Performance of Compressive Parameter Estimation via $K$-Median Clustering

Dian Mo, Student Member, IEEE, Marco F. Duarte, Senior Member, IEEE

Abstract

In recent years, compressive sensing (CS) has attracted significant attention in parameter estimation tasks, including frequency estimation, time delay estimation, and localization. In order to use CS in parameter estimation, parametric dictionaries (PDs) collect observations for a sampling of the parameter space and yield sparse representations for 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 the signal representations, alleviating these coherence issues. However, the resulting approaches depend heavily on a careful setting of the maximum allowable coherence; furthermore, their guarantees applied on the coefficient vector recovery do not translate in general to the parameter estimation task. In this paper, we propose the use of the earth mover’s distance (EMD), as applied to a pair of true and estimate PD coefficient vectors, to measure the error in sparsity-based parameter estimation. We formally analyze the connection between the aforementioned EMD and the parameter estimation error. Additionally, we leverage the relationship between $K$-median clustering and EMD-based sparse approximation to develop improved PD-based parameter estimation algorithms. We theoretically show that the EMD provides a better-suited metric for the performance of PD-based parameter estimation. Finally, we present numerical experiments that verify our theoretical results and show the performance improvements for the proposed compressive parameter estimation algorithms.

Index Terms

Compressive sensing, parameter estimation, parametric dictionary, earth mover’s distance, $K$-median clustering.

I. INTRODUCTION

Improvements in sensing hardware provide practitioners with increasing sampling frequency and resolution, which, however, results in very large observation data. To satisfy the demand of continuously growing observations, innovative frameworks have been proposed to acquire signals via a small number of measurements, including compressive sensing (CS) \cite{2,4}. CS simultaneously acquires and compresses sparse signals using random projections. In particular, if there exists a basis or dictionary in which the signal of interest can be represented sparsely, i.e.,
only a few coefficients are nonzero, then it is possible to successfully reconstruct the signal from a small number of measurements.

Recently, the application of CS has been extended from signal recovery to parameter estimation through the design of parametric dictionaries (PDs) \[5\,16\]. A PD collects observations of parametric signals corresponding to a set of parameters in order to provide sparse representations for signals of interest. The resulting connection between parameter estimation and sparse signal recovery has made it possible for compressive parameter estimation (CPE) to be implemented via standard (sparsity-based) CS recovery algorithms, where the dictionary coefficients obtained from signal recovery can be interpreted by matching the parameters with the nonzero coefficients. This CS approach has been previously formulated for landmark CPE problems, including localization and bearing estimation \[5\,10\], time delay estimation (TDE) \[11\,12\], and frequency estimation (FE) \[13\,16\].

Unfortunately, since only in the contrived case when the unknown parameters are all contained in the sampling set of the parameter space can the PD-based CPE be perfect, dense sampling of the parameter space is needed to improve the parameter estimation resolution \[6\]. However, dense sampling introduces high coherence between PD elements, i.e., the maximum inner product between any pair of PD elements is large, which is known to be harmful to standard CS recovery algorithms \[17\]. Previous approaches address this coherence problem by leveraging structured sparsity models \[13\] to inhibit highly coherent PD elements from appearing simultaneously in the recovered signal’s representation, an approach that has been termed band exclusion in the literature \[6\,12\,15\,16\,19\]. However, the performance of the resulting algorithms is highly dependent on the careful setting of the maximum allowable coherence among the chosen PD elements and restricts the minimum difference between any two parameters that can be observed simultaneously in a signal.

Additionally, almost all standard CS recovery algorithms used in previous PD-based CPE methods guarantee stable recovery of the PD coefficients. This stability refers to bounds on the error between the true and the estimated PD coefficient vectors measured via the Euclidean distance. The use of this error metric is intrinsically linked to the core hard thresholding operator, which sets all entries of an input vector to zero except for those with largest magnitudes and returns the optimal sparse approximation to the input vector in terms of Euclidean distance. However, such guarantee has a very limited impact on PD-based CPE, since only in the most demanding case of perfect recovery can the above guarantee be linked to the accurate parameter estimation. Accurate parameter estimation, which is translated to accurate estimation of the nonzero entries or the support of the PD coefficient vector, can be guaranteed in certain settings where the distance between PD coefficient vectors is measured by the difference between their supports. The most well-known distance that provides a good match to this criterion is the Hamming distance \[20\,23\]; however, this choice of distance will simply reveal the number of errors made in parameter estimation, rather than their magnitude.

In contrast, the earth mover’s distance (EMD) \[24\,26\], which quantifies the difference between two vectors by the minimum amount and distance of flow among the entries of one vector that is required to match the other one, is a very attractive option to measure the error in CPE due to the fact that the EMD of two PD coefficient vectors is indicative of the parameter estimation error when the entries of the PD coefficient vectors or the PD elements
are sorted by the value of the corresponding parameters.

In this paper, we propose a new method for CPE that uses PDs and sparsity and replaces the hard thresholding operator with $K$-median clustering. Existing results show that $K$-median clustering can return the optimal sparse approximation to an input vector in terms of EMD \cite{25, 26}, and so it provides a suitable alternative to the standard soft and hard thresholding operators. We also provide an analysis of (i) the connection between the EMD between a pair of true and estimate PD coefficient vectors and the corresponding parameter estimation error, and (ii) the relationship between $K$-median clustering and EMD-based sparse approximation. First, we theoretically show that the EMD between PD coefficient vectors provides an upper bound of the parameter estimation error. Second, we formulate theorems that provide performance guarantees for PD-based parameter estimation when clustering methods are used; these guarantees refer to the autocorrelation function of the parametric signal model, which measures the magnitude of the inner product between signal observations corresponding to different parameters. Third, we analyze the effect of the decay of the parametric signal’s correlation function on the performance of clustering-based parameter estimation methods, and relate these guarantees to the performance of the clustering methods under CS compression and signal noise. Finally, we introduce and analyze the joint use of thresholding and clustering methods to address performance loss resulting from compression and noise.

This paper is organized as follows. We provide an summary of compressive sensing and compressive parameter estimation and the issues present in existing work in Section II. In Section III, we present and analyze the use of EMD and clustering methods for PD-based parameter estimation; furthermore, we formulate and analyze an algorithm for PD-based sparse approximation in the EMD sense that employs $K$-median clustering and provides increased accuracy for parameter estimation. In Section IV, we present numerical simulations that verify our results on clustering method in the example applications of time delay estimation and frequency estimation. Finally, we provide our conclusions in Section V. The proofs of several theorems are included in the appendix.

II. BACKGROUND

A. Compressive Sensing

Compressive sensing (CS) has emerged as a technique integrating sensing and compression for signals that are known to be sparse or compressible in some basis. A discrete-time signal $x \in \mathbb{C}^N$ is $K$-sparse in a basis or a frame $\Psi$ when the signal can be represented by the basis or frame as $x = \Psi c$ with $\|c\|_0 \leq K$, where the $\ell_0$ norm $\| \cdot \|_0$ counts the number of nonzero entries. The dimension-reducing measurement matrix $\Phi \in \mathbb{R}^{M \times N}$ compresses the discrete signal $x$ to obtain the measurements $y = \Phi x \in \mathbb{C}^M$. Though, in general, it is ill-posed to recover the signal $x$ from the measurements $y$ when $M < N$ since $\Phi$ has a nontrivial null space, CS theory shows that it is possible to recover the signal from a small number of measurements when the signal is sparse and the measurement matrix satisfies the restricted isometry property (RIP) \cite{3, 27}, i.e., for any $K$-sparse signals $x_1$ and $x_2$,

\[(1 - \delta)\|x_1 - x_2\|_2 \leq \|\Phi x_1 - \Phi x_2\|_2 \leq (1 + \delta)\|x_1 - x_2\|_2. \tag{1}\]
In words, a measurement matrix satisfies the RIP when it can approximately preserve the Euclidean distance between any pair of $K$-sparse signals.

Given these conditions, the CS recovery problem usually can be solved via optimization methods or greedy algorithms. All existing greedy algorithms either iteratively obtain an improved estimation of the coefficient vectors, such as Iterative Hard Thresholding (IHT) [28], Compressed Sampling Matching Pursuit (CoSaMP) [29], and Subspace Pursuit (SP) [30], or iteratively identify the nonzero entries of the coefficient vectors, such as Orthogonal Matching Pursuit (OMP) [31, 32]. The mentioned greedy algorithms rely on the hard thresholding operator, which set all entries of an input vector to zero except those with largest magnitudes. The sparse vector resulting from the thresholding operator provides the optimal $K$-sparse approximation for the input vector in the sense that the output has minimum Euclidean distance to the input vector among all possible $K$-sparse vectors.

While classical CS processes signals by exploiting the fact that they can be described as sparse in some basis or frame, the locations of the nonzero entries of PD coefficient vectors often have underlying structure. To capture the additional structure in sparse signals, model-based CS replaces the thresholding operator in standard greedy algorithms by a corresponding structured sparse approximation, which, similarly, finds the optimal structured sparse approximations for a input vector in the sense that the output vector exhibits the desired structure and is closest to the input vector among all possible structured sparse vectors [15, 18, 19, 33].

B. Compressive Parameter Estimation

Parameter estimation problems are usually defined in terms of a parametric signal class, which is defined via a mapping $\psi : \Theta \rightarrow X$ from the parameter space $\Theta \subset \mathbb{R}^D$ to the signal space $X \subset \mathbb{C}^N$. The signal observed in a parameter estimation problem contains $K$ unknown parametric components $x = \sum_{i=1}^{K} c_i \psi(\theta_i)$, and the goal is to obtain estimates of the parameters $\hat{\theta_i}$ from the signal $x$. By introducing a parametric dictionary (PD) as a collection of samples from the signal space $\Psi = [\psi(\omega_1), \psi(\omega_2), \ldots, \psi(\omega_L)] \subseteq \psi(\Theta)$, which corresponds to a set of samples from the parameter space $\Omega = \{\omega_1, \omega_2, \ldots, \omega_L\} \subseteq \Theta$. Thus, the signal can be expressed as a linear combination of the dictionary elements $x = \Psi c$ when all the unknown parameters are contained in the sampling set of the parameter space; that is, when $\theta_i \in \Omega$ for each $i = 1, \ldots, K$. Therefore, finding the unknown parameters reduces to finding the PD elements appearing in the signal representation or, equivalently, finding the nonzero entries or support of the sparse PD coefficient vector $c$ for which we indeed have $y = \Phi \Psi c$. The search for the vector $c$ can be solved using the CS recovery process.

The PD-based compressive parameter estimation (CPE) can be perfect only if the parameter sample set $\Omega$ is dense and large enough to contain all of the unknown parameters $\{\theta_i\}_{i=1}^{K}$. If this stringent case is not met for some unknown parameter $\theta_k$, a denser sampling of the parameter space decreases the distance between the parametric signal $\psi(\theta_k)$ of the unknown parameter $\theta_k$ and the parametric signal $\psi(\omega_l)$ of the nearest parameter sample $\omega_l$, such that we can approximate $\theta_k$ by $\omega_l$. However, highly dense sampling increases the similarity between the adjacent PD elements and the coherence in the PD [34], which is measured by the maximum normalized inner product of
dictionary elements, i.e.,

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

(2)

Additionally, dense sampling increases the difficulty of distinguishing PD elements and severely hampers the performance of CSP [17, 35]. Prior work addressed such issues by using a coherence-inhibiting structured sparse approximation where the resulting $K$ nonzero entries of coefficient vectors correspond to the dictionary elements that have sufficiently low coherence, i.e., $|\langle \psi(\theta_i), \psi(\theta_j) \rangle| \leq \nu$ for $i, j = 1, \ldots, K$, in order to inhibit the highly coherent dictionary elements from appearing in signal representation simultaneously; this approach is known as band exclusion [6, 12, 15, 16, 19]. The maximum allowed coherence $\nu$ that defines the restriction on the choice of PD elements is essential to successful performance: setting its value too large results in the selection of coherent PD elements, while setting its value too small tightens up requirements on the minimum separation of the parameters.

Another issue that arises in PD-based CPE is that existing CS recovery algorithms commonly used in this setting can only guarantee stable recovery of the PD coefficient vector when the error is measured by the $\ell_2$ norm, i.e., when the estimated coefficient vector is close to the true coefficient vector in Euclidean distance. Such a guarantee is linked to the core hard/soft thresholding operation, which sets all entries of an input vector to zero except for those with the largest magnitudes and returns the optimal sparse approximation to the input vector with respect to the $\ell_2/\ell_1$ norm. However, the guarantee provides control on the performance of parameter estimation only in the most demanding case of exact recovery, when parameter estimation would be perfect. Otherwise, if the exact coefficient vector recovery cannot be met, such a PD coefficient vector recovery guarantee is meaningless for parameter estimation since the $\ell_2$ norm cannot precisely measure the difference between the supports of the PD coefficient vectors. The Hamming distance between pairs of PD coefficient vectors can measure the size of the discrepancy between their corresponding supports; furthermore, this distance has been previously leveraged as an error measure in the CS literature [20-23]. Nonetheless, in our case of interest, a Hamming distance metric would only capture the number of errors made in parameter estimation but would still not be able to quantify the magnitude of such errors.

Alternatively, the earth mover’s distance (EMD) has recently been used in CS to measure the error of the coefficient vectors in terms of the similarity between their supports [25, 26]. In particular, the EMD between two PD coefficient vectors with the same $\ell_1$ norm optimizes the work of the flow (i.e., the amount of the flow and the distance of flow) among one vector to make the two vectors equivalent [1]. Based on the fact that the work of the flow between any two entries of a PD coefficient vector is isometric with the distance between the two corresponding parameter values, it is reasonable that the EMD between pairs of PD coefficient vectors efficiently measures the corresponding error of parameter estimation. We elaborate our study of this property in Section III-A.

1The standard definition of EMD assumes that the two vectors have the same $\ell_1$ norm. Nonetheless, one can add an additional cost due to norm mismatch that is equal to the mismatch times the length of the signal [56].
C. K-Median Clustering

Cluster analysis partitions a set of data points based on the similarity information between each pair of points, which is usually expressed in terms of a distance [37, 38]. Clustering is the 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 and 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 the $K$ centroids $\{q_1, q_2, \ldots, q_K\}$ of the clusters such that each cluster contains all points that are more similar (i.e., closer) to their centroid than to other centroids:

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

One can define a clustering quality measure as the total sum of weighted distances between points and centroids:

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

Different choices of distances can result in different procedures to obtain the centroids [38]. If the Euclidean distance ($\ell_2$) is used, then each cluster’s centroid will be the mean of its elements, and so the clustering is called $K$-mean clustering. If the Manhattan distance ($\ell_1$) is used, then each cluster’s centroid will be the median of its elements, and so the clustering is called $K$-median clustering. In the special case where all points $p_i \in \mathbb{R}$, i.e., all the points are along a line, and the absolute value is used as a distance, i.e., $d(p_i, p_j) = |p_i - p_j|$, the object function defined in (4) becomes

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

One can solve for the centroids by differentiating the measure function and setting it to zero; the result is

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

where $\text{sign}(x)$ returns the sign of $x$. Equation (6) illustrates that the resulting centroids are the medians of the elements for each cluster and the points on the two sides of the centroids have maximally balanced weight:

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

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

D. Earth Mover’s Distance

The earth mover’s distance (EMD) between two vectors $(c, \hat{c})$ relies on the notion of mass assigned to each entry of the involved vectors and equal to each entry’s magnitude, with the goal to transfer mass among the entries of the first vector in order to match the mass of the entries of the second vector. We assume that the vectors $c$ and $\hat{c}$ are nonnegative and sparse, with nonzero entries $\{c_1, c_2, \ldots, c_K\}$ and $\{\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_K\}$, and with
supports $\Omega = \{\omega_1, \ldots, \omega_K\}$ and $\hat{\Omega} = \{\hat{\omega}_1, \ldots, \hat{\omega}_K\}$, respectively. EMD($c, \hat{c}$) represents the difference between the two vectors by finding the minimum sum of mass flows $f_{ij}$ from entry $c_i$ to entry $\hat{c}_j$ multiplied by the distance $d_{i,j} = |\omega_i - \hat{\omega}_j|$ that can be applied to the first vector $c$ to yield the second vector $\hat{c}$. This is a typical linear programming problem that can be written as

$$\text{EMD}(c, \hat{c}) = \min_{f_{ij}} \sum_{i,j} f_{ij}d_{ij}$$

such that

$$\sum_{j=1}^{K} f_{ij} = c_i, \quad i = 1, 2, \ldots, K;$$

$$\sum_{i=1}^{K} f_{ij} = \hat{c}_j, \quad j = 1, 2, \ldots, K;$$

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

(8)

E. EMD-Optimal Sparse Approximation

EMD-optimal sparse approximation plays a crucial role in our proposed CPE approach. This process makes it easy to integrate EMD into a CS framework to formulate a new CPE algorithm [15].

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

$$V_i = \{l \in I : |l - s_i| \leq |l - s_j|\}. \quad (9)$$

The EMD defined in (8) can be written as

$$\text{EMD}(v, \hat{c}) = \sum_{i=1}^{K} \sum_{j \in V_i} |v_i||j - s_i|.$$  \quad (10)

It is important to note that (10) has the same formula as (5), which is the objective function used in $K$-median clustering. Thus, one can pose a $K$-median clustering problem to minimize the value of (10) over all possible supports $S$. To that end, define $L$ points in a one-dimensional space with weights $|v_1|, |v_2|, \ldots, |v_L|$ and locations $1, 2, \ldots, L$. It is easy to see that if we denote the set of centroid positions obtained by performing $K$-median clustering for this problem as $S$, then the set $S$ corresponds to the support of the $K$-sparse signal that is closest to the vector $v$ when measured with the EMD. One can then simply compute the sets in (9) and define the EMD-optimal $K$-sparse approximation $\hat{c}$ to the vector $v$ as $\hat{c}_{s_i} = \sum_{j \in S_i} v_j$ for $s_i \in S$, with all other entries equal to zero. Thus, it is computationally feasible to provide sparse approximations in the EMD sense [25, 26].
III. CLUSTERING METHODS FOR PARAMETER ESTIMATION

A. Estimation Error

The goal of parameter estimation is to obtain parameter estimates such that the total errors of the unknown parameters and the estimated parameters is minimized. Computing the estimation error between a set of \( K \) one-dimensional true parameter values \( \theta = \{\theta_1, \theta_2, \ldots, \theta_K\} \subset \mathbb{R} \) and a set of \( K \) one-dimensional parameter value estimates \( \hat{\theta} = \{\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_K\} \subset \mathbb{R} \) is an assignment problem that minimizes the cost of assigning each true parameter to a parameter estimate, when the cost of assigning the true parameter \( \theta_i \in \theta \) to the parameter estimate \( \hat{\theta}_j \in \hat{\theta} \) is the absolute distance between the two values \( t_{ij} = |\theta_i - \hat{\theta}_j| \). The resulting parameter estimation error (PEE) can be obtained by solving a linear program

\[
Pee(\theta, \hat{\theta}) = \min_{\{g_{ij}\}} \sum_{i,j} g_{ij} t_{ij}
\]

such that
\[
\sum_{j=1}^{K} g_{ij} = 1, \quad i = 1, 2, \ldots, K; \quad \sum_{i=1}^{K} g_{ij} = 1, \quad j = 1, 2, \ldots, K; \quad g_{ij} \in \{0, 1\}, \quad i, j = 1, 2, \ldots, K.
\]

(11)

Here, obtaining \( g_{ij} = 1 \) indicates that the true parameter \( \theta_i \) will be assigned to the parameter estimate \( \hat{\theta}_j \). It is easy to show that the integer programming problem (11) has the same solution as its linear programming relaxation \[39\]

\[
Pee(\theta, \hat{\theta}) = \min_{\{g_{ij}\}} \sum_{i,j} g_{ij} t_{ij}
\]

such that
\[
\sum_{j=1}^{K} g_{ij} = 1, \quad i = 1, 2, \ldots, K; \quad \sum_{i=1}^{K} g_{ij} = 1, \quad j = 1, 2, \ldots, K; \quad g_{ij} \geq 0, \quad i, j = 1, 2, \ldots, K.
\]

(12)

When the sampling interval of the parameter space that generates the PD is constant and equal to \( \Delta \), it is easy to see that, for a pair of true and estimated PD vectors and the corresponding true and estimated parameters, the computation of the PEE (12) and the EMD (8) involve solutions that obey \( t_{ij} = \Delta \cdot d_{ij} \). This straightforward comparison demonstrates the close relationship between the EMD and PEE, which is formally stated in following theorem and proven in Appendix A.

**Theorem 1.** Assume that \( \Delta \) is the sampling interval of the parameter space that generates the PD used for parameter estimation. If \( c \) and \( \hat{c} \) are two \( K \)-sparse PD coefficient vectors corresponding to two sets of parameters \( \theta \) and \( \hat{\theta} \), then the EMD between the two coefficient vectors provides an upper bound of the PEE between the two sets of parameters:

\[
Pee(\theta, \hat{\theta}) \leq \frac{\Delta}{c_{\min}} \text{EMD}(c, \hat{c}),
\]

(13)
where \( c_{\text{min}} \) is the smallest component magnitude among the nonzero entries of \( c \) and \( \hat{c} \).

Theorem 1 gives the intuition that if an EMD-optimal sparse approximation method is available that provides a guarantee on the stable recovery of coefficient vectors in terms of the EMD, one can then potentially use it to provide guarantees for a compressive parameter estimation (CPE) algorithm.

B. The Role of Correlation in PD-based Parameter Estimation

We follow the convention of greedy algorithms for CS, where a proxy of the coefficient vector is obtained via the correlation of the observations with the dictionary elements, i.e., \( v = \Psi^* x \), where \( \Psi^* \) denotes the conjugate transpose of \( \Psi \). The resulting proxy vector \( v \) can be expressed as a linear combination of shifted correlation functions. The magnitudes of these components will be proportional to the magnitude of the corresponding signal components.

The correlation value between the dictionary elements corresponding to parameters \( \theta_1 \) and \( \theta_2 \) is defined as

\[
\lambda(\omega) = \lambda(\theta_2 - \theta_1) = \langle \psi(\theta_2), \psi(\theta_1) \rangle = \psi^*(\theta_1)\psi(\theta_2),
\]

where \( \omega = \theta_1 - \theta_2 \) measures the difference between parameters. In many parameter estimation problems, such as frequency estimation (FE) and time delay estimation (TDE), the correlation function has bounded variation such that the cumulative correlation function, defined as

\[
\Lambda(\theta) = \sum_{\omega \in \Omega, \omega \leq \theta} |\lambda(\omega)|,
\]

is bounded. The cumulative correlation function is a nondecreasing function with infimum \( \Lambda(-\infty) = 0 \) and supremum \( \Lambda(\infty) = \sum_{\omega \in \Omega} |\lambda(\omega)| \).

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

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

\[
y = x = \sum_{i=1}^{K} c_i \psi(\theta_i).
\]

Therefore, the proxy entries \( v_j \) correspond to inner products between the observation vector \( y \) and the PD elements \( \psi(\omega_j) \) corresponding to the sampled parameters \( \omega_j \in \Omega \), and thus can be expressed as a linear combination of
Fig. 1: Examples of correlation in PD constructions. (a) Correlation function $\lambda(\omega)$ and (b) normalized cumulative correlation function $\Lambda(\omega)/\Lambda(\infty)$ as a function of the discretized parameter $\omega/\Delta$ for time delay estimation (TDE) and frequency estimation (FE).

shifted correlation functions. For $j = 1, 2, \cdots, L$, the $j$-th entry of proxy vector is

$$v_j = v(\omega_j) = \langle x, \psi(\omega_j) \rangle = \sum_{i=1}^{K} c_i \langle \psi(\theta_i), \psi(\omega_j) \rangle = \sum_{i=1}^{K} c_i \lambda(\omega_j - \theta_i).$$

(17)

It is easy to see that the proxy function will feature local maxima at the sampled parameter values from the PD that are closest to the true parameter values. Thus, the goal of the parameter estimation finally reduces to the search of local maxima in the proxy vector over all parameter samples represented in the PD, which often is addressed via optimal sparse approximation of the proxy.

When the proxy function has fast decay (usually the case when the coherence of the PD is very small) it is possible to find the local maxima of the proxy function via the soft/hard thresholding operator, as used in standard greedy algorithms for sparse signal recovery. However, when the proxy function decays slowly (usually the case when the PD elements are highly coherent), the thresholding operator will unavoidably focus its search around the the peak of the proxy function with the largest magnitude, unless additional approaches like band exclusion are implemented. In contrast, EMD-based sparse approximation identifies the local maxima of the proxy directly by exploiting the fact that these local maxima correspond to the $K$-median clustering centroids, when certain conditions are met. Thus, we can propose an EMD and PD-based parameter estimation algorithm, shown as Algorithm 1 which leverages a standard iterative, Lloyd-style $K$-median clustering algorithm [37].

C. Performance Analysis

There are some conditions that the signal $x$ should satisfy to minimize the estimation error when using Algorithm 1.
Algorithm 1: Sparse Parameter Estimator $C(v, \Omega, K)$

Input: PD proxy vector $v$, set of PD parameter values $\Omega$, target sparsity $K$

Output: parameter estimates $\hat{\theta}$, sampled indices $S$

1: Initialize: set $L$ be the length of samples $\Omega$, choose $S$ as a random $K$-element subset of $\{1, \ldots, L\}$

2: repeat
3: $g_i = \text{arg} \min_{j=1,\ldots,K} |i - s_j|$ for each $i = 1, \ldots, L$ \{label each parameter sample\}
4: $s_j = \text{median}\{i \cdot |c_i| : g_i = j\}$ for each $j = 1, \ldots, K$ \{update medians\}
5: $\hat{\theta} = \Omega_S$ \{find parameter estimates\}
6: until $S$ does not change or maximum number of iterations is reached

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

  $$\zeta = \min_{i \neq j} |\theta_i - \theta_j|.$$  \hfill(18)

- **Parameter Range:** Any parameter observed should be sufficiently far away from the bounds of the parameter range. It is often convenient to restrict the feasible parameters and sampled parameters in a small range, i.e, $\Omega$ is bounded. According to (7), which implies a balance of the proxy $v$ around the the local maxima when using clustering method, there will be a bias in the estimation due to the missed portion of the symmetric correlation function $\lambda$ when the unknown parameter is too close to the bound of the parameter range $\Omega$. Therefore, the condition should also consider the minimum off-bound distance $\epsilon$, formally written as

  $$\epsilon = \min_{1 \leq i \leq K} \{\min(\theta_i - \min(\Omega), \max(\Omega) - \theta_i)\}.$$  \hfill(19)

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

  $$r = \max_{i \neq j} \frac{c_i}{c_j}.$$  \hfill(20)

With these conditions, we can formulate the following theorem that guarantees the performance of clustering method in CPE, which is proven in Appendix B.

**Theorem 2.** Assume that the signal $x$ given in (16) involves $K$ parameters $\theta_1, \theta_2, \ldots, \theta_K$ and has dynamic range as defined in (20). For a sufficiently small sampling interval $\Delta$, and for any allowed error $\sigma > 0$, if the minimum separation distance as (18) satisfies

$$\zeta \geq 2\Lambda^{-1}\left(2\Lambda(0)\left(1 - \frac{\Lambda(\sigma)/\Lambda(0) - 1}{(2K-2)r + 1}\right)\right) + 2\sigma,$$  \hfill(21)
and the minimum off-bound distance as (19) satisfies
\[ \epsilon \geq \Lambda^{-1} \left( 2\Lambda(0) \left( 1 - \frac{\Lambda(\sigma)/\Lambda(0) - 1}{2K} \right) \right), \] (22)
then the estimates \( \hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_K \) returned from Algorithm 1 have estimation error
\[ \left| \theta_k - \hat{\theta}_k \right| \leq \sigma, \quad k = 1, 2, \ldots, K. \] (23)

Though Theorem 2 is derived for the case of a nonnegative real correlation function and is asymptotic on the parameter sampling interval \( \Delta \), our numerical simulations in the sequel show that the predicted relationship between minimum separation and estimation error are observed in practical problems of modest sizes. As the theorem predicts, small estimation error requires large minimum separation distance. Additionally, the theorem reveals a linear relationship between the normalized cumulative correlation for the minimum separation distance \( \Lambda(\zeta)/\Lambda(\infty) \) and the normalized cumulative correlation for the maximum observed error \( \Lambda(\sigma)/\Lambda(0) \).

Theorem 2 also makes explicit the dependence of the parameter estimation performance on the cumulative correlation function \( \Lambda \). This dependence illustrates the wide difference in performances between TDE and FE: the minimum separation distance required by TDE is much smaller than that of FE, even though the value of the function \( \Lambda(\zeta) \) is the same in both cases, due to the contrasting rates of decay of the function \( \Lambda \), as shown in Figure 1.

D. Effect of Compression and Measurement Noise

We will see that the addition of CS and measurement noise make the estimation problem even harder, since in both of these cases there is a decrease in the rate of decay of the cumulative correlation function \( \Lambda \). When the measurement matrix \( \Phi \) is used to obtain the observed measurements \( y \) from the signal \( x \) such that
\[ y = \Phi x = \sum_{i=1}^{K} c_i \Phi \psi(\theta_i), \]
the proxy becomes
\[ v_j = v(\omega_j) = \langle \Phi^* y, \psi(\omega_j) \rangle = \langle y, \Phi \psi(\omega) \rangle = \sum_{i=1}^{K} c_i \langle \Phi \psi(\theta_i), \Phi \psi(\omega_j) \rangle. \] (24)

Only if \( \Phi^* \Phi = I \) can (24) be identical to (17). We define the compressed correlation function as
\[ \lambda_{\Phi}(\omega) = \lambda_{\Phi}(\theta_1 - \theta_2) = \langle \Phi \psi(\theta_2), \Phi \psi(\theta_1) \rangle = \psi^*(\theta_1) \Phi^* \Phi \psi(\theta_2). \] (25)

So the proxy can again be expressed as the linear combination of shifted copies of the redefined correlation function (25):
\[ v_j = v_j(\omega_j) = \sum_{i=1}^{K} c_i \lambda_{\Phi}(\omega_j - \theta_i). \] (26)

Although in general we will have \( \lambda \neq \lambda_{\Phi} \), we can use the preservation property of inner products through random projections [40]. That is, when \( \Phi \) has independent and identically distributed (i.i.d.) random entries and sufficiently many rows, there exists a constant \( \delta > 0 \) such that for all pairs \( (\theta_i, \theta_j) \) of interest we have
\[ (1 - \delta)\lambda(\theta_i - \theta_j) \leq \lambda_{\Phi}(\theta_i - \theta_j) \leq (1 + \delta)\lambda(\theta_i - \theta_j). \] (27)
The parameter $\delta$ decays as the compression rate of the matrix $M/N$ increases, and the manifolds we consider here are known to be amenable to large amounts of compression [41]. Such a relationship indicates that the compression can greatly affect the correlation function and the performance of clustering methods for CPE.

E. Quantifying the Role of Correlation Decay

We choose to focus on simple bounds for the correlation function $\lambda$ to analyze its role in the performance of EMD and PD-based parameter estimation. Similarly to [42], we use bounding functions to measure and control the decay rate of the correlation function $\lambda$. We approximate the correlation function $\lambda(\omega)$ with the exponential function $\kappa(\omega) = \exp(-a|\omega|)$ so that the exponential function provides an upper bound of the actual correlation function, i.e., $\lambda(\omega) \leq \kappa(\omega)$, and the performance obtained from the exponential function provides an upper bound of the performance from the correlation function. In the exponential function, $a$ is the parameter that controls the decay rate: the larger $a$ is, the faster that the correlation function decays.

It is easy to see that the decay rate of the redefined correlation function (25) will be smaller than that of the original correlation function (14). We assume that $\lambda(\omega) = \exp(-a|\omega|)$ and that a bound $\lambda(\omega) \leq \exp(-b|\omega|)$ exists; in this case, $b < a$ due to the fact that (27) provides us with the following upper bound:

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

(28)

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

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

$$v_t(\omega_i) = \begin{cases} v(\omega_i), & |v(\omega_i)| > t, \\ 0, & |v(\omega_i)| \leq t. \end{cases}$$

(29)

As demonstrated by the following theorem, proven in Appendix C, the thresholding operator reduces the required minimum separation distance for accurate estimation.

**Theorem 3.** Under the setup of Theorem 2, assume that the correlation function defined in (25) is given by $\lambda(\omega) = \exp(-a|\omega|)$. For a sufficiently small sampling interval $\Delta$, and for any allowed error $\sigma > 0$, if $t$ is the threshold given in (29), the dynamic range given in (20) is equal to $r$, and the minimum separation distance given in (18) satisfies

$$\zeta \geq \frac{1}{a} \ln \left( \frac{8r^2}{t^2/\left(r_{c\min}\right)^2 - \exp(-2a\sigma)} + 1 \right),$$

(30)
where $c_{\text{min}}$ is the minimum component magnitude, then the estimates $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_K$ returned from performing Algorithm 1 on the thresholded proxy $v_t$ have estimation error

$$|\theta_k - \hat{\theta}_k| \leq \sigma, \quad k = 1, 2, \ldots, K. \quad (31)$$

Theorem 3 extends Theorem 2 by including the use of thresholding as a tool to combat the slow decay speed of the correlation function, due to an ill-posed estimation problem, the use of CS, or the presence of noise in the measurements. One can instinctively see that the presence of noise in the measurements will also slow the decay of correlation function, which according to the theorem will require larger minimum separation or careful thresholding. In practice, the decay coefficient $a$ can usually be obtained by finding the minimum value such that the exponential function $\kappa(\omega)$ provides an upper bound of correlation function $\lambda(\omega)$. Although Theorem 3 is based on an approximation of the actual CPE problem setup, our numerical results in the sequel shows its validation in practical settings for TDE and FE.

IV. Numerical Experiments

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

For TDE, the parametric signal model describes a sampled version of a chirp:

$$\psi(\omega)[n] = \begin{cases} \frac{2}{3Tf_s} \exp \left( j2\pi \left( f_c + \frac{nT_s - \omega}{T} f_a \right) (nT_s - \omega) \right) \left( 1 + \cos \left( 2\pi \frac{nT_s - \omega}{T} \right) \right), & nT_s - \omega \in [0, T], \\ 0, & \text{otherwise,} \end{cases} \quad (32)$$

for $n = 1, \ldots, N$. Here $T = 1$ $\mu$s is the length of the chirp, $f_c = 1$ MHz is the chirp’s name frequency, $f_a = 20$ MHz is the chirp’s frequency sweep, $f_s = \frac{1}{T} = 50$ MHz is the sampling frequency of the discrete version of the chirp, and $N = 500$ samples are taken. The parameter space range is from $\theta_{\text{min}} = 0$ to $\theta_{\text{max}} = 10$ $\mu$s with sampling step $\Delta$. The PD for TDE contains all chirp signals corresponding to the sampled parameters $\Psi = [\psi(0), \psi(\Delta), \ldots, \psi(\theta_{\text{max}})]$.

For FE, the parametric signals are the $N$-dimensional signals with entries as

$$\psi(\theta)[n] = \frac{1}{\sqrt{N}} \exp \left( j2\pi \theta \frac{n}{N} \right), \quad n = 0, 1, \ldots, N - 1. \quad (33)$$

The parameter space ranged is from $\theta_{\text{min}} = 0$ to $\theta_{\text{max}} = 100$ Hz with sampling step $\Delta$. As before, the PD for FE contains all parametric signals corresponding to the sampled parameter $\Psi = [\psi(0), \psi(\Delta), \ldots, \psi(\theta_{\text{max}})]$.

In the first experiment, we illustrate the relationship between minimum separation distance and the maximum allowable error. We measure the performance for each minimum separation distance by the maximum estimation error over 1000 randomly chosen signals with unknown parameters. For TDE, the sampling step of the parameter space is $\Delta = 0.001$ $\mu$s, so that the PD contains observations for 10001 parameter samples, and we let the minimum separation vary over the range $[0.07 \mu$s, $0.1 \mu$s]. For FE, the sampling step of the parameter space is $\Delta = 0.05$ Hz,
so that the PD contains observations for 2001 parameter samples, and we let the minimum separation vary over
the range $[35\text{ Hz}, 70\text{ Hz}]$.

Figure 2 show the normalized cumulative correlation of maximum error $\Lambda(\sigma)/\Lambda(0)$ as a function of the the
normalized cumulative correlation for the minimum separation distance $\Lambda(\zeta)/\Lambda(\infty)$ for both TDE and FE; recall
that $\Lambda(0) = \Lambda(\infty)/2$. The figure also shows the relationship between the minimum separation and the maximum
error without the use of the cumulative correlation function for both example cases. The almost linear relationship
between $\Lambda(\zeta)/\Lambda(\infty)$ and $\Lambda(\sigma)/\Lambda(0)$ for both the TDE and FE cases numerically verifies the result of Theorem 2. The difference between the performance results for TDE and FE validates the conclusion that FE requires a
significantly larger minimum separation than TDE. From Figure 2(a), we known that it is impossible to get an
arbitrarily small estimation error even if the minimum separation keeps increasing, as the estimation error cannot
be smaller than the sampling step. To achieve more precise estimation, the use of additional methods such as
interpolation are needed. Nonetheless, it is easy to see that the specific behavior particular to a parameter estimation
problem is captured succinctly by the normalized cumulative correlation function, where the linear dependence is
consistently observed.

In the second experiment, we illustrate the application of Theorem 2 in the TDE problem. We vary the chirp’s
frequency sweep $f_a$ of the chip between 2 Hz and 20 Hz to generate different rates of decay of the correlation
function and obtain the decay parameter $a$ as the smallest value that enables the exponential function $\exp(-a|\omega|)$ to bound the correlation function $\lambda(\omega)$. We then measure the performance of TDE in the same manner as before
(maximum error over 1000 randomly drawn signals) by determining the minimum separation $\zeta$ for which when
maximum allowed estimation error is equal to the PD parameter sampling step $\Delta = 0.02\mu s$. The results in Figure 3(a) show the reciprocal relationship between the normalized minimum separation $\zeta/\Delta$ and the decay parameter $a$. Additionally, Figure 3(b) shows the logarithmic relationship between the normalized minimum separation $\zeta/\Delta$ and the dynamic range of the component magnitudes $r$ when $f_a = 10\text{ Hz}$ and $t = 0.9$. Finally, Figure 3(c) shows the negative logarithm relationship between the normalized minimum separation $\zeta/\Delta$ and the threshold $t$ with $f_a = 10\text{ Hz}$ and $r = 1$. All figures are indicative of agreement between Theorem 3 and practical results. Perhaps
the most important application of Theorem 3 focuses on the choice of threshold $t$ for the particular problem of
interest, which can improve the performance of the clustering method in compressive parameter estimation (CPE).

To deal with the problems of slow decay or large dynamic range, one can try to increase the threshold value on
the proxy rather than increasing the minimum separation to improve the estimation performance.

Our third and fourth experiments test the performance of clustering methods in CPE. $K$-median clustering is
incorporated into subspace pursuit (SP), a standard sparse recovery algorithm introduced in [30]. The resulting
clustering subspace pursuit (CSP), as shown in Algorithm 2 is compared with the band-exclusion subspace pursuit
(BSP) used in [12, 16] for TDE and FE. Similarly, CSP can be armed with polar interpolation to significantly improve
the estimation precision in a manner similar to band-excluding interpolation subspace pursuit (BISP) [1, 43]; we
call the resulting algorithm CISP.

Our third experiment tests the CSP, BSP, CISP, and BISP algorithms on 1000 independent randomly-generated
Fig. 2: Performance results for parameter estimation as a function of the parameter separation. The figures on the left correspond to TDE, while the figures on the right correspond to FE. Top: Normalized cumulative correlation function for the maximum PEE \( \Lambda(\sigma) / \Lambda(0) \) as a function of the normalized cumulative correlation function for the minimum separation \( \Lambda(\zeta) / \Lambda(\infty) \), showing a linear dependence. Bottom row: Normalized maximum error \( \sigma / \Delta \) as a function of normalized minimum separation \( \zeta / \Delta \).

TDE problems with minimum separation \( \zeta = 0.2 \mu s \) where CS measurements are taken under additive white Gaussian noise (AWGN). The maximum allowed coherence for BSP and BISP is chosen as \( \nu = 0.001 \), and the threshold for CSP and CISP is set as \( t = 0 \) (i.e., no thresholding takes place). Figure 4(a) shows the average parameter error as a function of the CS compression rate \( \kappa = M/N \) when no noise is added. The parameter error refers to minimum match cost between true parameters and estimated parameters solved by Hungarian algorithm [44]. The results indicate that clustering-based algorithms match the performance of their band-exclusion counterparts for most compression rates, without the need to carefully tune a band exclusion parameter (see [1] for a discussion). Additionally, Figure 4(b) shows the parameter estimation error as a function of the measurement’s signal-to-noise ratio (SNR). CSP and CISP are shown to achieve the same noise robustness as BSP and BISP, respectively, as their curves match almost exactly.
Fig. 3: Performance results for TDE parameter estimation for a variety of rates of decay for the parametric signal’s correlation function’s. The figures show the normalized minimum separation necessary for accurate estimation (σ ≤ Δ) as a function of (a) decay coefficient, (b) dynamic range, and (c) threshold value.

Algorithm 2 Clustering Subspace Pursuit (CSP)

Input: measurement vector $y$, measurement matrix $\Phi$, sparsity $K$, set of sampled parameters $\Omega$, threshold $t$

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

1: Initialize: $\hat{x} = 0$, $S = \emptyset$, generate PD $\Psi$ from $\Omega$.
2: repeat
3: $y_r = y - \Phi \hat{x}$ \hspace{1cm} \{Compute residual\}
4: $v = (\Phi \Psi)^+ y_r$ \hspace{1cm} \{Obtain proxy from residual\}
5: $v(|v| \leq t) = 0$ \hspace{1cm} \{Threshold proxy\}
6: $S = S \cup C(v, \Omega, K)$ \hspace{1cm} \{Augment parameter estimates from proxy\}
7: $c = (\Phi \Psi S)^+ y$ \hspace{1cm} \{Obtain proxy on parameter estimates\}
8: $S = C(c, \Omega S, K)$ \hspace{1cm} \{Refine parameter estimates\}
9: $\hat{x} = \Psi_S c_S$ \hspace{1cm} \{Assemble signal estimate\}
10: $\hat{\theta} = \Omega_S$ \hspace{1cm} \{Assemble parameter estimates\}
11: until a convergence criterion is met

In our fourth and final experiment, we repeat the third experiment on the FE problem instead, with 1000 independent randomly drawn signals for each setup with minimum separation ζ = 2Hz. The maximum allowed coherence for BSP and BISP is $\nu = 0.2$, and the threshold for CSP and CISP is set to $t = 0.4$. The figure shows that with the proper threshold, CSP can have the same performance as BSP.

V. CONCLUSION

In this paper, we have introduced and analyzed the relationship between the earth mover’s distance (EMD) applied to PD coefficient vectors and the parameter estimation error (PEE) obtained from sparsity-based methods. We also leveraged the relationship between EMD-based sparse approximation and $K$-median clustering algorithms in the
design of new compressive parameter estimation (CPE) algorithms. Based on the relationship between the EMD and the PEE, we have analytically shown that the EMD between PD coefficient vectors provides an upper bound of the PEE obtained from methods that use PDs and EMD. Furthermore, by noting the connection between EMD-sparse approximation and $K$-median clustering, we formulate the algorithms that return best sparse approximation in terms of EMD and we have derived three theoretical results that provide performance guarantees of EMD-based parameter estimation algorithms under some requirements for the signals observed. Our experimental result shows the validation of our analysis in several practical settings, and provides methods to control the effect of coherence, compression, and noise in the performance of CPE.
ACKNOWLEDGEMENTS

We thank Michael Wakin and Armin Eftekhari for helpful comments and for pointing us to [41].

APPENDIX A

PROOF OF THEOREM 1

We first consider the case where the two vectors \( c \) and \( \widehat{c} \) have the same \( \ell_1 \) norm, so that the standard definition of the EMD applies. Let \( f^* = [f_{11}^*, f_{12}^*, \ldots, f_{1K}^*, f_{21}^*, \ldots, f_{KK}^*]^T \) be the vector containing all \( f_{ij}^* \) that solves the optimization problem (8), \( g^* \) be the similarly-defined binary vector that solves the optimization problem (12), and \( r = f^* - c_{\min}g^* \). Similarly, let \( d \) and \( t \) be the similarly-defined vectors collecting all ground distances \( d_{ij} \) and \( t_{ij} \).

Then from (12) and (8), we have

\[
\text{EMD}(c, \widehat{c}) = d^T f^* = d^T (c_{\min}g^* + r) = c_{\min}d^T g^* + d^T r = \frac{c_{\min}}{\Delta} d^T g^* + d^T r = \frac{c_{\min}}{\Delta} \text{PEE}(\theta, \widehat{\theta}) + d^T r. \tag{34}
\]

Note that the first term in (34) is the value of the objective function in the optimization problem (8) when all entries of both \( c \) and \( \widehat{c} \) have magnitude \( c_{\min} \). The second term corresponds to the contribution to the objective function due to magnitudes that are larger than \( c_{\min} \). We show now that this second term is nonnegative.

When the magnitude of the entry \( \omega_i \) of \( c \) increases from \( c_{\min} \) to \( c_i \), at least one of the outgoing flow \( f_{ij} \) will need to increase. This implies that the corresponding flow \( f_{ij}^* \geq c_{\min}g_{ij}^* \). Thus we will have for some \( j \) that

\[
r_{ij} = f_{ij}^* - c_{\min}g_{ij}^* \geq 0.
\]

So, having shown that \( r \) is nonnegative, we have that \( d^T r \geq 0 \). Then it is possible to rewrite (34) as

\[
\text{EMD}(c, \widehat{c}) \geq \frac{c_{\min}}{\Delta} \text{PEE}(\theta, \widehat{\theta}), \tag{35}
\]

which proves the theorem. The result is still valid when \( \|c\|_1 \neq \|\widehat{c}\|_1 \) since the additional mismatch penalty further increases the value of the EMD. \( \square \)

APPENDIX B

PROOF OF THEOREM 2

Asymptotically, when the sampling step of the parameter \( \Delta \to 0 \), the proxy defined as (17) becomes a continuous function such that

\[
v(\omega) = \sum_{i=1}^{K} \lambda(\omega - \theta_i) \tag{36}
\]

for all \( \omega \in \Omega \). In additional, the balanced weight properties around the cluster centroid \( \widehat{\theta}_i \), as defined in (7), reduces to the equality

\[
\int_{p \in C_j, p \leq \widehat{\theta}_i} w(p)dp = \int_{p \in C_j, p \geq \widehat{\theta}_i} w(p)dp, \tag{37}
\]

where \( p \) is the position function and \( w \) is the weight function. Additionally, the cumulative correlation function in (15) converges to the integral

\[
\Lambda(\theta) = \int_{-\infty}^{\theta} \lambda(\omega)d\omega. \tag{38}
\]
Without loss of generality, if \( \theta_{\min} = \min(\Omega) \) and \( \theta_{\max} = \max(\Omega) \), assume that parameter values are sorted so that
\[
\theta_{\min} + \epsilon \leq \theta_1 < \theta_2 < \cdots < \theta_K \leq \theta_{\max} - \epsilon.
\] (39)

When the entries of the proxy \( v \) is clustered into \( K \) groups according to the centroids \( \hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_K \), as shown in Algorithm 1, the point \((\hat{\theta}_i + \hat{\theta}_{i+1})/2\) is the upper bound for cluster \( i \) and the lower bound for cluster \( i + 1 \), since it has the same distance to both centroids. We will show how large the minimum separation \( \zeta \) and minimum off-bound distance \( \epsilon \) need to be such that the maximum estimation error is \( e \), i.e., \( \max_k |\hat{\theta}_k - \theta_k| = e \).

We first consider the cases \( 2 \leq k \leq K - 1 \): the \( k \)-th cluster with centroid \( \hat{\theta}_k \) includes the parameter range \([\hat{\theta}_{k-1} + \hat{\theta}_k, \hat{\theta}_{k+1} + \hat{\theta}_k]/2\], i.e.,
\[
\int_{\hat{\theta}_{k-1} + \hat{\theta}_k}^{\hat{\theta}_k} v(\omega) \, d\omega = \int_{\hat{\theta}_k}^{\hat{\theta}_{k+1} + \hat{\theta}_k} v(\omega) \, d\omega,
\]
\[
2 \int_{-\infty}^{\hat{\theta}_k} v(\omega) \, d\omega = \int_{-\infty}^{\hat{\theta}_{k-1} + \hat{\theta}_k} v(\omega) \, d\omega + \int_{-\infty}^{\hat{\theta}_{k+1} + \hat{\theta}_k} v(\omega) \, d\omega,
\]
\[
2 \sum_{i=1}^{K} c_i \Lambda \left( \hat{\theta}_k - \theta_i \right) = \sum_{i=1}^{K} c_i \Lambda \left( \frac{\hat{\theta}_{k-1} + \hat{\theta}_k}{2} - \theta_i \right) + \sum_{i=1}^{K} c_i \Lambda \left( \frac{\hat{\theta}_k + \hat{\theta}_{k+1}}{2} - \theta_i \right),
\]
\[
2c_k \Lambda \left( \hat{\theta}_k - \theta_k \right) = \sum_{i=1}^{K} c_i \Lambda \left( \frac{\hat{\theta}_{k-1} + \hat{\theta}_k}{2} - \theta_i \right) + \sum_{i=1}^{K} c_i \Lambda \left( \frac{\hat{\theta}_k + \hat{\theta}_{k+1}}{2} - \theta_i \right) - 2 \sum_{i \neq k} c_i \Lambda \left( \hat{\theta}_k - \theta_i \right).
\] (40)

Since \( \hat{\theta}_k - \theta_k \geq -\epsilon \) and \( \theta_{k+1} - \theta_k \geq \zeta \), for \( k = 2, 3, \ldots, K - 1 \), we obtain a lower bound of the left hand side.
of (40) by repeatedly using the fact that \( \Lambda(\omega) \) is nondecreasing:

\[
\begin{align*}
\sum_{i=1}^{K} c_i \Lambda \left( \frac{\hat{\theta}_{k-1} + \hat{\theta}_k}{2} - \theta_i \right) + \sum_{i=1}^{K} c_i \Lambda \left( \frac{\hat{\theta}_k + \hat{\theta}_{k+1}}{2} - \theta_i \right) - 2 \sum_{i \neq k} c_i \Lambda \left( \hat{\theta}_k - \theta_i \right) \\
\geq \sum_{i=1}^{K} c_i \Lambda \left( \frac{\hat{\theta}_{k-1} + \hat{\theta}_k}{2} - \theta_{k-1} \right) + \sum_{i=1}^{K} c_i \Lambda \left( \frac{\hat{\theta}_k + \hat{\theta}_{k+1}}{2} - \theta_k \right) + \sum_{i=1}^{k} c_i \Lambda \left( \hat{\theta}_k - \theta_1 \right) - 2 \sum_{i=1}^{k} c_i \Lambda \left( \hat{\theta}_k - \theta_{k+1} \right)
\end{align*}
\]

Plugging in this lower bound, we have that

\[
2 \Lambda \left( \hat{\theta}_k - \theta_k \right) \geq (2(K - 1)r + 1) \Lambda \left( \frac{\zeta}{2} - e \right) - 2(K - 1)r \Lambda(\infty). \tag{41}
\]

Similarly, the upper bound of the left hand side of (40) is

\[
2 \Lambda \left( \hat{\theta}_k - \theta_k \right) \leq (2(K - 1)r + 2) \Lambda(\infty) - (2(K - 1)r + 1) \Lambda \left( \frac{\zeta}{2} - e \right). \tag{42}
\]

If \( \zeta \) satisfies

\[
\Lambda \left( \frac{\zeta}{2} - e \right) \geq \Lambda(\infty) \left( 1 - \frac{\Lambda(\sigma)/\Lambda(0) - 1}{2(K - 1)r + 1} \right), \tag{43}
\]

it is easy to verify that

\[
\begin{align*}
\Lambda \left( \hat{\theta}_k - \theta_k \right) &\geq (2(K - 1)r + 1) \Lambda \left( \frac{\zeta}{2} - e \right) - 2(K - 1)r \Lambda(\infty) \\
&\geq (2(K - 1)r + 1) \Lambda(\infty) - 2\Lambda(\sigma) + \Lambda(\infty) - 2(K - 1)r \Lambda(\infty) \\
&\geq 2\Lambda(\infty) - 2\Lambda(\sigma) \geq 2\Lambda(-\sigma), \tag{44}
\end{align*}
\]
and
\[
2\Lambda \left( \hat{\theta}_k - \theta_k \right) \leq (2(K-1)r + 2) \Lambda(\infty) - (2(K-1)r + 1) \Lambda \left( \frac{\zeta}{2} - \epsilon \right)
\]
\[
\leq (2(K-1)r + 2) \Lambda(\infty) - (2(K-1)r + 1) \Lambda(\infty) + 2\Lambda(\sigma) - \Lambda(\infty)
\]
\[
\geq 2\Lambda(\sigma),
\]
which imply \(-\sigma \leq \hat{\theta}_k - \theta_k \leq \sigma\) for \(k = 2, 3, \ldots, K-1\).

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

From the weight balance property, we have
\[
\int_{\theta_{\min}}^{\hat{\theta}_1} v(\omega) \, d\omega = \int_{\hat{\theta}_1}^{\theta_{\min}} v(\omega) \, d\omega,
\]
\[
2 \int_{-\infty}^{\hat{\theta}_1} v(\omega) \, d\omega = \int_{-\infty}^{\theta_{\min}} v(\omega) \, d\omega + \int_{-\infty}^{\hat{\theta}_2} v(\omega) \, d\omega,
\]
\[
2 \sum_{i=1}^{K} c_i \Lambda(\hat{\theta}_1 - \theta_i) = \sum_{i=1}^{K} c_i \Lambda(\theta_{\min} - \theta_i) + \sum_{i=1}^{K} c_i \Lambda \left( \frac{\hat{\theta}_1 + \hat{\theta}_2}{2} - \theta_i \right),
\]
\[
2c_1 \Lambda(\hat{\theta}_1 - \theta_1) = \sum_{i=1}^{K} c_i \Lambda(\theta_{\min} - \theta_i) + \sum_{i=1}^{K} c_i \Lambda \left( \frac{\hat{\theta}_1 + \hat{\theta}_2}{2} - \theta_i \right) - 2 \sum_{i=2}^{K} c_i \Lambda(\hat{\theta}_1 - \theta_i).
\]

If \(\epsilon\) satisfies
\[
\Lambda(\epsilon) \geq \Lambda(\infty) \left( 1 - \frac{\Lambda(\sigma)/\Lambda(0) - 1}{2Kr} \right),
\]
then we have the following result from (46):
\[
2\Lambda(\hat{\theta}_1 - \theta_1) \leq \sum_{i=1}^{K} c_i \Lambda(\theta_{\min} - \theta_i) + \sum_{i=1}^{K} c_i \Lambda \left( \frac{\hat{\theta}_1 + \hat{\theta}_2}{2} - \theta_i \right) - 2 \sum_{i=2}^{K} c_i \Lambda(\hat{\theta}_1 - \theta_i)
\]
\[
\leq \sum_{i=1}^{K} c_i \Lambda(\theta_{\min} - \theta_i) + \Lambda \left( \frac{\hat{\theta}_1 + \hat{\theta}_2}{2} - \theta_1 \right) + \sum_{i=2}^{K} c_i \Lambda \left( \frac{\hat{\theta}_1 + \hat{\theta}_2}{2} - \theta_2 \right) - 2 \sum_{i=2}^{K} c_i \Lambda(\hat{\theta}_1 - \theta_K)
\]
\[
\leq Kr \Lambda(-\epsilon) + \Lambda(\infty) + (K-1)r \Lambda \left( -\frac{\zeta}{2} + \epsilon \right)
\]
\[
\leq Kr \left( \Lambda(\infty) - \Lambda(\epsilon) \right) + \Lambda(\infty) + (K-1)r \left( \Lambda(\infty) - \Lambda \left( \frac{\zeta}{2} - \epsilon \right) \right)
\]
\[
\leq Kr \frac{2\Lambda(\sigma) - \Lambda(\infty)}{2Kr} + \Lambda(\infty) + (K-1)r \frac{2\Lambda(\sigma) - \Lambda(\infty)}{2(K-1)r + 1}
\]
\[
\leq 2\Lambda(\sigma).
\]

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

In summary, when all estimation errors are smaller than \(\sigma\), it is straightforward for us to replace the \(\epsilon\) in (43) by \(\sigma\) to get the expected condition on \(\zeta\):
\[
\Lambda \left( \frac{\zeta}{2} - \sigma \right) \geq \Lambda(\infty) \left( 1 - \frac{\Lambda(\sigma)/\Lambda(0) - 1}{2(K-1)r + 1} \right).
\]
\[
\square
When the redefined correlation function is $\lambda_\theta(\omega) = \exp(-a|\omega|)$, the proxy function given in (26) is

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

Without loss of generality, assume parameter values are sorted so that $\theta_1 < \theta_2 < \cdots < \theta_K$ and all component magnitudes no smaller than 1.

Assume only the proxy in the parameter range $[l_j, u_j]$ around each $\theta_j$ will be preserved after thresholding with level $t$. We have $l_1 < \theta_1 < u_1 < \cdots < \theta_{j-1} < u_{j-1} < l_j < \theta_j < u_j < \cdots < \theta_K$. So the proxy at $\omega = u_j$ has value equal to the threshold, i.e.,

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

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

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

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

$$T_j = A_j \frac{1}{U_j} + \frac{1}{U_j} + B_j U_j,

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

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

and $T_j = t/c_j$, $A_j$ and $B_j$ satisfy

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

One solution of the quadratic equation (50) is

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

In this solution we can see that $U_j$ decreases as $T_j$ increases. The alternative solution is omitted, since $U_j$ will increase as $T_j$ increases.
Similarly,

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

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

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

\[
T_j = A_j L_j + \frac{1}{L_j} + B_j \frac{1}{L_j},
\]

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

\[
L_j = \frac{T_j - \sqrt{T_j^2 - 4 A_j (1 + B_j)}}{2A_j}.
\]

Let \( \hat{\theta}_j \in [l_j, u_j] \) be the estimated parameter for \( \theta_j \). Asymptotically, when the sampling step of the parameter space \( \Delta \) goes to zero, the balance weight properties implies

\[
\int_{l_j}^{u_j} \sum_{i=1}^{K} c_i \exp(-a |\omega - \theta_i|) d\omega = \int_{l_j}^{u_j} \sum_{i=1}^{K} c_i \exp(-a |\omega - \theta_i|) d\omega.
\]

When \( \hat{\theta}_j \leq \theta_j \), we have

\[
\frac{a}{c_j} \int_{l_j}^{u_j} \sum_{i=1}^{K} c_i \exp(-a |\omega - \theta_i|) d\omega
\]

\[
= a \sum_{i=1}^{j-1} \frac{c_i}{c_j} \int_{l_j}^{u_j} \exp(-a (\omega - \theta_i)) d\omega + a \int_{l_j}^{u_j} \exp(-a (\theta_j - \omega)) d\omega + a \sum_{i=j+1}^{K} \frac{c_i}{c_j} \int_{l_j}^{u_j} \exp(-a (\theta_i - \omega)) d\omega
\]

\[
= a A_j \int_{l_j}^{u_j} \exp(-a (\omega - \theta_j)) d\omega + a \int_{l_j}^{u_j} \exp(-a (\theta_j - \omega)) d\omega + a B_j \int_{l_j}^{u_j} \exp(-a (\theta_j - \omega)) d\omega
\]

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

\[
= -A_j E_j + B_j \frac{1}{E_j} + \frac{1}{E_j} + A_j L_j - \frac{1}{L_j} - B_j \frac{1}{L_j},
\]

(56)
where $E_j = 1/\lambda(\hat{\theta}_j - \theta_j) = \exp(a |\hat{\theta}_j - \theta_j|) = \exp(a (\theta_j - \hat{\theta}_j)) \geq 1$, and

$$
\frac{a}{c_{\hat{\theta}_j}} \sum_{i=1}^{K} c_i \exp(-a |\omega - \theta_i|) \, d\omega
$$

$$
= a \sum_{i=1}^{j-1} \frac{c_i}{c_{\hat{\theta}_j}} \int_{\hat{\theta}_j}^{\theta_i} \exp(-a (\omega - \theta_i)) \, d\omega + a \int_{\hat{\theta}_j}^{\theta_j} \exp(-a (\omega - \theta_j)) \, d\omega + a \sum_{i=j+1}^{K} \frac{c_i}{c_{\hat{\theta}_j}} \int_{\hat{\theta}_j}^{\theta_i} \exp(-a (\theta_i - \omega)) \, d\omega
$$

$$
= a \sum_{i=1}^{j-1} \frac{c_i}{c_{\hat{\theta}_j}} \int_{\hat{\theta}_j}^{\theta_i} \exp(-a (\omega - \theta_i)) \, d\omega + a \int_{\hat{\theta}_j}^{\theta_j} \exp(-a (\theta_j - \omega)) \, d\omega + a \sum_{i=j+1}^{K} \frac{c_i}{c_{\hat{\theta}_j}} \int_{\hat{\theta}_j}^{\theta_i} \exp(-a (\theta_i - \omega)) \, d\omega
$$

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

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

(57)

After plugging (56) and (57) into (55) and moving all terms with $L_j$ or $U_j$ to the right side and moving other terms to the left side, we obtain

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

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

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

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

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

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

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

(58)

One can show a similar result when $\theta_j > \hat{\theta}_j$ and $E_j = \exp(a |\theta_j - \hat{\theta}_j|) = \exp(a (\theta_j - \hat{\theta}_j))$, so that (55) is reduced to

$$
\frac{A_j - B_j}{S_j} = \begin{cases} 
A_j E_j - B_j \frac{1}{E_j} + 1 - \frac{1}{E_j} & \text{if } \theta_j \leq \hat{\theta}_j \\
A_j \frac{1}{E_j} - B_j E_j - 1 + \frac{1}{E_j} & \text{if } \theta_j > \hat{\theta}_j
\end{cases}
$$

(59)

where

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

(60)

and

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

(61)
with the relationship (51). We now show that if
\[ S_j \geq \sqrt{ \left( \frac{t}{r c_{\text{min}}} \right) - 8 \left( \frac{r}{\exp(a\zeta) - 1} \right)^2 } \geq \exp(-a\sigma), \] (62)
or
\[ \zeta \geq \frac{1}{a} \ln \left( \sqrt{\frac{8n^2}{t^2/r c_{\text{min}}^2 - \exp(-2a\sigma)}} + 1 \right), \] (63)
as we expected, the estimation error is small.

When \( \theta_j \leq \hat{\theta}_j \),
\[ (A_j - B_j)/S_j = A_j E_j - B_j/E_j + 1 - 1/E_j \geq (A_j - B_j)/E_j, \] (64)
which requires \( A_j \geq B_j \) due to the fact that \( S_j \leq E_j \). So
\[ A_j E_j - B_j \frac{1}{E_j} + 1 - \frac{1}{E_j} = \frac{A_j - B_j}{S_j} \leq (A_j - B_j) \exp(a\sigma) \leq A_j \exp(a\sigma) - B_j \exp(-a\sigma) + 1 - \exp(-a\sigma), \] (65)
since \( \exp(-a\sigma) \leq 1 \leq \exp(a\sigma) \). This implies that \( E_j \leq \exp(a\sigma) \) and \( \hat{\theta}_j - \theta_j \leq \sigma \). Similarly, when \( \theta_j > \hat{\theta}_j \), \( A_j \leq B_j \) and \( \hat{\theta}_j - \theta_j \geq -\sigma \). □

REFERENCES

[1] D. Mo and M. F. Duarte, “Compressive Parameter Estimation with Earth Mover’s Distance via K-Median Clustering,” in Proc. SPIE Wavelets and Sparsity XV, vol. 8858, San Diego, CA, Aug. 2013.
[2] R. G. Baraniuk, “Compressive Sensing,” IEEE Signal Proc. Mag., vol. 24, no. 4, pp. 118–121, Jul. 2007.
[3] E. J. Candès, “Compressive Sampling,” in Proc. Int. Congr. Math. (ICM), vol. 3, Madrid, Spain, Aug. 2006, pp. 1433–1452.
[4] D. L. Donoho, “Compressed Sensing,” IEEE Trans. Inf. Theory, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.
[5] V. Cevher, A. C. Gurbuz, J. H. McClellan, and R. Chellappa, “Compressive Wireless Arrays for Bearing Estimation,” in IEEE Int. Conf. Acoustics, Speech and Signal Proc. (ICASSP), Las Vegas, NV, Apr. 2008, pp. 2497–2500.
[6] M. F. Duarte, “Localization and Bearing Estimation via Structured Sparsity Models,” in IEEE Stat. Signal Proc. Workshop (SSP), Ann Arbor, MI, Aug. 2012, pp. 333–336.
[7] M. A. Herman and T. Strohmer, “High Resolution Radar via Compressed Sensing,” IEEE Trans. Inf. Theory, vol. 57, no. 6, pp. 2275–2284, Jun. 2009.
[8] H. J. Rad and G. Leus, “Sparsity-Aware TDOA Localization of Multiple Sources,” IEEE Trans. Signal Proc., vol. 61, no. 19, pp. 4021–4025, Oct. 2013.
[9] S. Sen, G. Tang, and A. Nehorai, “Multiobjective Optimization of OFDM Radar Waveform for Target Detection,” IEEE Trans. Inf. Theory, vol. 59, no. 2, pp. 639–652, Feb. 2011.
[10] I. Stojanovic, M. Çetin, and W. C. Karl, “Compressed Sensing of Monostatic and Multistatic SAR,” IEEE Geosci. Remote Sens. Lett, vol. 10, no. 6, pp. 1444–1448, 2013.
[11] A. Eftekhar, J. Romberg, and M. B. Wakin, “Matched Filtering From Limited Frequency Samples,” IEEE Trans. Inf. Theory, vol. 59, no. 6, pp. 3475–3496, Jun. 2013.
[12] K. Fyhn, M. F. Duarte, and S. H. Jensen, “Compressive Time Delay Estimation Using Interpolation,” in IEEE Global Conf. Signal and Information Proc. (GlobalSIP), Austin, TX, Dec. 2013, pp. 624–624.

[13] C. D. Austin, R. L. Moses, J. N. Ash, and E. Ertin, “On the Relation Between Sparse Reconstruction and Parameter Estimation With Model Order Selection,” IEEE J. Sel. Topics Signal Process., vol. 4, no. 3, pp. 560–570, Jun. 2010.

[14] S. Bourguignon, H. Carfantan, and J. Idier, “A Sparsity-Based Method for the Estimation of Spectral Lines From Irregularly Sampled Data,” IEEE J. Sel. Topics Signal Process., vol. 1, pp. 575–585, Dec. 2007.

[15] M. F. Duarte and R. G. Baraniuk, “Spectral Compressive Sensing,” Appl. and Comput. Harmonic Anal., vol. 35, no. 1, pp. 111–129, Jan. 2013.

[16] K. Fyhn, H. Dadkhahi, and M. F. Duarte, “Spectral Compressive Sensing with Polar Interpolation,” in IEEE Int. Conf. Acoustics, Speech and Signal Proc. (ICASSP), Vancouver, Canada, May 2013, pp. 6225–6229.

[17] J. A. Tropp, “Greed is Good: Algorithmic Results for Sparse Approximation,” IEEE Trans. Inf. Theory, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.

[18] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, “Model-Based Compressive Sensing,” IEEE Trans. Inf. Theory, vol. 56, no. 4, pp. 1982–2001, Apr. 2010.

[19] A. Fannjiang and W. Liao, “Coherence Pattern-Guided Compressive Sensing with Unresolved Grids,” SIAM J. Imaging Sci., vol. 5, no. 1, pp. 179–202, Feb. 2012.

[20] W. U. Bajwa, R. Calderbank, and S. Jafarpour, “Why Gabor Frames? Two Fundamental Measures of Coherence and Their Role in Model Selection,” J. of Comm. and Networks, vol. 12, no. 4, pp. 289–307, Aug. 2012.

[21] A. K. Fletcher, S. Rangan, and V. K. Goyal, “Necessary and Sufficient Conditions for Sparsity Pattern Recovery,” IEEE Trans. Inf. Theory, vol. 55, no. 12, pp. 5758–5772, Dec. 2009.

[22] A. Karbasi, A. Hormati, S. Mohajer, and M. Vetterli, “Support Recovery in Compressed Sensing: An Estimation Theoretic Approach,” in IEEE Int Symp. Inf. Theory (ISIT), Seoul, South Korea, Jul. 2009, pp. 679–683.

[23] G. Reeves and M. Gastpar, “The Sampling Rate-Distortion Tradeoff for Sparsity Pattern Recovery in Compressed Sensing,” IEEE Trans. Inf. Theory, vol. 58, no. 5, pp. 3065–3092, May 2012.

[24] Y. Rubner, C. Tomasi, and L. J. Guibas, “A Metric for Distributions with Applications to Image Databases,” in Int. Conf. Comput. Vision (ICCV), Bombay, India, Jan. 1998, pp. 59–66.

[25] R. Gupta, P. Indyk, and E. Price, “Sparse recovery for Earth Mover Distance,” in 48th Ann. Allerton Conf. Commun, Control, and Computing, Monticello, IL, Sep. 2010, pp. 1742–1744.

[26] P. Indyk and E. Price, “K-Median Clustering, Model-Based Compressive Sensing, and Sparse Recovery for Earth Mover Distance,” in Proce. 43th Ann. ACM Symp. Theory of Computing, ser. STOC ’11, San Jose, CA, Jun. 2011, pp. 627–636.

[27] E. J. Candès and T. Tao, “Near-Optimal Signal Recovery From Random Projections: Universal Encoding Strategies?,” IEEE Trans. Inf. Theory, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.

[28] T. Blumensath and M. E. Davies, “Iterative Thresholding for Sparse Approximations,” J. Fourier Analysis and Applicat., vol. 14, no. 5, pp. 629–654, Dec. 2008.
[29] D. Needell and J. A. Tropp, “CoSaMP: Iterative signal recovery from incomplete and inaccurate samples,” Appl. and Comput. Harmonic Anal., vol. 26, no. 3, pp. 301–321, May 2009.
[30] W. Dai and O. Milenkovic, “Subspace Pursuit for Compressive Sensing Signal Reconstruction,” IEEE Trans. Inf. Theory, vol. 55, no. 5, pp. 2230–2249, May 2009.
[31] S. G. Mallat and Z. Zhang, “Matching Pursuits With Time-Frequency Dictionaries,” IEEE Trans. Inf. Theory, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.
[32] J. A. Tropp and A. C. Gilbert, “Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit,” IEEE Trans. Inf. Theory, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.
[33] T. Blumensath and M. E. Davies, “Sampling Theorems for Signals From the Union of Finite-Dimensional Linear Subspaces,” IEEE Trans. Inf. Theory, vol. 55, no. 4, pp. 1872–1882, Apr. 2009.
[34] D. L. Donoho and M. Elad, “Optimally Sparse Representation in General (Non-Orthogonal) Dictionaries via $\ell_1$ Minimization,” in Proc. Natl. Acad. Sci. U.S.A., vol. 100, no. 5, Mar. 2003, pp. 2197–2202.
[35] H. Rauhut, K. Schnass, and P. Vanderheynst, “Compressed Sensing and Redundant Dictionaries,” IEEE Trans. Inf. Theory, vol. 54, pp. 2210–2219, May 2008.
[36] O. Pele and M. Werman, “Fast and Robust Earth Mover’s Distance,” in Int. Conf. Comput. Vision (ICCV), Kyoto, Japan, Sep. 2009, pp. 460–467.
[37] P. S. Bradley, O. L. Mangasarian, and W. N. Street, “Clustering via Concave Minimization,” in Advances in Neural Inf. Proc. Systems (NIPS), vol. 9, Denver, CO, Dec. 1996, pp. 368–374.
[38] P. N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining, 2nd ed. Pearson Education, Limited, Dec. 2014.
[39] F. Giles and W. Pulleyblank, “Total Dual Integrality and Integer Polyhedra,” Linear Algebra and its Applications, vol. 20, pp. 191–196, Jun. 1979.
[40] S. S. Vempala, The Random Projection Method. American Math. Soc., 2005.
[41] A. Eftekhar and M. B. Wakin, “New Analysis of Manifold Embeddings and Signal Recovery from Compressive Measurements,” Appl. and Comput. Harmonic Anal., Apr. 2014, to appear.
[42] G. Tang and B. Recht, “Atomic Decomposition of Mixtures of Translation-Invariant Signals,” in Comput. Advances in Multi-Sensor Adaptive Proc. (CAMSAP), Saint Martin, France, Dec. 2013.
[43] C. Ekanadham, D. Tranchina, and E. P. Simoncelli, “Recovery of Sparse Translation-Invariant Signals With Continuous Basis Pursuit,” IEEE Trans. Inf. Theory, vol. 59, no. 10, pp. 4735–4744, Oct. 2011.
[44] J. Munkres, “Algorithms for the Assignment and Transportation Problems,” J. Society for Industrial and Applied Mathematics, vol. 5, no. 1, pp. 32–38, Mar. 1957.