, pp. 679–683.

[23] Galen Reeves and Michael Gastpar, "The sampling rate-distortion tradeoff for sparsity the sampling rate-distortion tradeoff for sparsity pattern recovery in compressed sensing," *IEEE Transactions on Information Theory*, vol. 58, no. 5, pp. 3065–3092, May 2012.

[24] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas, "A metric for distributions with applications to image databases," in *International Conference on Computer Vision*, Jan. 1998, pp. 59–66.

[25] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, Nov. 2000.

[26] Rishi Gupta, Piotr Indyk, and Eric Price, "Sparse recovery for earth mover distance," in *Annual Allerton Conference on Communication, Control, and Computing*, 2010, pp. 1742–1744.

[27] Piotr Indyk and Eric Price, "K-median clustering, model-based compressive sensing, and sparse recovery for earth mover distance," in *Proceedings of Annual ACM Symposium on Theory of computing*, San Jose, CA, Aug. 2011, pp. 627–636.

[28] Emmanuel J. Candes and Terence Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?," *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.

[29] Richard G. Baraniuk, Mark Davenport, Ronald DeVore, and Michael B. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, Dec. 2008.

[30] Emmanuel J. Candes and Terence Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[31] Emmanuel J. Candes and Terence Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[32] David L. Donoho and Michael Elad, "Optimally sparse representation in general (non-orthogonal) dictionaries via ℓ^1 minimization," in *Proceedings of the National Academy of Sciences of the United States of America*, 2003, number 5 in 100, pp. 2197–2202.

[33] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, Aug. 1998.

[34] Mario A. T. Figueiredo, Robert D. Nowak, and Stephen J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 586–597, Dec. 2007.

[35] Thomas Blumensath and Mike E. Davies, "Iterative thresholding for sparse approximations," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 629–654, Dec. 2008.

[36] Thomas Blumensath and Mike E. Davies, "Iterative hard thresholding for compressed sensing," *Applied and Computational Harmonic Analysis*, vol. 27, no. 3, pp. 265–274, Nov. 2009.

[37] Stéphane G. Mallat and Zhifeng Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.

[38] Joel A. Tropp and Anna C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.

[39] Deanna Needell and Joel A. Tropp, "Cosamp: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301–321, May 2009.

[40] Wei Dai and Olgica Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2230–2249, May 2009.

[41] Thomas Blumensath and Mike E. Davies, "Sampling theorems for signals from the union of finite-dimensional linear subspaces," *IEEE Transactions on Information Theory*, vol. 55, no. 4, pp. 1872–1882, Apr. 2009.

[42] Yue M. Lu and Minh N. Do, "Sampling signals from a union of subspaces," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 41–47, Mar. 2008.

[43] Richard G. Baraniuk and Michael B. Wakin, "Random projections of smooth manifolds," *Foundations of Computational Mathematics*, vol. 9, no. 1, pp. 51–77, Feb. 2009.

[44] Holger Rauhut, Karin Schnass, and Pierre Vandergheynst, "Compressed sensing and redundant dictionaries," *IEEE Transactions on Information Theory*, vol. 54, no. 5, pp. 2210–2219, May 2008.

[45] Moses S. Charikar, "Similarity estimation techniques from rounding algorithms," in *Proceedings of the ACM Symposium on Theory of Computing*, Montreal, Quebec, Canada, 2002, pp. 380–388.

[46] Piotr Indyk and Nitin Thaper, "Fast color image retrieval via embeddings," *Workshop on Statistical and Computational Theories of Vision*, 2003.

[47] Paul S. Bradley, Olvi L. Mangasarian, and W. N. Street, "Clustering via concave minimization," in *Advances in Neural Information Processing Systems*, Cambridge, MA, 1997, vol. 9, pp. 368–374, MIT Press.

[48] Pang Ning Tan, Michael Steinbach, and Vipin Kumar, *Introduction to Data Mining*, Addison Wesley, 1 edition, May 2005.

[49] Chaitanya Ekanadham, Daniel Tranchina, and Eero P. Simoncelli, "Recovery of sparse translation-invariant signals with continuous basis pursuit," *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 4735–4744, Oct. 2011.

[50] F.R. Giles and W.R. Pulleyblank, "Total dual integrality and integer polyhedra," *Linear algebra and its applications*, vol. 20, pp. 191–196, Jun. 1979.

[51] Joel A. Tropp, Jason N. Laska, Marco F. Duarte, Justin K. Romberg, and Richard G. Baraniuk, "Beyond nyquist: Efficient sampling of sparse bandlimited signals," *IEEE Transactions on Information Theory*, vol. 56, pp. 520–544, 2010.