Grid-less DOA Estimation Using Sparse Linear Arrays Based on Wasserstein Distance

Mianzhi Wang, Student Member, IEEE, Zhen Zhang, Student Member, IEEE, and Arye Nehorai, Life Fellow, IEEE

Abstract—Sparse linear arrays, such as nested and co-prime arrays, are capable of resolving $O(M^2)$ sources by exploiting their so-called difference coarray model. One popular approach to exploit the difference coarray model is to construct an augmented covariance matrix from the sample covariance matrix. By applying common direction-of-arrival (DOA) estimation algorithms to this augmented covariance matrix, more sources than the number of sensors can be identified. In this letter, inspired by the optimal transport theory, we develop a new approach to construct this augmented covariance matrix. We formulate a structured covariance estimation problem that minimizes the Bures-Wasserstein distance between the sample covariance matrix and the subsampled augmented covariance matrix, which can be either casted to a semi-definite programming problem, or directly solved using gradient-based methods. Our approach contributes to a new grid-less DOA estimation algorithm for sparse linear arrays. Numerical examples show that our approach achieves state-of-art estimation performance.

Index Terms—Direction-of-arrival estimation, sparse linear arrays, co-prime and nested arrays, Wasserstein distance, convex optimization

I. INTRODUCTION

DIRECTION-OF-ARRIVAL (DOA) estimation, which recovers the direction information of multiple incoming waves using the measurements of sensor arrays, is one of the major problems in array signal processing [1]–[3]. With uniform linear arrays (ULAs), classical DOA estimation methods, such as MUtiple SIgnal Classification (MUSIC) [4], can resolve up to $M-1$ sources using $M$ sensors. With sparse linear arrays, such as minimum redundancy arrays (MRAs) [5], [6], nested arrays [7], [8], and co-prime arrays [9]–[12], up to $O(M^2)$ uncorrelated sources can be resolved using $M$ sensors by exploiting the difference coarray model.

Various DOA estimation algorithms can be applied to exploit the difference coarray model. These methods can be grouped into two categories: grid-based methods and grid-less methods. Grid-based methods discretize the parameter space of interest into a fine grid and recover the DOAs using sparse enforcing techniques [13]–[18]. On the contrary, grid-less methods do not require the parameters of interest lie on a predefined grid. Therefore, their estimation errors are not limited by the grid size. Many grid-less methods rely on the construction of the covariance matrix of the difference coarray model, namely the augmented covariance matrix, whose size is usually much larger than the sample covariance matrix of the physical array model. Based on this augmented covariance matrix, more sources than the number of sensors can be resolved. Various methods of constructing such an augmented covariance matrix are developed based on spatial-smoothing [7], [19], nuclear-norm minimization [20], maximum-likelihood function with Toeplitz parameterization [21], asymptotically efficient covariance fitting criterion [22], and virtual array interpolation [23].

In this letter, we explore a new way of constructing the augmented covariance matrix inspired by the Wasserstein distance between zero-mean Gaussians, which originates in the optimal transport theory [24], [25]. The Wasserstein distance between zero-mean Gaussians introduces a metric over the space of covariance matrices. Using this metric, we formulate a structured covariance matrix estimation problem to construct the augmented covariance matrix, which can be casted to a semi-definite programming (SDP) problem. We can then estimate the DOAs based on the constructed augmented covariance matrix. To the best knowledge, this is the first time optimal transport theory has been applied to sparse linear arrays, which opens up a new direction for developing DOA estimators for such arrays. Our formulation provides a new grid-less and regularization-free DOA estimation method for sparse linear arrays that achieves state-of-art performance.

Notations: We use $CN(\mu, \Sigma)$ to denote a complex circularly-symmetric Gaussian distribution whose mean is $\mu$ and covariance is $\Sigma$. Given a matrix $A$, we use $\|A\|_F$ to denote its Frobenius norm. Given a Hermitian matrix $H$, $H \succeq 0$ means $H$ is positive semi-definite (PSD). We use $\otimes$, $\odot$, and $\circ$ to denote the Kronecker product, the Khatri-Rao product, and the Hadamard product, respectively. We use $\mathbb{H}^n$ and $\mathbb{T}^n$ to denote the set of $n \times n$ Hermitian and Toeplitz matrices, respectively. Given a set of $n \times n$ matrices, $\mathbb{S}_+^n$, we use $\mathbb{S}_+^n$ to denote its subset of PSD matrices.

II. A REVIEW OF DIRECTION FINDING USING SPARSE LINEAR ARRAYS

We consider an $M$-sensor sparse linear array whose sensors are placed on a grid with grid size $d_0$. We denote the sensor locations by $\mathcal{D} = \{d_1d_0, d_2d_0, \ldots, d_Md_0\}$, where $d_i$ are integers. Without loss of generality, we assume that the first sensor is located at the origin (i.e., $d_1 = 0$), and that $d_1 < d_2 < \cdots < d_M$. Such a sparse linear array can be viewed as a thinned ULA of $M_0 = d_M + 1$ sensors (note that $\{d_1, d_2, \ldots, d_M\} \subseteq \{d_1, d_1 + 1, \ldots, d_M\}$). Assume that $K$ far-field narrow-band sources impinge on the array from the
directions \( \theta = [\theta_1, \theta_2, \ldots, \theta_K]^T \). The \( N \) snapshots received by the array can be expressed as

\[
y(t) = SA_0(\theta)x(t) + n(t), \quad t = 1, 2, \ldots, N.
\]

Here \( x(t) \) denotes the source signals, and \( n(t) \) denotes the additive noise. \( A_0(\theta) = [a_0(\theta_1), a_0(\theta_2), \ldots, a_0(\theta_K)] \) is the steering matrix of an \( M_0 \)-sensor ULA, where

\[
a_0(\theta_k) = \begin{bmatrix} 1 & e^{j2\pi d_0 \sin \theta_k} & \cdots & e^{j2\pi (M_0-1)d_0 \sin \theta_k} \end{bmatrix},
\]

and \( \lambda \) denotes the wavelength of the carrier wave. \( S \in \mathbb{R}^{M \times M_0} \) is a sensor selection matrix encoding the thinning process, where \( S_{mn} \in \{0,1\} \) and \( S_{mn} = 1 \) only if the \( n \)-th sensor in the full ULA is selected as the \( m \)-th sensor in the sparse linear array. Typically, the inter-element spacing \( d_0 \) is chosen to be \( \lambda/2 \) to avoid grating lobes.

We consider the stochastic model [26] and assume that the sources are uncorrelated. Under this signal model,

1) The source signals, \( x(t) \sim CN(0, P) \), where \( P = \text{diag}(p_1, p_2, \ldots, p_K) \) and \( p_k \) denotes the power of the \( k \)-th source;
2) The additive noise vectors, \( n(t) \), are independently and identically distributed as \( CN(0, \sigma^2 I_M) \);
3) Both the source signals and the noise are temporally uncorrelated, and the source signals are uncorrelated from the additive noise.

Consequently, \( y(t) \) are i.i.d. complex circularly-symmetric Gaussians with mean zero and covariance \( R \), where

\[
R = E[yy^H] = SR_0S^T = SA_0PA_0^H \sigma^2 I_M.
\]

Here \( R_0 = A_0PA_0^H + \sigma^2 I \) is the covariance matrix of an \( M_0 \)-sensor ULA. Because \( A_0 \) is the steering matrix of a ULA and \( P \) is diagonal, \( R_0 \) is a Toeplitz matrix.

Vectorizing \( R \) leads to

\[
r = \text{vec}(R) = (S \otimes S)(A_0^\dagger \otimes A_0)P + \sigma \text{vec}(I_M).
\]

Here \((S \otimes S)(A_0^\dagger \otimes A_0)\) can be viewed as the steering matrix of a virtual array whose sensor locations are given by \( D_{co} = \{ (d_m - d_n)_{d0} | m, n = 1, 2, \ldots, M \} \). This virtual array is usually referred as “different coarray”. Consequently, \( r \) resembles a measurement vector of this difference coarray. If \( md0 \in D_{co} \forall m = 0, 1, \ldots, M_0 - 1 \), we say the difference coarray is “hole-free”. It has been shown that, given the sample covariance matrix \( \hat{R} = \frac{1}{N} \sum_{t=1}^{N} y(t)y^H(t) \), by exploiting the difference coarray structure of “hole-free” sparse linear arrays, one can construct an augmented covariance matrix that provides a good estimate of \( R_0 \) [7]. Because \( M_0 \) is typically much larger than \( M \), by applying MUSIC or other spectral estimation algorithms, we can resolve many more sources than the number of sensors.

**Remark 1:** Some sparse linear arrays, such as co-prime arrays, are not “hole-free”, and only a smaller submatrix of \( R_0 \) can be estimated from \( \hat{R} \). As long as the size of this submatrix is greater than \( M \), more sources than the number of sensors can still be resolved. For brevity, we consider only “hole-free” arrays in the following discussion, which can be easily extended to arrays that are not “hole-free”.

In general, the process of recovering \( R_0 \) from \( \hat{R} \) can be described with the following optimization problem:

\[
\min_{\hat{R}_0} \mathcal{d}(\hat{R}, SR_0S^T) + L(R_0), \quad \text{s.t. } R_0 \in \mathcal{C},
\]

where \( \mathcal{d}(\cdot, \cdot) \) is some similarity measure defined over \( \mathbb{H}_+^M \), \( L(R_0) \) captures the regularization terms, and \( \mathcal{C} \) denotes the constraints.

One simple method of estimating \( R_0 \) consists of only two steps: redundancy averaging and direct augmentation [27]–[29]. Denoting this estimated \( \hat{R}_0 \) by \( \hat{R}_{DA} \), it is straightforward to show that \( \hat{R}_{DA} \) is the solution to (P1) by choosing \( \mathcal{d}(\cdot, \cdot) \) as

\[
\mathcal{d}_F(R_1, R_2) = \| R_1 - R_2 \|_F,
\]

removing the regularization \( L(R_0) \), and setting \( \mathcal{C} = T_{+}^{M_0} \). By applying MUSIC to \( \hat{R}_{DA} \), we recover the DA-MUSIC algorithm [19], [29]. One issue with \( \hat{R}_{DA} \) is that it is not guaranteed to be PSD. To tackle this issue, spatial smoothing (SS) instead of direct augmentation can be performed as the second step. The resulting estimate, denoted by \( \hat{R}_{SS} \), is equal to \( \hat{R}_{DA} \hat{R}_{DA}^H/M_0 \). By applying MUSIC to \( \hat{R}_{SS} \), we recover the SS-MUSIC algorithm [7]. It is also possible to introduce regularizations to further exploit the structure of \( R_0 \). Because \( R_0 - \sigma I \) is low-rank, one popular choice is the nuclear norm [20], [30]. Additional constraints, such as the ones based on correlation subspaces [31], [32], can be introduced if there exists some prior knowledge on the source directions. It should be noted that \( \mathcal{d}_F \) does not capture the geometry of the space of covariance matrices. Therefore, DOA estimates obtained from \( \hat{R}_{DA} \) or \( \hat{R}_{SS} \) are generally statistically inefficient [29].

Because \( y(t) \) follows a zero-mean complex Gaussian distribution, another popular choice of \( \mathcal{d}(\cdot, \cdot) \) is the KL-divergence between two zero-mean complex Gaussian distributions:

\[
d_{KL}(R_1, R_2) = \mathcal{d}_{KL}(\hat{R}_0, \hat{R}, \mathcal{C}) = \frac{1}{2} \log \frac{|\hat{R}|}{|\hat{R}|} - \mathcal{d}(\hat{R}_0, \hat{R}, \mathcal{C}) - \frac{1}{2} \log 2\pi e |\hat{R}|.
\]

After removing the constants and the regularization term, and setting \( \mathcal{C} = T_{+}^{M_0} \cap \mathbb{H}_+^{M_0} \), (P1) becomes

\[
\min_{\hat{R}_0} \mathcal{d}(\hat{R}, SR_0S^T) + L(R_0), \quad \text{s.t. } R_0 \in T_{+}^{M_0} \cap \mathbb{H}_+^{M_0},
\]

which corresponds to the maximum-likelihood based structured covariance estimation problem. The problem (P2) is non-trivial, whose suboptimal solutions can be obtained via fixed-point like iterations [21], [33]. To tackle the difficulty in solving (P2), the authors of [16] and [22] suggested using the following covariance fitting criterion instead of \( d_{KL} \):

\[
d_{SPA}(R_1, R_2) = \| R_1^{-\frac{1}{2}} (R_1 - R_2) R_2^{-\frac{1}{2}} \|_F.
\]

In [22], the authors showed that resulting optimization problem can be converted to a SDP problem, which can be numerically solved with off-the-shelf solvers such as SDPT3 [34].

### III. WASSERSTEIN DISTANCE INSPIRED COVARIANCE MATRIX ESTIMATION

In this section, we introduce our new Wasserstein distance based approach, show its SDP formulation, and derive the
gradients of the resulting optimization problem. Given two probability distributions, \( \mu \) and \( \nu \), on \( \mathbb{R}^n \), the Kantorovich’s two formulation of the optimal transport problem minimizes the following objective function:

\[
\inf_{q \in \mathcal{Q}(\mu, \nu)} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|x - y\|^2 \, dq(x, y),
\]

where the cost function is chosen to be \( \| \cdot \|_2^2 \). \( \mathcal{Q}(\mu, \nu) \) is the set of joint probability distributions on \( \mathbb{R}^n \times \mathbb{R}^n \) whose marginals are given by \( \mu \) and \( \nu \), respectively. The square root of (7) defines the so-called 2-Wasserstein distance over the probability distributions on \( \mathbb{R}^n \) [24]. When both \( \mu \) and \( \nu \) are proper zero-mean Gaussians\(^1\) with covariance matrix \( \hat{R} \) and \( R \), respectively, the optimization problem (7) has a closed-form solution [35], \( d_{\text{W}G}(\hat{R}, R) \), which is given by

\[
d_{\text{W}G}(\hat{R}, R) = \text{tr} \left( \hat{R} + R - 2(\hat{R}^T R)^{\frac{1}{2}} \right),
\]

where \( (\cdot)^{\frac{1}{2}} \) denotes the principal matrix square root\(^2\). It has been shown that \( d_{\text{W}G}(\hat{R}, R) \) introduces a metric on \( \mathbb{H}^n_+ \), namely the Bures-Wasserstein metric [36]. It is straightforward to verify that \( d_{\text{W}G}(\hat{R}, R) \) is generally not equal to \( d_{\text{W}G}(\hat{R} - R, 0) \). Hence, \( d_{\text{W}G}(\hat{R}, R) \) captures the non-Euclidean geometry of covariance matrices.

Combining (P1) and (8), we can formulate a structured covariance estimation problem while considering the geometry of covariance matrices:

\[
\min_{\hat{R}} \left\{ \text{tr} \left( \hat{R} + R - 2(\hat{R}^T R)^{\frac{1}{2}} \right) \right\},
\text{ s.t. } R = S \hat{R} S^T, \quad \hat{R} \in \mathbb{S}_+^n \cap \mathbb{H}^n_+.
\]

Note that \( R \succeq 0 \) guarantees that \( R = S \hat{R} S^T \succeq 0 \).

We next analyze the convexity of (P3). Here the constraint set \( \mathbb{S}_+^n \cap \mathbb{H}^n_+ \) is a convex cone, and \( \text{tr}(R) \) is linear function of \( \hat{R} \). Because that \( \text{tr}(X^2) \) is strictly concave when \( X \in \mathbb{H}^n_+ \) [36], and that \( \hat{R}^T R \) is a linear function of \( \hat{R} \), \( \text{tr} \left( (\hat{R}^T R)^{\frac{1}{2}} \right) \) is concave in \( \hat{R} \). Therefore, the optimization problem (P3) is convex. More specifically, when the array is “hole-free” and \( \hat{R} \) is positive definite, \( \hat{R}^T S \hat{R} S^T R \) defines a bijection. Consequently, (P3) is strictly convex and admits a unique solution.

Analytically solving (P3) turns out to be non-trivial due to the matrix square root operation. It has been shown in [35] that \( d_{\text{W}G}(\hat{R}, R) \) can be obtained by solving the following optimization problem:

\[
\min_{\hat{R}} \text{tr} \left( \hat{R} + R - 2(\hat{R}^T R)^{\frac{1}{2}} \right), \quad \text{s.t. } \begin{bmatrix} \hat{R} & V \\ V^T & R \end{bmatrix} \succeq 0.
\]

Therefore, we can cast (P3) into the following SDP problem:

\[
\min_{\hat{R}, V} \text{tr} \left( \hat{R} + S \hat{R} S^T - V - V^H \right), \quad \text{s.t. } \begin{bmatrix} S \hat{R} S^T & V \\ V^T & R \end{bmatrix} \succeq 0, \quad \hat{R} \in \mathbb{S}_+^n \cap \mathbb{H}^n_+.
\]

(10)\(^3\) where \( \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n \) are the eigenvalues of \( \Lambda \) ordered non-increasingly.

\( \hat{R}^T S \hat{R} S^T R \) defines a bijection. Consequently, (P3) is strictly convex and admits a unique solution.

1An \( n \)-dimensional complex circularly-symmetric Gaussian distribution can be fully characterized by a 2\( n \)-dimensional real Gaussian distribution. Therefore, similar results follow for complex \( \hat{R}_1 \) and \( R_2 \) [35].

2Given a PSD matrix \( A \) with eigendecomposition \( \Lambda \Lambda^H \), its principal matrix square root is \( E \text{diag}(\lambda_1^\frac{1}{2}, \ldots, \lambda_n^\frac{1}{2}) \Lambda^H \).

3The full gradient of (P5) with respect to \( c_i \) is given by \( g_i^{(1)} + g_i^{(2)} + g_i^{(3)} \), and gradient-based methods can then be applied to solve (P5).
Fig. 1: (a) MSE vs. number of snapshots, $N$, for different DOA estimation methods, averaged from 3000 Monte Carlo trials; (b) MSE vs. SNR for different DOA estimation methods, averaged from 3000 Monte Carlo trials; (c) CPU time vs. $N_1$ for three different algorithms. In (a) and (b), the solid lines with markers denote the MSEs, and the dashed line denotes the CRB.

Noting that $g_i^{(1)}$ can be precomputed, the worst case computational complexity is mainly determined by $g_i^{(2)}$ and $g_i^{(3)}$. Because eigendecomposition and matrix multiplication are both $O(n^3)$ operations, $g_i^{(2)}$ and $g_i^{(3)}$ cost at most $O(M_i^3)$ operations. Because $i$ ranges from 1 to $2M_0 - 1$, the worst case computational complexity of gradient-based methods is $O(kM_0^3)$, where $k$ is the number of descent steps.

We close this section by considering an interesting case: $K = 1$ and $\sigma = 0$. In such a case, $R = pa(\theta)a(\theta)$ is a rank one matrix parameterized by a single parameter, $\theta$, and $R^{\frac{1}{2}} = \sqrt{p/Ma(\theta)a(\theta)}$. Given the sample covariance matrix, $\hat{R}$, we have

$$\min_{\theta} d_{WG}^2(\hat{R}, R) \iff \min_{\hat{\theta}} pM - \frac{p}{M} \text{tr}\left( (aa^H \hat{R}a) \hat{\theta}^2 \right)$$

$$\iff \max_{\theta} (a^H \hat{R}a) \hat{\theta}^2 \text{tr}\left( \frac{1}{\sqrt{M}}(aa^H) \right)$$

$$\iff \max_{\theta} a^H \hat{R}a,$$

which is the objective function of a conventional beamformer [2]. This result gives a geometric interpretation of the conventional beamformer. By maximizing $a^H \hat{R}a$, we are finding the DOA that minimizes the Bures-Wasserstein distance between the sample covariance matrix $\hat{R}$ and the structured covariance matrix $R = pa(\theta)a^H(\theta)$.

IV. NUMERICAL EXAMPLES

We consider a nested array generated by the parameter pair $(4, 4)$: $\{0, 1, 2, 3, 4, 9, 14, 19\}d_0$. We uniformly place nine off-grid sources, which is greater than the number of sensors, within the range $(-70\pi/219, 76\pi/219)$. We consider the following methods to estimate $R_0$:

1) SS: Spatial smoothing after redundancy averaging [29].
2) NN: Nuclear-norm regularized reconstruction [30]. As suggested in the paper, we set the regularization parameter to $2.5 \times 10^{-3}/(\log N)^2 \log M$.
3) SPA: Sparse and parametric approach [22].
4) WG: Solving the SDP formulation (P4).
5) WGGrad: Solving (P5) using gradient descent with backtracking. We set $\mu = 0.01$ and stop when the relative change in the objective function is less than $10^{-7}$.

Once $R_0$ is estimated, the DOAs are obtained using root-MUSIC [40]. For reference, we included the Cramér-Rao bound (CRB) for sparse linear arrays [29], [41]–[43] for comparison.

Fig. 1(a) plots the mean-squared errors (MSEs) of the five methods and the CRB versus the number of snapshots, where the signal-to-noise ratio (SNR) is set to $0$ dB in all trials. Fig. 1(b) plots the MSEs of the five methods and the CRB versus the SNR, where the number of snapshots is fixed to 50 in all trials. We observe that the performance of WGGrad well agrees with that of WG in all cases. SPA has the best overall performance because $d_{SPA}$ asymptotically approximates the maximum-likelihood function [22], [44], [45]. The performance of our proposed methods, WS and WGGrad, closely matches that of the SPA, and is significantly better than SS and NN.

To compare the CPU time of WS, WSG, and SPA, we consider nested arrays generated by the parameter pair $(N_1, N_1)$ and vary $N_1$ from 3 to 12. The number of snapshots, $N$, is fixed to 50 in all trials. All trials are performed on a server with an Intel Xeon E5-2650v3 CPU. We plot the results in Fig. 1(c). We observe that as the size of the array grows, SPA and WG consumes significantly more CPU time than WGGrad, as predicted by our complexity analysis in Section III. Therefore, it is computationally more efficient to use WGGrad when the sparse linear array is large.

V. CONCLUDING REMARKS

In this letter, we proposed a new grid-less DOA estimation algorithm based on the Bures-Wasserstein metric between covariance matrices. We first construct an augmented covariance matrix by minimizing the Bures-Wasserstein distance between the sample covariance matrix and the subsampled augmented covariance matrix, and then estimate the DOAs by applying root-MUSIC to the augmented covariance matrix. We showed that the minimization problem can be solved with either SDP or gradient-based methods. It is observed through numerical examples that our algorithm achieves state-of-art DOA estimation performance. Our result will inspire further research on the application of optimal transport theory to array signal processing. Possible future research includes analyzing the connection between our algorithm and the SPA algorithm as well as developing more efficient numerical methods.