Marius Pesavento, Minh Trinh-Hoang, and Mats Viberg

Three More Decades in Array Signal Processing Research

An optimization and structure exploitation perspective

©SHUTTERSTOCK.COM/TRIFF

Digital Object Identifier 10.1109/MSP.2023.3255558
Date of current version: 1 June 2023

The signal processing community is currently witnessing the emergence of sensor array processing and direction-of-arrival (DoA) estimation in various modern applications, such as automotive radar, mobile user and millimeter wave indoor localization, and drone surveillance, as well as in new paradigms, such as joint sensing and communication in future wireless systems. This trend is further enhanced by technology leaps and the availability of powerful and affordable multiantenna hardware platforms.

Introduction

New multiantenna technology has led to the widespread use of such systems in contemporary sensing and communication systems as well as a continuous evolution toward larger multiantenna systems in various application domains, such as massive multiple, input-multiple-output (MIMO) communications systems comprising hundreds of antenna elements. The massive increase of the antenna array dimension leads to unprecedented resolution capabilities, which opens new opportunities and challenges for signal processing. For example, in large MIMO systems, modern array processing methods can be used to estimate and track the physical path parameters, such as DoA, direction of departure, time delay of arrival, and Doppler shift, of tens or hundreds of multipath components with extremely high precision [1]. This parametric approach for massive MIMO channel estimation and characterization benefits from the enhanced resolution capabilities of large array systems and efficient array processing techniques. Direction-based MIMO channel estimation, which has not been possible in small MIMO systems due to the limited number of antennas, not only significantly reduces the complexity but also improves the quality of MIMO channel prediction as the physical channel parameters generally evolve on a much smaller timescale than the MIMO channel coefficients.

The history of advances in superresolution DoA estimation techniques is long, starting from the early parametric multisource methods, such as the computationally expensive

 1053-5888/23©2023IEEE

maximum likelihood (ML) techniques, to the early subspace-based techniques, such as Pisarenko and MUSIC [2]. Inspired by the seminal review article, "Two Decades of Array Signal Processing Research: The Parametric Approach" by Krim and Viberg, published in *IEEE Signal Processing Magazine* [3], we are looking back at another three decades in array signal processing research under the classical narrow-band array processing model based on second-order statistics. We revisit major trends in the field and retell the story of array signal processing from a modern optimization and structure exploitation perspective. In our overview, through prominent examples, we illustrate how different DoA estimation methods can be cast as optimization problems with side constraints originating from prior knowledge regarding the structure of the measurement system. Due to space limitations, our review of the DoA estimation research in the past three decades is by no means complete. For didactic reasons, we mainly focus on developments in the field that easily relate to the traditional multisource estimation criteria in [3] and choose simple illustrative examples.

Model structure, e.g., in the form of a favorable spatial sampling pattern, is exploited for various purposes: either to reduce the computational complexity and to make the estimation computationally tractable or to generally improve the estimation quality.

As many optimization problems in sensor array processing are notoriously difficult to solve exactly due to their nonlinearity and multimodality, a common approach is to apply problem relaxation and approximation techniques in the development of computationally efficient and close-to-optimal DoA estimation methods. The DoA estimation approaches developed in the last 30 years differ in the prior information and model assumptions that are maintained and relaxed during the approximation and relaxation procedure in the optimization.

Along the line of constrained optimization, problem relaxation, and approximation, recently, the partial relaxation (PR) technique has been proposed as a new optimization-based DoA estimation framework that applies modern relaxation techniques to traditional multisource estimation criteria to achieve new estimators with excellent estimation performance at affordable computational complexity. In many senses, it can be observed that the estimators designed under the PR framework admit new insights into existing methods of this well-established field of research [4].

The introduction of sparse optimization techniques for DoA estimation and source localization in the late 2000s marks another methodological leap in the field [5], [6], [7], [8], [9]. These modern optimization-based methods became extremely popular due to their advantages in practically important scenarios where classical subspace-based techniques for DoA estimation often experience a performance breakdown, e.g., in the case of correlated sources, when the number of snapshots is low, or when the model order is unknown. Sparse representation-based methods have been successfully extended to incorporate and exploit various forms of structures, e.g., application-dependent row- and rank-sparse structures [10], [11], that induce joint sparsity to enhance estimation performance in the case of multiple snapshots. In particular array geometries, additional structures, such as Vandermonde and shift invariance, can be used to obtain efficient parameterizations of the array sensing matrix that avoid the usual requirement of sparse reconstruction methods to sample the angular field of view (FoV) on a fine DoA grid [12], [13].

Despite the success of sparsity-based methods, it is, however, often neglected that these methods also have their limitations, such as estimation biases resulting from off-grid errors and the impact of the sparse regularization, high computational complexity, and memory demands as well as sensitivity to the choice of the so-called hyperparameters. In fact, for many practical estimation scenarios, sparse optimization techniques are often outperformed by classical subspace techniques in terms of both the resolution of sources and computational complexity. From the theoretical perspective, performance guarantees of sparse methods are generally available only under the condition of the minimum angular separation between the source signals [9]. Therefore, it is important to be aware of these limitations and to appreciate the benefits of both traditional and modern optimization-based DoA estimation methods.

The narrow-band far-field point source signals with perfectly calibrated sensor arrays and centralized processing architectures have been fundamental assumptions in the past. With the trend of wider reception bandwidth on the one hand, and larger aperture and distributed array on the other hand, the aforementioned assumptions appeared restrictive and often impractical. Distributed sensor networks have emerged as a scalable solution for source localization where sensors exchange data locally within their neighborhood and in-network processing is used for distributed source localization with low communication overhead [14]. Furthermore, DoA estimation methods for partly calibrated subarray systems have been explored [15], [16].

Model structure, e.g., in the form of a favorable spatial sampling pattern, is exploited for various purposes: either to reduce the computational complexity and to make the estimation computationally tractable or to generally improve the estimation quality. In this article, we revisit the major trends of structure exploitation in sensor array signal processing. Along this line, we consider advanced spatial sampling concepts designed in recent years, including minimum redundancy [17], augmentable [18], nested [19], and coprime arrays [20], [21]. The aforementioned spatial sampling patterns were designed to facilitate new DoA estimation methods with the capability of resolving significantly more sources than sensors in the array. This is different from conventional sampling patterns, e.g., uniform linear

array (ULA), where the number of identifiable sources is always smaller than the number of sensors.

Signal model

In this overview article, we consider the narrow-band point source signal model. Under this signal model, we are interested in estimating the DoAs, i.e., the parameter vector $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_N]^{\mathsf{T}}$, of N far-field narrow-band sources impinging on a sensor array composed of M sensors from noisy measurements.

We assume that the DoA θ_n lies in the FoV Θ, i.e., $\theta_n \in \Theta$. Let $\boldsymbol{x}(t) = \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{s}(t) + \boldsymbol{n}(t)$ denote the linear array measurement model at time instant t where $\boldsymbol{s}(t)$ and $\boldsymbol{n}(t)$ denote the signal waveform vector and the sensor noise vector, respectively. The sensor noise $\boldsymbol{n}(t)$ is commonly assumed to be a zero-mean spatially white complex circular Gaussian random process with a covariance matrix $\nu \boldsymbol{I}_M$. The steering matrix $\boldsymbol{A}(\boldsymbol{\theta}) \in \mathcal{A}_N$ lives on an N-dimensional array manifold $\mathcal{A}_N$, which is defined as

The introduction of sparse optimization techniques for DoA estimation and source localization in the late 2000s marks another methodological leap in the field.

$$\mathcal{A}_N = \{\boldsymbol{A} = [\boldsymbol{a}(\vartheta_1), \ldots, \boldsymbol{a}(\vartheta_N)] \,|\, \vartheta_1 < \ldots < \vartheta_N \text{ and } \vartheta_n \in \Theta \text{ for all } n = 1, \ldots, N\}. \tag{1}$$

In (1), the steering vector $\boldsymbol{a}(\theta) = [e^{-\mathrm{j}\pi d_1 \cos(\theta)}, e^{-\mathrm{j}\pi d_2 \cos(\theta)}, \ldots, e^{-\mathrm{j}\pi d_M \cos(\theta)}]^{\mathsf{T}}$ denotes, e.g., the array response of a linear array with sensor positions $d_1, \ldots, d_M$ in half wavelength for a narrow-band signal impinging from the direction θ. The steering matrix $\boldsymbol{A}(\boldsymbol{\theta}) = [\boldsymbol{a}(\theta_1), \ldots, \boldsymbol{a}(\theta_N)]$ must satisfy certain regularity conditions so that the estimated DoAs can be uniquely identifiable up to a permutation from the noiseless measurement. Mathematically, the unique identifiability condition requires that if $\boldsymbol{A}(\boldsymbol{\theta}^{(1)})\boldsymbol{s}^{(1)}(t) = \boldsymbol{A}(\boldsymbol{\theta}^{(2)})\boldsymbol{s}^{(2)}(t)$ for $t = 1, \ldots, T$, then $\boldsymbol{\theta}^{(1)}$ is a permutation of $\boldsymbol{\theta}^{(2)}$. Generally, this condition must be verified for any sensor structure and the corresponding FoV. Specifically, it can be shown that if the array manifold is free from ambiguities, i.e., if any oversampled steering matrix $\boldsymbol{A}(\boldsymbol{\theta}) \in \mathcal{A}_K$ of dimension $M \times K$ with $K \geq M$ has a Kruskal rank $q(\boldsymbol{A}(\boldsymbol{\theta})) = M$, then N DoAs with $N < M$ can be uniquely determined from the noiseless measurement [3]. Equivalently, any set of M column vectors $\{\boldsymbol{a}(\theta_1), \ldots, \boldsymbol{a}(\theta_M)\}$ with M distinct DoAs $\theta_1, \ldots, \theta_M \in \Theta$ is linearly independent.

In the so-called conditional signal model, the waveform vector $\boldsymbol{s}(t)$ is assumed to be deterministic such that $\boldsymbol{x}(t) \sim \mathcal{N}_{\mathrm{C}}(\boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{s}(t), \nu \boldsymbol{I}_M)$. The unknown noise variance ν and the signal waveform $\boldsymbol{S} = [\boldsymbol{s}(1), \ldots, \boldsymbol{s}(T)]$ are generally not of interest in the context of DoA estimation, but they are necessary components of the signal model. In contrast, in the unconditional signal model, the waveform is assumed to be zero-mean complex circular Gaussian such that $\boldsymbol{x}(t) \sim \mathcal{N}_{\mathrm{C}}(\boldsymbol{0}_M, \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{P}\boldsymbol{A}^{\mathsf{H}}(\boldsymbol{\theta}) + \nu \boldsymbol{I}_M)$, where the noise variance ν and the waveform covariance matrix $\boldsymbol{P} = \mathbb{E}\{\boldsymbol{s}(t)\,\boldsymbol{s}^{\mathsf{H}}(t)\}$ are considered as unknown parameters. We assume, if not stated otherwise, that the signals are not fully correlated, i.e., $\boldsymbol{P}$ is nonsingular.

Note that in practical wireless communication or radar applications, the received signal may be broadband. Such scenarios require extensions of the narrow-band signal model, e.g., to subband processing or the multidimensional harmonic retrieval, which is, however, out of scope of this article.

Cost function and concentration

Parametric methods for DoA estimation can generally be cast as optimization problems with multivariate objective functions that depend on a particular data matrix $\boldsymbol{Y}$ obtained from the array measurements $\boldsymbol{X} = [\boldsymbol{x}(1), \ldots, \boldsymbol{x}(T)]$ through a suitable mapping, the unknown DoA parameters of interest $\boldsymbol{\theta}$, and the unknown nuisance parameters, which we denote by the vector $\boldsymbol{\alpha}$. Hence, the parameter estimates are computed as the minimizer of the corresponding optimization problem with the objective function $f(\boldsymbol{Y} \,|\, \boldsymbol{A}(\boldsymbol{\theta}), \boldsymbol{\alpha})$ as follows:

$$A(\hat{\boldsymbol{\theta}}) = \underset{\boldsymbol{A}(\boldsymbol{\theta}) \in \mathcal{A}_N}{\operatorname{argmin}} \min_{\boldsymbol{\alpha}} f(\boldsymbol{Y} \,|\, \boldsymbol{A}(\boldsymbol{\theta}), \boldsymbol{\alpha}). \tag{2}$$

Remark that in (2), we make no restriction on how the data matrix $\boldsymbol{Y}$ is constructed from the measurement matrix $\boldsymbol{X}$. For example, in the most trivial case, the data matrix $\boldsymbol{Y}$ can directly represent the array measurement matrix, i.e., $\boldsymbol{Y} = \boldsymbol{X}$. However, for other optimization criteria, the data matrix $\boldsymbol{Y}$ can be the sample covariance matrix, i.e., $\boldsymbol{Y} = \hat{\boldsymbol{R}} = (1/T)\boldsymbol{X}\boldsymbol{X}^{\mathsf{H}}$ as a sufficient statistics, or even the signal eigenvectors $\boldsymbol{Y} = \hat{\boldsymbol{U}}_{\mathrm{s}}$ (or the noise eigenvectors $\boldsymbol{Y} = \hat{\boldsymbol{U}}_{\mathrm{n}}$) obtained from the eigendecomposition $\hat{\boldsymbol{R}} = \hat{\boldsymbol{U}}_{\mathrm{s}}\hat{\boldsymbol{\Lambda}}_{\mathrm{s}}\hat{\boldsymbol{U}}_{\mathrm{s}}^{\mathsf{H}} + \hat{\boldsymbol{U}}_{\mathrm{n}}\hat{\boldsymbol{\Lambda}}_{\mathrm{n}}\hat{\boldsymbol{U}}_{\mathrm{n}}^{\mathsf{H}}$ where $\hat{\boldsymbol{\Lambda}}_{\mathrm{s}} = \operatorname{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_N)$ contains the N-largest eigenvalues of $\hat{\boldsymbol{R}}$. In Table 1, some prominent examples of multisource estimation methods are listed: deterministic ML (DML) [2, Sec. 8.5.2], weighted subspace fitting (WSF) [22], and covariance matching estimation techniques (COMET) [23].

As we are primarily interested in estimating the DoA parameters $\boldsymbol{\theta}$, a common approach is to concentrate the objective function with respect to all (or only part of) the nuisance parameters $\boldsymbol{\alpha}$. In the case that a closed-form minimizer of the nuisance parameters w.r.t. the remaining parameters exists, the expression of this minimizer can be inserted back to the original objective function to obtain the concentrated optimization problem. More specifically, let $\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta})$ denote the minimizer of the full problem for the nuisance parameter vector $\boldsymbol{\alpha}$ as a function of $\boldsymbol{\theta}$, i.e., $\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}) = \operatorname{argmin}_{\boldsymbol{\alpha}} f(\boldsymbol{Y} \,|\, \boldsymbol{A}(\boldsymbol{\theta}), \boldsymbol{\alpha})$. The concentrated objective function $g(\boldsymbol{Y} \,|\, \boldsymbol{A}(\boldsymbol{\theta})) = f(\boldsymbol{Y} \,|\, \boldsymbol{A}(\boldsymbol{\theta}), \hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}))$ then depends only on the DoAs $\boldsymbol{\theta}$. Apart from the reduction of dimensionality, the concentrated versions of multisource optimization problems often admit appealing interpretations. In Table 1, the concentrated criteria corresponding to the previously considered full-parameter multisource criteria are provided. We observe, e.g., in the case of the concentrated

DML and the WSF criteria, that at the optimum, the residual signal energy contained in the nullspace of the steering matrix is minimized.

Due to the complicated structure of the array manifold $\mathcal{A}_N$ in (1), the concentrated objective function $g(\boldsymbol{Y}|\boldsymbol{A}(\boldsymbol{\theta}))$ is, for common choices in Table 1, highly nonconvex and multimodal w.r.t. the DoA parameters $\boldsymbol{\theta}$. Consequently, the concentrated cost function contains a large number of local minima in the vicinity of the global minimum. This can, e.g., be observed in Figure 1, where the cost function of the DML estimator is depicted. While multisource estimation criteria generally show unprecedented asymptotic as well as threshold performance for low sample size, signal-to-noise ratio (SNR), and closely spaced sources, their associated computational cost is unsuitable in many practical applications. The exact minimization generally requires an N-dimensional search over the FoV, which becomes computationally prohibitive even for low source numbers, e.g., $N = 3$.

In the past three decades and beyond, significant efforts have been made to devise advanced DoA estimation algorithms that exhibit good tradeoffs between performance and complexity. While some very efficient methods have been proposed in a different context and based on pure heuristics, in this feature article, we focus on optimization-based estimators that stem, in some way or another, from multisource optimization problems for the classical array processing model (compare Table 1). Considering the array processing literature, a vast amount of estimators proposed in the past years can be derived from multisource optimization problems. Optimization-based estimators have the advantage that they are not only well-motivated but also intuitively interpretable and flexible for generalization to more sophisticated

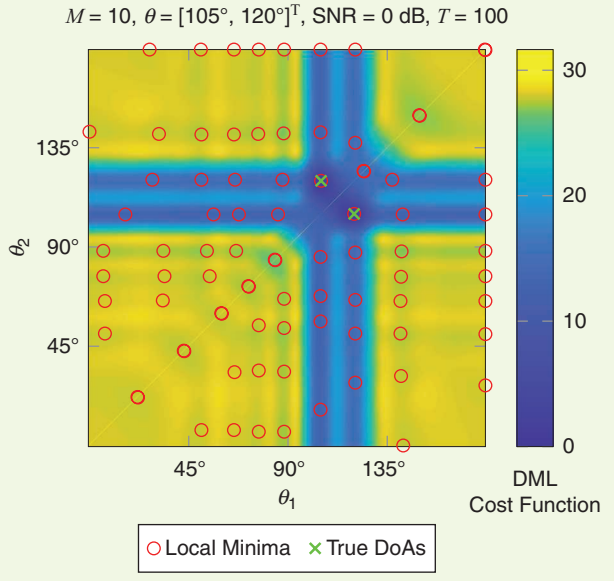

FIGURE 1. An example of the DML cost function for two sources evaluated over the FoV. Multiple local minima are observed. Consequently, local optimization search cannot guarantee to converge to the global minimum.

Table 1. Conventional DoA estimators.

		Full Dimension	Partial Relaxation	Single-Source Approximation
		DML	**PR-DML**	**Conventional Beamformer**
Signal fitting	Original	$\underset{\substack{\boldsymbol{A}\in\mathcal{A}_N \\ \boldsymbol{S}\in\mathbb{C}^{N\times T}}}{\operatorname{argmin}} \|\boldsymbol{X}-\boldsymbol{A}\boldsymbol{S}\|_{\mathrm{F}}^2$	${}^{N}\underset{\boldsymbol{a}\in\mathcal{A}_1}{\operatorname{argmin}} \min_{\boldsymbol{s},\boldsymbol{B},\boldsymbol{J}} \left\|\boldsymbol{X}-\boldsymbol{a}\boldsymbol{s}^{\mathsf{T}}-\boldsymbol{B}\boldsymbol{J}\right\|_{\mathrm{F}}^2$	${}^{N}\underset{\boldsymbol{a}\in\mathcal{A}_1}{\operatorname{argmin}} \min_{\boldsymbol{s}\in\mathbb{C}^T} \left\|\boldsymbol{X}-\boldsymbol{a}\boldsymbol{s}^{\mathsf{T}}\right\|_{\mathrm{F}}^2$
	Concentrated	$\underset{\boldsymbol{A}\in\mathcal{A}_N}{\operatorname{argmin}} \operatorname{tr}(\boldsymbol{\Pi}_{\boldsymbol{A}}^{\perp}\boldsymbol{X}\boldsymbol{X}^{\mathsf{H}})$	${}^{N}\underset{\boldsymbol{a}\in\mathcal{A}_1}{\operatorname{argmin}} \sum_{k=N}^{M}\lambda_k(\boldsymbol{\Pi}_{\boldsymbol{a}}^{\perp}\boldsymbol{X}\boldsymbol{X}^{\mathsf{H}})$	${}^{N}\underset{\boldsymbol{a}\in\mathcal{A}_1}{\operatorname{argmin}} \operatorname{tr}(\boldsymbol{\Pi}_{\boldsymbol{a}}^{\perp}\boldsymbol{X}\boldsymbol{X}^{\mathsf{H}})$
		WSF	**PR-WSF**	**Variant of Weighted MUSIC***
Subspace fitting	Original	$\underset{\substack{\boldsymbol{A}\in\mathcal{A}_N, \\ \boldsymbol{V}\in\mathbb{C}^{N\times N}}}{\operatorname{argmin}} \left\|\hat{\boldsymbol{U}}_{\mathrm{s}}\boldsymbol{W}^{\frac{1}{2}}-\boldsymbol{A}\boldsymbol{V}\right\|_{\mathrm{F}}^2$	${}^{N}\underset{\boldsymbol{a}\in\mathcal{A}_1}{\operatorname{argmin}} \min_{\boldsymbol{v},\boldsymbol{B},\boldsymbol{Q}} \left\|\hat{\boldsymbol{U}}_{\mathrm{s}}\boldsymbol{W}^{\frac{1}{2}}-\boldsymbol{a}\boldsymbol{v}^{\mathsf{T}}-\boldsymbol{B}\boldsymbol{Q}\right\|_{\mathrm{F}}^2$	${}^{N}\underset{\boldsymbol{a}\in\mathcal{A}_1}{\operatorname{argmin}} \min_{\boldsymbol{v}\in\mathbb{C}^N} \left\|\hat{\boldsymbol{U}}_{\mathrm{s}}\boldsymbol{W}^{\frac{1}{2}}-\boldsymbol{a}\boldsymbol{v}^{\mathsf{T}}\right\|_{\mathrm{F}}^2$
	Concentrated	$\underset{\boldsymbol{A}\in\mathcal{A}_N}{\operatorname{argmin}} \operatorname{tr}(\boldsymbol{\Pi}_{\boldsymbol{A}}^{\perp}\hat{\boldsymbol{U}}_{\mathrm{s}}\boldsymbol{W}\hat{\boldsymbol{U}}_{\mathrm{s}}^{\mathsf{H}})$	${}^{N}\underset{\boldsymbol{a}\in\mathcal{A}_1}{\operatorname{argmin}} \sum_{k=N}^{M}\lambda_k(\boldsymbol{\Pi}_{\boldsymbol{a}}^{\perp}\hat{\boldsymbol{U}}_{\mathrm{s}}\boldsymbol{W}\hat{\boldsymbol{U}}_{\mathrm{s}}^{\mathsf{H}})$	${}^{N}\underset{\boldsymbol{a}\in\mathcal{A}_1}{\operatorname{argmin}} \operatorname{tr}(\boldsymbol{\Pi}_{\boldsymbol{a}}^{\perp}\hat{\boldsymbol{U}}_{\mathrm{s}}\boldsymbol{W}\hat{\boldsymbol{U}}_{\mathrm{s}}^{\mathsf{H}})$
		COMET	**PR-CCF**	**Variant of Capon Beamformer****
Covariance fitting	Original	$\underset{\substack{\boldsymbol{A}\in\mathcal{A}_N \\ \boldsymbol{P},\nu}}{\operatorname{argmin}} \left\|\hat{\boldsymbol{R}}^{-\frac{1}{2}}(\hat{\boldsymbol{R}}-\boldsymbol{R})\hat{\boldsymbol{R}}^{-\frac{1}{2}}\right\|_{\mathrm{F}}^2$ subject to $\boldsymbol{R}=\boldsymbol{A}\boldsymbol{P}\boldsymbol{A}^{\mathsf{H}}+\nu\boldsymbol{I}$	${}^{N}\underset{\boldsymbol{a}\in\mathcal{A}_1}{\operatorname{argmin}} \min_{\sigma_{\mathrm{s}}^2\geq 0,\boldsymbol{G}} \left\|\hat{\boldsymbol{R}}-\sigma_{\mathrm{s}}^2\boldsymbol{a}\boldsymbol{a}^{\mathsf{H}}-\boldsymbol{G}\boldsymbol{G}^{\mathsf{H}}\right\|_{\mathrm{F}}^2$ subject to $\hat{\boldsymbol{R}}-\sigma_{\mathrm{s}}^2\boldsymbol{a}\boldsymbol{a}^{\mathsf{H}}-\boldsymbol{G}\boldsymbol{G}^{\mathsf{H}}\succeq\boldsymbol{0}$ $\operatorname{rank}(\boldsymbol{G})\leq N-1$	${}^{N}\underset{\boldsymbol{a}\in\mathcal{A}_1}{\operatorname{argmin}} \min_{\sigma_{\mathrm{s}}^2\geq 0} \left\|\hat{\boldsymbol{R}}-\sigma_{\mathrm{s}}^2\boldsymbol{a}\boldsymbol{a}^{\mathsf{H}}\right\|_{\mathrm{F}}^2$ subject to $\hat{\boldsymbol{R}}-\sigma_{\mathrm{s}}^2\boldsymbol{a}\boldsymbol{a}^{\mathsf{H}}\succeq\boldsymbol{0}$
	Concentrated	See [23, eq. (35)]	${}^{N}\underset{\boldsymbol{a}\in\mathcal{A}_1}{\operatorname{argmin}} \sum_{k=N}^{M}\lambda_k^2\left(\hat{\boldsymbol{R}}-\frac{1}{\boldsymbol{a}^{\mathsf{H}}\hat{\boldsymbol{R}}^{-1}\boldsymbol{a}}\boldsymbol{a}\boldsymbol{a}^{\mathsf{H}}\right)$	${}^{N}\underset{\boldsymbol{a}\in\mathcal{A}_1}{\operatorname{argmin}} \left\|\hat{\boldsymbol{R}}-\frac{1}{\boldsymbol{a}^{\mathsf{H}}\hat{\boldsymbol{R}}^{-1}\boldsymbol{a}}\boldsymbol{a}\boldsymbol{a}^{\mathsf{H}}\right\|_{\mathrm{F}}^2$

$\boldsymbol{\Pi}_{\boldsymbol{A}}^{\perp} = \boldsymbol{I}-\boldsymbol{A}(\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathsf{H}}$ denotes the orthogonal projector onto the nullspace spanned by the columns of $\boldsymbol{A}$. The code for the different variants of the PR methods can be downloaded at https://github.com/PartialRelaxationMethods.

*Conventional weighted MUSIC algorithm (e.g., see [2, eq. (9.258)]) applies the weighting on the noise subspace. This variant applies the weighting on the signal subspace.

**Note that the optimizer $\hat{\sigma}_{\mathrm{s}}^2$ is the spectrum of the Capon Beamformer. The null spectrum of this estimator contains both the spectra of the Conventional Beamformer and the Capon Beamformer.

realistic array signal models. Interestingly, some of the estimators in Table 1 were initially derived by heuristics and are reintroduced here from the perspective of multisource optimization problems.

Modern convex optimization for DoA estimation

The progress in modern convex optimization theory and the emergence of efficient constrained optimization solvers with the turn of the millennium, such as, e.g., the SeDuMi software for solving semidefinite programs, had a significant impact on the research across disciplines in the signal processing, communication, and control communities. In fact, it comes as no surprise that the advances in sensor array signal processing of the past three decades are well aligned with this trend that facilitates advanced constrained optimization-based design approaches. Three closely related universal concepts have been intensively used in array signal processing to make optimization-based estimation procedures numerically stable and computationally feasible. These are: 1) structure exploitation, 2) approximation, and 3) relaxation.

- *Structure exploitation:* This refers to techniques that make use of particular redundancies in the measurement system to introduce convenient data reorganizations and reparameterizations. Examples are methods particularly designed for uniform, shift-invariant, and coprime array geometries.
- *Problem approximation:* These techniques provide local approximations of the multidimensional multimodal nonconvex objective function with the goal to decompose a complex problem into several subproblems. Each subproblem, whose minimizer is much generally simpler to obtain than that of the original problem, is solved in parallel or sequentially, ideally in closed form. Examples are the expectation-maximization algorithm, the orthogonal matching pursuit (OMP), and the single-source approximation methods.
- *Problem relaxation:* Problem relaxation techniques in DoA estimation aim at simplifying the complicated manifold structure associated with the estimation problem. The manifold relaxation is carried out, e.g., to convexify the constraint sets in the associated optimization problems such that numerical methods can be applied.

Approximation and relaxation techniques have in common that they are used to deliberately ignore some parts of the problem structure at the expense of the optimality or performance of the solution. The objective is to simplify the problem so that efficient suboptimal solutions can be obtained that, in many cases, are close to optimal and often even admit performance guarantees. The DoA estimators reviewed in this overview article apply one or more of the aforementioned optimization concepts, as explained in more detail in the following sections.

Single-source approximation

Spectral-based DoA estimation methods, like the popular MUSIC algorithm, belong to the class of single-source approximation methods. In contrast to the full parameter search of minimizing the multisource objective $f(\boldsymbol{Y}|\boldsymbol{A}(\boldsymbol{\theta}), \boldsymbol{\alpha})$ over the N-source signal model with the array manifold $\mathcal{A}_N$ and nuisance parameter vector $\boldsymbol{\alpha}$, the optimization problem in the single-source approximation approach is simplified, and the optimization is carried out only over a single-source model with array manifold, i.e., $\boldsymbol{A}(\boldsymbol{\theta}) \to \boldsymbol{a}(\theta) \in \mathcal{A}_1$ and nuisance parameters $\boldsymbol{\alpha} \to \boldsymbol{\alpha}_1$. It is important to note that, while the number of signal components considered in the optimization is reduced in the single-source approximation approach, the data term $\boldsymbol{Y}$ in the objective remains unchanged. The locations $\boldsymbol{a}(\hat{\theta})$ of the N-deepest minima of the so-called null spectrum $f(\boldsymbol{Y}|\boldsymbol{a}(\theta), \boldsymbol{\alpha}_1)$ evaluated for all steering vectors $\boldsymbol{a}(\theta) \in \mathcal{A}_1$ with angles in the FoV are considered as the steering vector of the estimated DoAs. By using the compact notation ${}^{N}\operatorname{argmin}\, g(\cdot)$ to represent the spectral search of the cost function $g(\cdot)$ for the N-deepest local minima, the single-source approximation is formulated as follows:

$$\{\boldsymbol{a}(\hat{\theta})\} = {}^{N}\operatorname*{argmin}_{\boldsymbol{a}(\theta)\in\mathcal{A}_1} \min_{\boldsymbol{\alpha}_1} f(\boldsymbol{Y}|\boldsymbol{a}(\theta), \boldsymbol{\alpha}_1). \tag{3}$$

For clarity, the concept of the single-source approximation and the corresponding spectral search are visualized in Figure 2. As summarized in Table 1, classical spectral search methods, such as the conventional beamformer, Capon beamformer, and MUSIC, can be reformulated as single-source approximations of the corresponding multisource criteria.

Partial Relaxation methods

Similar to the conventional parametric methods, the PR approach considers the signals from all potential source directions in the multisource cost function. However, to make the problem tractable, the array structures of some signal components are relaxed. More precisely, instead of enforcing the steering matrix $\boldsymbol{A} = [\boldsymbol{a}(\theta_1), \ldots, \boldsymbol{a}(\theta_N)]$ to be an element in the highly structured array manifold $\mathcal{A}_N$, as in the multisource criteria in (2), without the loss of generality, we maintain the manifold structure of only the first column $\boldsymbol{a}(\theta_1)$ of $\boldsymbol{A}$, which corresponds to the signal of consideration. On the other hand, the manifold structure of the remaining sources $[\boldsymbol{a}(\theta_2), \ldots, \boldsymbol{a}(\theta_N)]$, which are considered as interfering sources, is relaxed to an arbitrary matrix $\boldsymbol{B} \in \mathbb{C}^{M\times(N-1)}$ [4]. Mathematically, we assume that $\boldsymbol{A} \in \bar{\mathcal{A}}_N$ where the relaxed array manifold $\bar{\mathcal{A}}_N$ is parameterized as

$$\bar{\mathcal{A}}_N = \{\boldsymbol{A}(\vartheta) = [\boldsymbol{a}(\vartheta), \boldsymbol{B}] \,|\, \boldsymbol{a}(\vartheta) \in \mathcal{A}_1, \boldsymbol{B} \in \mathbb{C}^{M\times(N-1)}\}. \tag{4}$$

We remark that every matrix element in the relaxed array manifold $\bar{\mathcal{A}}_N$ in (4) still retains the specific structure from the geometry of the sensor array in its first column, hence the name PR. However, only one DoA can be estimated from the first column of the matrix minimizer if the cost function of (2) is minimized on the relaxed array manifold $\bar{\mathcal{A}}_N$ of (4). Therefore, we perform the spectral search similarly to the single-source approximation in the "Single-Source Approximation" section as follows. First, we fix the data matrix

$\boldsymbol{Y}$ and minimize and concentrate the objective function in (2) with respect to $\boldsymbol{B}$ and other nuisance parameters $\boldsymbol{\alpha}$ to obtain the concentrated cost function. Then, we evaluate the concentrated cost function for different values of $\boldsymbol{a}(\vartheta) \in \mathcal{A}_1$ to determine the locations of the N-deepest local minima. The concept of the PR approach is illustrated in Figure 3. Using similar notation as in the single-source approximation approach, the PR approach admits the following general optimization problem:

$$\begin{aligned}\{\boldsymbol{a}(\hat{\theta})\} &= {}^{N}\underset{\boldsymbol{A}(\theta)\in\bar{\mathcal{A}}_N}{\operatorname{argmin}}\ f(\boldsymbol{Y}\,|\,\boldsymbol{A}(\theta),\boldsymbol{\alpha}) \\ &= {}^{N}\underset{\boldsymbol{a}(\theta)\in\mathcal{A}_1}{\operatorname{argmin}}\ \min_{\boldsymbol{B}\in\mathbb{C}^{M\times(N-1)}}\ \min_{\boldsymbol{\alpha}} f(\boldsymbol{Y}\,|\,[\boldsymbol{a}(\theta),\boldsymbol{B}],\boldsymbol{\alpha}). \qquad (5)\end{aligned}$$

The rationale for the PR approach lies in the fact that, when a candidate DoA θ coincides with one of the true DoAs θ_n, then with $\boldsymbol{B}$ modeling the steering vectors of the remaining DoAs, a perfect fit to the data is attained at a high SNR or large number of snapshots T. When the candidate θ is, however, different from all true DoAs θ_n, the number of degrees of freedom in $\boldsymbol{B}$ is not sufficiently large to represent the contribution of all N-source signals. By applying different cost functions to the general optimization problem in (5), multiple novel estimators in the PR framework are obtained in [4]. A summary of estimators under the PR framework and their relations with conventional multisource and single-source approximation-based DoA estimators are provided in Table 1.

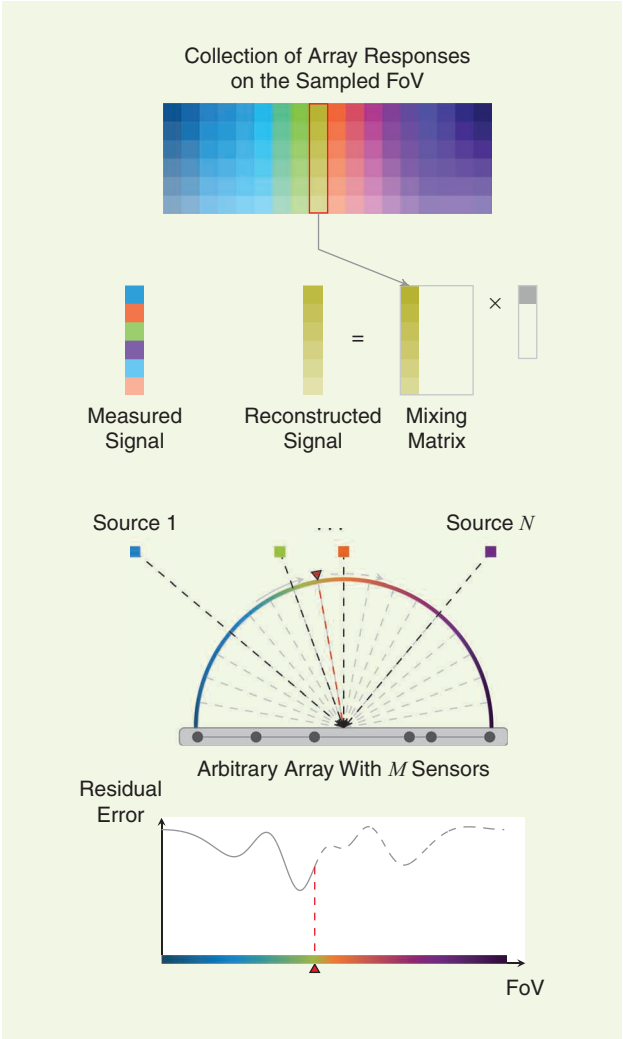

FIGURE 2. An illustration of the single-source approximation concept. The optimization is carried out only over a single-source model with an array manifold, i.e., $\boldsymbol{A}(\boldsymbol{\theta}) \rightarrow \boldsymbol{a}(\theta) \in \mathcal{A}_1$ (note that the mixing matrix has only one nonzero column corresponding to the candidate DoA $\boldsymbol{a}(\theta)$). The influence of the remaining source signals during the spectral search is neglected, which is denoted by zero columns in the mixing matrix. The data term $\boldsymbol{Y}$ in the objective function of the single-source approximation method is, however, identical to that of the corresponding multisource optimization problem.

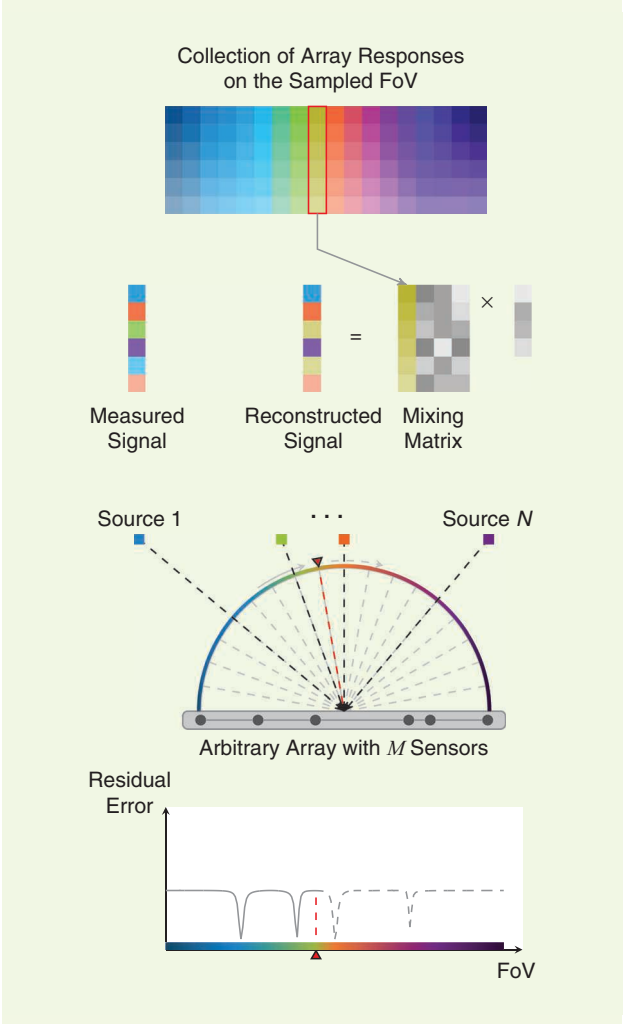

FIGURE 3. An illustration of the PR concept. The optimization is carried out over the relaxed array manifold $\bar{\mathcal{A}}_N$, where the structure of the first column in $\boldsymbol{a}(\theta_1)$ is maintained and the structure of the remaining columns is relaxed to an arbitrary complex matrix, $[\boldsymbol{a}(\theta_2),\ldots,\boldsymbol{a}(\theta_N)] \rightarrow \boldsymbol{B} \in \mathbb{C}^{M\times(N-1)}$. Unlike the single-source approximation, the influence of the remaining source signals during the spectral search is considered by the unstructured matrix $\boldsymbol{B}$ (depicted by gray columns in the mixing matrix), which generally leads to an improvement of the DoA estimation when sources are closely spaced.

Sequential techniques

While the PR methods show excellent threshold performance in scenarios with a low number of uncorrelated sources, their performance quickly degrades as the number of sources increases. This phenomenon can be explained in short as follows: the approximation error associated with the manifold relaxation of the interfering sources increases with the number of sources. The same holds true for the single-source approximation methods. As the approximation error increases, the capability of incorporating the influence of multiple structured source signals in the optimization problem decreases, and thus, a degradation in the estimation performance is observed. To overcome the degradation effect in scenarios with large source numbers, sequential estimation techniques have been proposed in which the parameters of multiple sources are estimated one after the other.

We revise three closely related and most widely known sequential estimation techniques: the MP technique, OMP, and the orthogonal least-squares (OLS) [24], [25] method. These methods have in common that the DoAs for N-sources are estimated sequentially and that the approximation is successively improved. In each iteration, the DoA of one additional source is estimated based on minimizing a function approximation of a given multisource criterion (compare Table 1), while the source DoAs estimated in the previous iterations are kept fixed at the value of their respective estimates.

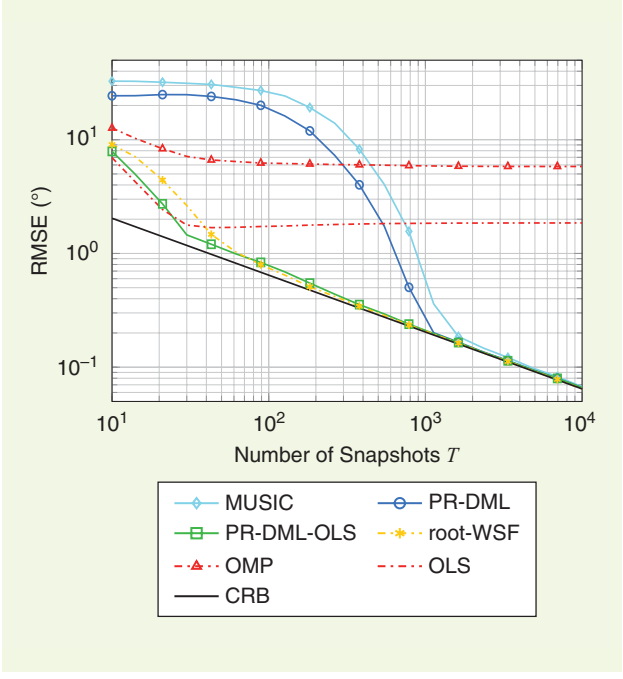

FIGURE 4. A performance evaluation of the sequential DoA estimation techniques for four uncorrelated source signals at $\boldsymbol{\theta} = [90°, 93°, 135°, 140°]^{\mathrm{T}}$ with an array composed of $M = 10$ sensors and SNR = 3 dB. OMP and OLS are biased as the first DoA is estimated according to the conventional beamformer, which cannot resolve two closely spaced sources at 90° and 93° regardless of the number of available snapshots T. On the other hand, PR-DML-OLS is asymptotically consistent, and its RMSE is close to the CRB.

Similar to the single-source approximation, the remaining sources whose DoA estimates have not yet been determined are ignored in the optimization.

The three methods differ, however, in the way the nuisance parameters corresponding to each source are treated in the optimization and in the corresponding parameter updating procedure. Concerning the MP algorithm, in each iteration, the nuisance parameters corresponding to the new source DoA are fixed and inserted as parameters in the objective in the following iterations. In contrast, in the OMP algorithm, the nuisance parameters of all estimated sources are updated in a refinement step after the DoA parameter of the current iteration is determined. The nuisance parameters are then inserted as parameters in the objective for the following estimation of the source DoA in the next iteration. The additional update step is generally associated with only a slight increase of the computational complexity. Nevertheless, this strategy effectively reduces error propagation effects.

OLS yields further performance improvements at the cost of more sophisticated estimate and update expressions. More precisely, in the OLS algorithm, the nuisance parameters corresponding to the sources of the previous and current iterations are treated as variables and optimized along with the DoA parameter of the new source in the current iteration. In Table 2, we provide the sequential estimation and update procedures of the MP, OMP, and OLS for the DML criterion (compare Table 1). At this point, we remark that the sequential estimation approach is general and can also be applied to other multisource criteria in Table 1, hence the WSF and the COMET criteria. Furthermore, sequential estimation can also be combined with the concept of PR that we introduced in the "PR Methods" section to further enhance the threshold performance and reduce the error propagation effects. As an example, we also provide the PR-DML-OLS method in Table 2. A numerical performance comparison of the sequential estimators is provided in Figure 4, where it can be observed that the OLS method shows improved performance as compared to OMP; however, both methods suffer from a bias. The PR-DML-OLS method is, in contrast, asymptotically consistent, and its root-mean-square error (RMSE) is close to the Cramér-Rao bound (CRB).

Sparse reconstruction methods

The nonlinear LS DML problem in Table 1 generally requires a multidimensional grid search over the parameter space to obtain the global minimum. More precisely, the objective function is evaluated at all possible combinations of N DoAs on a particular discretized FoV. Clearly, the complexity of this brute-force multidimensional search strategy grows exponentially with the number of sources. To reduce the computational cost associated with the nonlinear LS optimization, convex approximation methods based on sparse regularization have been proposed.

We assume that for a particular FoV discretization $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^K$ containing $K \gg M$ angles, a so-called oversampled

dictionary matrix $\tilde{\boldsymbol{A}} = \boldsymbol{A}(\tilde{\boldsymbol{\theta}})$ of dimensions $M \times K$ is constructed and the DML problem is equivalently formulated as the linear LS minimization problem with a cardinality constraint, i.e.,

$$\min_{\tilde{\boldsymbol{S}} \in \mathbb{C}^{K \times T}} \| \boldsymbol{X} - \tilde{\boldsymbol{A}} \tilde{\boldsymbol{S}} \|_{\mathrm{F}}^2 \text{ subject to } \| \tilde{\boldsymbol{S}} \|_{2,0} \le N \tag{6}$$

where $\|\boldsymbol{M}\|_{2,0} = \#(\{k \,|\, \|\boldsymbol{m}_k\|_2 \neq 0\})$ denotes the $\ell_{2,0}$ mixed pseudonorm of a matrix $\boldsymbol{M} = [\boldsymbol{m}_1, \ldots, \boldsymbol{m}_K]^{\mathsf{T}}$, i.e., the number of rows $\boldsymbol{m}_k$ with nonzero Euclidean norm $\|\boldsymbol{m}_k\|_2$ for $k = 1, \ldots, K$. This is illustrated in Figure 5. Given a solution $\tilde{\boldsymbol{S}}^{\star}$ of the optimization problem in (6), the DoAs estimates $\hat{\boldsymbol{A}}(\hat{\boldsymbol{\theta}})$ are determined from the support of $\tilde{\boldsymbol{S}}^{\star}$, i.e., the locations of the nonzero rows. Several approximation methods have been proposed to simplify the problem in (6) using sparse regularization. Sparse regularization approaches are directly devised from the Lagrangian function of the optimization problem in (6), i.e.,

$$\min_{\tilde{\boldsymbol{S}}} \| \boldsymbol{X} - \tilde{\boldsymbol{A}} \tilde{\boldsymbol{S}} \|_{\mathrm{F}}^2 + \mu \| \tilde{\boldsymbol{S}} \|_{2,0} \tag{7}$$

where the hyperparameter μ (also called the *regularization parameter*) balances the tradeoff between data matching and sparsity. For small values of μ, the mismatch between the model and the measurements is emphasized in the minimization, whereas for larger values of μ, the row sparsity of the solution is enhanced. Since the discretized FoV is given and, thus, the oversampled dictionary matrix $\tilde{\boldsymbol{A}}$ is constant, the data-matching term in the objective function of (7) is a simple linear LS function. Nevertheless, the sparse regularization term is both nonsmooth and nonconvex w.r.t. $\tilde{\boldsymbol{S}}$, and thus, the problem in (7) is difficult to solve directly. In [26], a convergent iterative fixed-point algorithm is proposed that solves a sequence of the smooth approximation problems of (7).

To make the optimization in (7) more tractable, a common approach is to convexify the regularizer in (7) by approximating the $\ell_{2,0}$ norm by the closest convex mixed-norm $\ell_{2,1}$, which is defined as $\|\boldsymbol{M}\|_{2,1} = \sum_{k=1}^{K} \|\boldsymbol{m}_k\|_2$ for $\boldsymbol{M} = [\boldsymbol{m}_1, \ldots, \boldsymbol{m}_K]^{\mathsf{T}}$. The resulting multiple measurement problem (MMP)

$$\min_{\tilde{\boldsymbol{S}}} \| \boldsymbol{X} - \tilde{\boldsymbol{A}} \tilde{\boldsymbol{S}} \|_{\mathrm{F}}^2 + \mu \| \tilde{\boldsymbol{S}} \|_{2,1} \tag{8}$$

is convex and thus can be solved efficiently [10]. One important drawback of the formulation in (8) is that the number of optimization variables grows linearly with the number of snapshots T and, therefore, also the associated computational complexity. Interestingly, the MMP can be equivalently expressed as the Sparse Row-Norm Reconstruction (SPARROW) problem as follows [11]:

$$\min_{\tilde{\boldsymbol{D}} \in \mathcal{D}_+} \operatorname{tr}\left(\left(\tilde{\boldsymbol{A}} \tilde{\boldsymbol{D}} \tilde{\boldsymbol{A}}^{\mathsf{H}} + \frac{\mu}{2\sqrt{T}} \boldsymbol{I}_M \right)^{-1} \hat{\boldsymbol{R}} \right) + \operatorname{tr}(\tilde{\boldsymbol{D}}). \tag{9}$$

We remark that in (9), the optimizing variable $\tilde{\boldsymbol{D}} = \operatorname{diag}(\tilde{d}_1, \ldots, \tilde{d}_K)$ is a nonnegative diagonal matrix whose

Table 2. Sequential DoA estimators.

	Iterative DoA Estimation Step	Nuisance Parameters Update
MP	$\hat{\theta}^{(k)} = \operatorname{argmin}_{\theta} \min_{s \in \mathbb{C}^T} \left\| \boldsymbol{X} - [\hat{\boldsymbol{A}}^{(k-1)}, \boldsymbol{a}(\theta)] \begin{bmatrix} \hat{\boldsymbol{S}}^{(k-1)} \\ \boldsymbol{s}^{\mathsf{T}} \end{bmatrix} \right\|_{\mathrm{F}}^2$ $\hat{\boldsymbol{A}}^{(k)} = [\hat{\boldsymbol{A}}^{(k-1)}, \boldsymbol{a}(\hat{\theta}^{(k)})]$	$\hat{\boldsymbol{s}}^{(k)} = \operatorname{argmin}_{\boldsymbol{s} \in \mathbb{C}^T} \left\| \boldsymbol{X} - [\hat{\boldsymbol{A}}^{(k-1)}, \boldsymbol{a}(\hat{\theta}^{(k)})] \begin{bmatrix} \hat{\boldsymbol{S}}^{(k-1)} \\ \boldsymbol{s}^{\mathsf{T}} \end{bmatrix} \right\|_{\mathrm{F}}^2$ $\hat{\boldsymbol{S}}^{(k)} = \begin{bmatrix} \hat{\boldsymbol{S}}^{(k-1)} \\ \hat{\boldsymbol{s}}^{(k)\mathsf{T}} \end{bmatrix}$
OMP	$\hat{\theta}^{(k)} = \operatorname{argmin}_{\theta} \min_{s \in \mathbb{C}^T} \left\| \boldsymbol{X} - [\hat{\boldsymbol{A}}^{(k-1)}, \boldsymbol{a}(\theta)] \begin{bmatrix} \hat{\boldsymbol{S}}^{(k-1)} \\ \boldsymbol{s}^{\mathsf{T}} \end{bmatrix} \right\|_{\mathrm{F}}^2$ $\hat{\boldsymbol{A}}^{(k)} = [\hat{\boldsymbol{A}}^{(k-1)}, \boldsymbol{a}(\hat{\theta}^{(k)})]$	$\hat{\boldsymbol{S}}^{(k)} = \operatorname{argmin}_{\boldsymbol{S} \in \mathbb{C}^{k \times T}} \| \boldsymbol{X} - [\hat{\boldsymbol{A}}^{(k-1)}, \boldsymbol{a}(\hat{\theta}^{(k)})] \boldsymbol{S} \|_{\mathrm{F}}^2$
OLS	$\hat{\theta}^{(k)} = \operatorname{argmin}_{\theta} \min_{\boldsymbol{S} \in \mathbb{C}^{k \times T}} \| \boldsymbol{X} - [\hat{\boldsymbol{A}}^{(k-1)}, \boldsymbol{a}(\theta)] \boldsymbol{S} \|_{\mathrm{F}}^2$ $\hat{\boldsymbol{A}}^{(k)} = [\hat{\boldsymbol{A}}^{(k-1)}, \boldsymbol{a}(\hat{\theta}^{(k)})]$	
PR-DML-OLS	$\hat{\theta}^{(k)} = \operatorname{argmin}_{\theta} \min_{\substack{\boldsymbol{B} \in \mathbb{C}^{M \times (N-k)} \\ \boldsymbol{S} \in \mathbb{C}^{N \times T}}} \| \boldsymbol{X} - [\hat{\boldsymbol{A}}^{(k-1)}, \boldsymbol{a}(\theta), \boldsymbol{B}] \boldsymbol{S} \|_{\mathrm{F}}^2$ $\hat{\boldsymbol{A}}^{(k)} = [\hat{\boldsymbol{A}}^{(k-1)}, \boldsymbol{a}(\hat{\theta}^{(k)})]$	

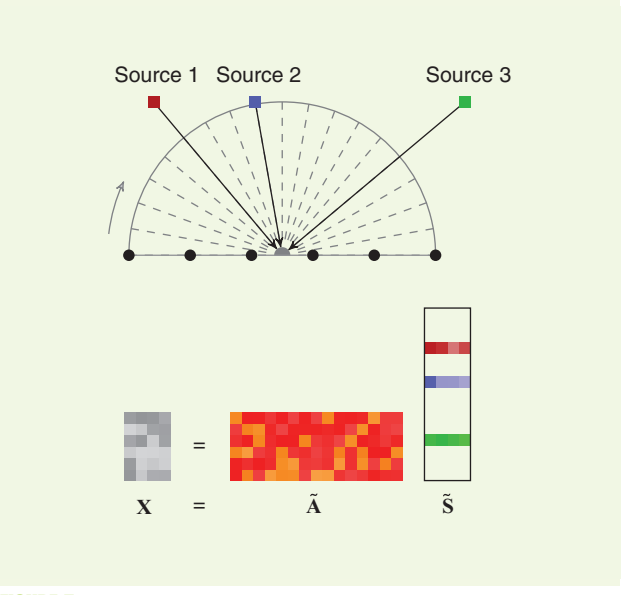

FIGURE 5. The concept of sparse reconstruction methods. In the noiseless case, the received signal $\mathbf{X}$ is decomposable into a product of a fixed oversampled steering matrix $\tilde{\mathbf{A}}$ and a row-sparse source signal matrix $\tilde{\mathbf{S}}$. The locations of the nonzero rows in $\tilde{\mathbf{S}}$ correspond to the DoAs.

dimension does not depend on the number of snapshots T. The first term in (9) can be interpreted as a data-matching term, while the second term induces sparsity. The optima $\tilde{\boldsymbol{S}}^\star$ and $\tilde{\boldsymbol{D}}^\star$ of the respective problems share the same support, and the diagonal elements of $\tilde{\boldsymbol{D}}^\star$ represent the scaled ℓ_2 row norms of $\tilde{\boldsymbol{S}}^\star$. The SPARROW problem is convex and can be efficiently solved using, e.g., a block-coordinate descent (BCD) algorithm. Remarkably, the support of the solution of (9) and therefore also the solution of the MMP (8) are fully encoded in the sample covariance matrix $\hat{\boldsymbol{R}}$, while the measurement matrix $\boldsymbol{X}$ is not explicitly required for the estimation of the DoAs. Making use of the Schur complement, the optimization problem in (9) can further be reformulated as the semidefinite problem (SDP).

Considering the array processing literature, a vast amount of estimators proposed in the past years can be derived from multisource optimization problems.

$$\min_{\tilde{\boldsymbol{D}},\boldsymbol{U}} \operatorname{Tr}(\boldsymbol{U}\hat{\boldsymbol{R}}) + \frac{1}{M}\operatorname{Tr}(\tilde{\boldsymbol{D}}) \quad \text{subject to} \begin{bmatrix} \boldsymbol{U} & \boldsymbol{I}_M^{\mathsf{H}} \\ \boldsymbol{I}_M & \tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{A}}^{\mathsf{H}} + \dfrac{\mu}{2\sqrt{T}}\boldsymbol{I}_M \end{bmatrix} \succeq 0. \tag{10}$$

An alternative SDP formulation also exists for the undersampled case, when the number of snapshots is smaller than the number of sensors, i.e., $M \geq T$. While from a computational point of view, the SDP formulations quickly become intractable when the dictionary $\tilde{\boldsymbol{A}}$ becomes large and the BCD solution is preferable for large K, the SDP formulations admit interesting extensions for ULAs and other structured arrays that do not require the use of a sampling grid and the explicit formation of the dictionary $\tilde{\boldsymbol{A}}$. The so-called gridless sparse reconstruction methods are motivated by the following observation: in the ULA case, the dictionary $\tilde{\boldsymbol{A}}$ is a Vandermonde matrix so that the matrix product $\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{A}}^{\mathsf{H}}$ with any diagonal matrix $\tilde{\boldsymbol{D}}$ can be substituted by a Toeplitz matrix $\operatorname{Toep}(\boldsymbol{u})$ with $\boldsymbol{u}$ denoting its first column. Inserting the compact Toeplitz reparameterization in the SDP problem (10) and making use of the property $\operatorname{Tr}(\tilde{\boldsymbol{D}}) = 1/M \operatorname{Tr}(\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{A}}^{\mathsf{H}}) = 1/M \operatorname{Tr}(\operatorname{Toep}(\boldsymbol{u}))$ the SDP reformulation becomes independent of a particular choice of the dictionary $\tilde{\boldsymbol{A}}$.

Consequently, the off-grid errors are avoided. An important question at this point is under which conditions the decomposition $\operatorname{Toep}(\boldsymbol{u}) = \tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{A}}^{\mathsf{H}}$ holds and whether the solution is unique. If such a decomposition exists with a unique solution, the gridless reformulation of the SDP is equivalent to the original grid-based formulation. The answer to this question is provided by the well-known Carathéodory's theorem, which states that the Vandermonde decomposition of any positive semidefinite low-rank Toeplitz matrix is always unique. Hence, provided that the solution $\operatorname{Toep}(\boldsymbol{u}^\star)$ is positive semidefinite and rank deficient, it can be uniquely factorized as $\operatorname{Toep}(\boldsymbol{u}^\star) = \boldsymbol{A}^\star \boldsymbol{D}^\star (\boldsymbol{A}^\star)^{\mathsf{H}}$ [27]. Given the generator vector $\boldsymbol{u}^\star$ retrieved from a low-rank Toeplitz matrix, the DoA estimates can be uniquely recovered, e.g., by solving the corresponding system of linear equations.

We remark that the gridless approach for sparse recovery in the MMP has first been introduced in the context of the atomic norm denoising problem [12], which can be considered as the continuous angle equivalent of the $\ell_{2,1}$ norm regularized LS matching problem (8). The associated SDP formulation in the ULA case with Toeplitz parameterization can be shown to be equivalent to the gridless version of (10).

We further remark that gridless sparse reconstruction methods are not limited to contiguous ULA structures. Also, other redundant array geometries can be exploited, such as shift-invariant arrays or thinned ULAs, i.e., incomplete ULAs with missing sensors ("holes"). In thinned ULAs, ambiguities may arise in the array manifold, and the model parameters may no longer be uniquely identifiable from the measurements. These ambiguities have, e.g., been characterized in [28], [29], and these references can provide guidelines for the choice of favorable thinned ULA geometries. Following a similar procedure as in the Toeplitz case, a substitution of the type $\boldsymbol{Q} = \tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{A}}^{\mathsf{H}}$ can be introduced where $\boldsymbol{Q}$ is no longer perfectly Toeplitz but contains other structured redundancies that can be expressed in the form of linear equality constraints in the problem (10). In these cases, estimation of signal parameters via rotational invariance techniques (ESPRIT) or root-MUSIC can be employed to estimate the DoAs from the minimizer $\boldsymbol{Q}^\star$. Even though unique factorization guarantees for $\boldsymbol{Q}^\star$ similar to Carathéodory's theorem do not exist, the generalized gridless recovery approach performs well in practice as long as the number of redundant entries in $\boldsymbol{Q}$ is sufficiently large.

While we focused in our overview on sparse regularization methods that are based on the DML cost function in Table 1 as the data-matching term, there exist numerous alternative approaches that use other matching terms. See [9] for a comprehensive overview of sparse DoA estimation techniques. A particularly interesting sparse DoA estimation method is the Sparse Iterative Covariance-Based Estimation Approach (SPICE) [8], which, as the name suggests, stems from a weighted version of the covariance matching criterion in Table 1. Remarkably, the SPICE formulation does not contain any hyperparameters to trade off between the data-matching quality versus the sparsity of the solution, which makes SPICE an attractive candidate among sparse reconstruction methods.

As mentioned previously, the traditional superresolution methods, such as the multisource estimation methods of Table 1 as well as the PR methods and MUSIC for uncorrelated sources, are capable of resolving arbitrary closely spaced source even with a finite number of sensors as long as the number of snapshots or the SNR is sufficiently large. It is important to note that such guarantees generally do not exist in convex sparse optimization methods [9], [12], [27]. Furthermore, sparse regularization-based DoA estimation

methods are known to suffer from bias, which marks one of their most important drawbacks. First, the bias can be originated from the grid mismatch of the source DoAs in the formation of the dictionary. Second, the sparse regularization term generally introduces a bias to the solution. While the former source of bias can be entirely avoided in the gridless sparse reconstruction formulations, the latter remains and can be reduced only by decreasing the regularization parameter μ, e.g., in (7). This in turn leads to an enlarged support set whose sizes are much larger the true number of sources N.

However, in the context of sparse regularization-based DoA estimation, if the model order N is known, it is often preferable to use comparably small values of the regularizers μ and to perform a local search for the N-largest maxima of the recovered row-norm vector $\tilde{\boldsymbol{d}}^{\star} = [\tilde{d}_1^{\star}, \ldots, \tilde{d}_K^{\star}]^{\mathrm{T}}$ in (9) to determine the DoA estimates. More specifically, the DoA estimates are the N entries in the sample DoA vector $\tilde{\boldsymbol{\theta}}$ that are indexed by $\{i\} = {}^{N}\mathrm{argmax}_k\, \tilde{d}_k^{\star}$.

Similar to the conventional parametric methods, the Partial Relaxation approach considers the signals from all potential source directions in the multisource cost function.

In conclusion, sparsity-based methods have their merit in difficult scenarios with low sample size or highly correlated and even coherent source signals where the sample covariance matrix does not exhibit the full signal rank N. In these scenarios, conventional subspace-based DoA estimation methods usually fail to resolve multiple sources. This is confirmed in the simulation example of Figure 6. Furthermore, sparsity-based methods can resolve multiple sources even in the single-snapshot case provided that the scene is sparse in the sense that the number of sources is small compared to the number of sensors in the array. A simple but interesting theorem for robust sparse estimation with noisy measurements regardless of the chosen sparse estimation approach is given by [30, Theorem 5].

Another benefit of sparse regularization methods is that, unlike parametric methods in DoA estimation, the knowledge of the number of sources N is not required for the estimation of the DoAs. In turn, the number of sources is implicitly determined from the sparsity of the solution. However, sparse reconstruction methods, with the exception of the hyperparameter-free SPICE method [8], are usually sensitive to a proper choice of the sparse regularization parameter μ. Furthermore, the associated computational complexity of sparse regularization methods, in particular for the SDP formulations in the grid-based and gridless cases, is higher than that of conventional subspace-based methods.

Exploitation of incomplete structural information

In many modern applications, such as the networks of aerial base stations, DoA estimation is carried out in a distributed fashion from signals measured at multiple subarrays, where the exact locations of subarrays are often unknown. Even in conventional centralized large sensor arrays, it is a challenging task to have a central synchronized clock among all sensors of the device and to maintain precise phase synchronization due to large distances in the array. Therefore, in practice, large array systems are partitioned into local subarrays, where due to the proximity, each subarray can be considered as perfectly calibrated, whereas the relative phase differences between subarrays are considered as unknown. In this setup, it is commonly assumed that the narrow-band assumption remains valid; hence, the waveforms do not essentially decorrelate during the travel time over the array.

DoA estimation in partly calibrated sensor arrays has first been considered in shift-invariant sensor array systems, which are composed of two identically oriented identical subarrays separated by a known displacement δ. For this configuration, the popular ESPRIT algorithm has been proposed [31]. In shift-invariant arrays, the overall array steering matrix $\boldsymbol{A}(\boldsymbol{\theta})$ can be partitioned into two potentially overlapping blocks, $\underline{\boldsymbol{A}}(\boldsymbol{\theta}) \in \mathbb{C}^{M_1 \times N}$ and $\overline{\boldsymbol{A}}(\boldsymbol{\theta}) \in \mathbb{C}^{M_2 \times N}$, respectively, representing the array response of the reference subarray and the shifted subarray. For notational simplicity, we assume that $M = 2M_1 = 2M_2$. Due to the shifting structure, the two subarray steering matrices are related through right multiplication with a diagonal phase shift matrix $\boldsymbol{D}(\boldsymbol{z}^{\delta}) = \mathrm{diag}(z_1^{\delta}, \ldots, z_N^{\delta})$ with unit-magnitude generators $z_n = e^{-\mathrm{j}\pi\cos(\theta_n)}$ that account for the known displacement δ measured in half wavelength, hence $\overline{\boldsymbol{A}}(\boldsymbol{\theta}) = \underline{\boldsymbol{A}}(\boldsymbol{\theta})\boldsymbol{D}(\boldsymbol{z})$.

Interestingly, ESPRIT as well as the enhanced Total-LS-ESPRIT (TLS-ESPRIT) method [31] can be reformulated as subspace-fitting techniques according to Table 1. Similar to the PR approach, a particular form of manifold relaxation is applied that maintains some part of the array structure and deliberately ignores other parts of the structure to admit a simple solution. The ESPRIT and TLS-ESPRIT estimators

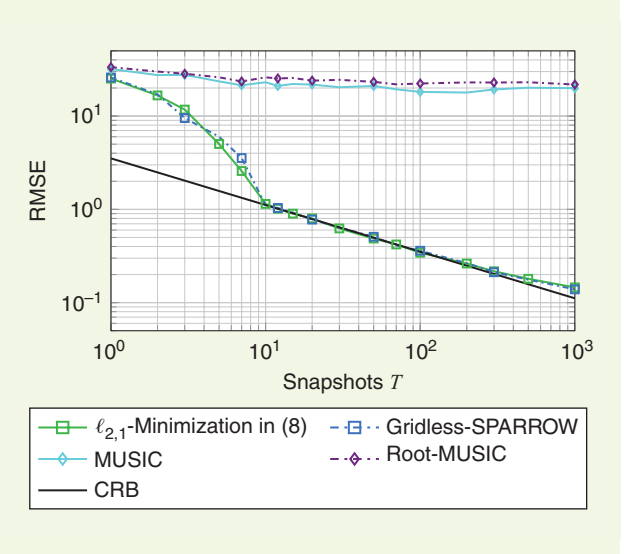

FIGURE 6. A performance evaluation of the sparse reconstruction-based DoA estimation techniques for two coherent sources at $\boldsymbol{\theta} = [90°, 120°]^{\mathrm{T}}$ with an array composed of $M = 6$ sensors and SNR = 10 dB.

are obtained from minimizing the subspace fitting functions $\|\hat{\boldsymbol{U}}_{\mathrm{s}}\boldsymbol{V}^{-1}-\boldsymbol{A}(\boldsymbol{\theta})\|_{\mathrm{F}}^2$ [22] and $\|\hat{\boldsymbol{U}}_{\mathrm{s}}-\boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{V}\|_{\mathrm{F}}^2$, respectively, for nonsingular $\boldsymbol{V}$, where the array manifold $\mathcal{A}_N$ of the fully calibrated subarray defined in (1) is relaxed to the ESPRIT manifold $\mathcal{A}_N^{\mathrm{ESPRIT}}=\{\boldsymbol{A}\in\mathbb{C}^{M\times N}|\,\boldsymbol{A}(\boldsymbol{\vartheta})=[\underline{\boldsymbol{A}}^{\mathsf{T}},\boldsymbol{D}(\boldsymbol{c})\underline{\boldsymbol{A}}^{\mathsf{T}}]^{\mathsf{T}},\underline{\boldsymbol{A}}\in\mathbb{C}^{M_1\times N},\boldsymbol{c}\in\mathbb{C}^N,\boldsymbol{\vartheta}=\cos^{-1}(-(\arg(\boldsymbol{c})/\pi\delta))\}$.

We remark that in the ESPRIT manifold $\mathcal{A}_N^{\mathrm{ESPRIT}}$, only part of the shift-invariance structure of $\mathcal{A}_N$ is maintained, and the particular subarray steering matrix structure in $\underline{\boldsymbol{A}}(\boldsymbol{\vartheta})=[\underline{\boldsymbol{a}}(\vartheta_1),\ldots,\underline{\boldsymbol{a}}(\vartheta_N)]$ is relaxed to an arbitrary complex matrix $\underline{\boldsymbol{A}}\in\mathbb{C}^{M_1\times N}$. In addition, the magnitude-one structure of the diagonal shift matrix $\boldsymbol{D}(\boldsymbol{z}^{\delta})$ is relaxed to an arbitrary diagonal matrix $\boldsymbol{D}(\boldsymbol{c})$. This implies that in the ESPRIT and TLS-ESPRIT algorithms, neither the subarray geometry nor potential directional gain factors between the two subarrays need to be known as long as the subarrays are identical and subarray displacements are known. Due to this particular manifold relaxation, the subspace fitting problems admit efficient closed-form solutions.

The concept of DoA estimation in subarray structures has been generalized in [32] to cover the case of multiple shift invariance arrays. In more general partly calibrated array scenarios, we assume that the sensor positions are generally unknown. Nevertheless, only several displacements between selected pairs of sensors in the array, the so-called lags of the array, are known. Let $\delta_1,\ldots,\delta_K$ denote known displacements in half wavelength in the array that are all pointing into the same direction. This is illustrated in Figure 7. The subarrays are flexibly defined by pairs of sensors that share a common lag δ_k (or their summations). Depending on the number of known lags among the sensor arrays, one particular sensor can belong to one or more subarrays.

For all known lags, we consider again the subspace fitting approach and apply the ESPRIT manifold relaxation technique. Hence, defining $\boldsymbol{T}=\boldsymbol{V}^{-1}$ and relaxing the structure of the subarrays, the objective becomes $\sum_{k=1}^{K}\left\|\left[\hat{\underline{\boldsymbol{U}}}_{\mathrm{s},k}\boldsymbol{T},\hat{\overline{\boldsymbol{U}}}_{\mathrm{s},k}\boldsymbol{T}\right]-\underline{\boldsymbol{A}}_k[\boldsymbol{I},\boldsymbol{D}(\boldsymbol{z}^{\delta_k})]\right\|_{\mathrm{F}}^2$, where $\underline{\boldsymbol{A}}_k$ is an arbitrary complex-valued matrix of known dimension that models the unknown subarray structure corresponding to the kth displacement δ_k and $\hat{\underline{\boldsymbol{U}}}_{\mathrm{s},k}$. The matrix $\hat{\overline{\boldsymbol{U}}}_{\mathrm{s},k}$ contains the corresponding rows of the signal eigenvectors in $\hat{\boldsymbol{U}}_{\mathrm{s},k}$. Inserting the LS minimizers $\hat{\underline{\boldsymbol{A}}}_{\mathrm{LS},k}=(1/2)\left(\hat{\underline{\boldsymbol{U}}}_{\mathrm{s},k}\boldsymbol{T}+\hat{\overline{\boldsymbol{U}}}_{\mathrm{s},k}\boldsymbol{T}\boldsymbol{D}^{*}(\boldsymbol{z}^{\delta_k})\right)$ back into the objective function, the concentrated objective function of the relaxed multiple shift-invariant ESPRIT is given by

$$\sum_{k=1}^{K}\mathrm{Tr}\Big(-\boldsymbol{T}^{\mathsf{H}}\hat{\underline{\boldsymbol{U}}}_{\mathrm{s},k}^{\mathsf{H}}\hat{\overline{\boldsymbol{U}}}_{\mathrm{s},k}\boldsymbol{T}\boldsymbol{D}(\boldsymbol{z}^{-\delta_k})+\boldsymbol{T}^{\mathsf{H}}\Big(\hat{\underline{\boldsymbol{U}}}_{\mathrm{s},k}^{\mathsf{H}}\hat{\underline{\boldsymbol{U}}}_{\mathrm{s},k}+\hat{\overline{\boldsymbol{U}}}_{\mathrm{s},k}^{\mathsf{H}}\hat{\overline{\boldsymbol{U}}}_{\mathrm{s},k}\Big)\boldsymbol{T}-\boldsymbol{T}^{\mathsf{H}}\hat{\overline{\boldsymbol{U}}}_{\mathrm{s},k}^{\mathsf{H}}\hat{\underline{\boldsymbol{U}}}_{\mathrm{s},k}\boldsymbol{T}\boldsymbol{D}(\boldsymbol{z}^{\delta_k})\Big).\tag{11}$$

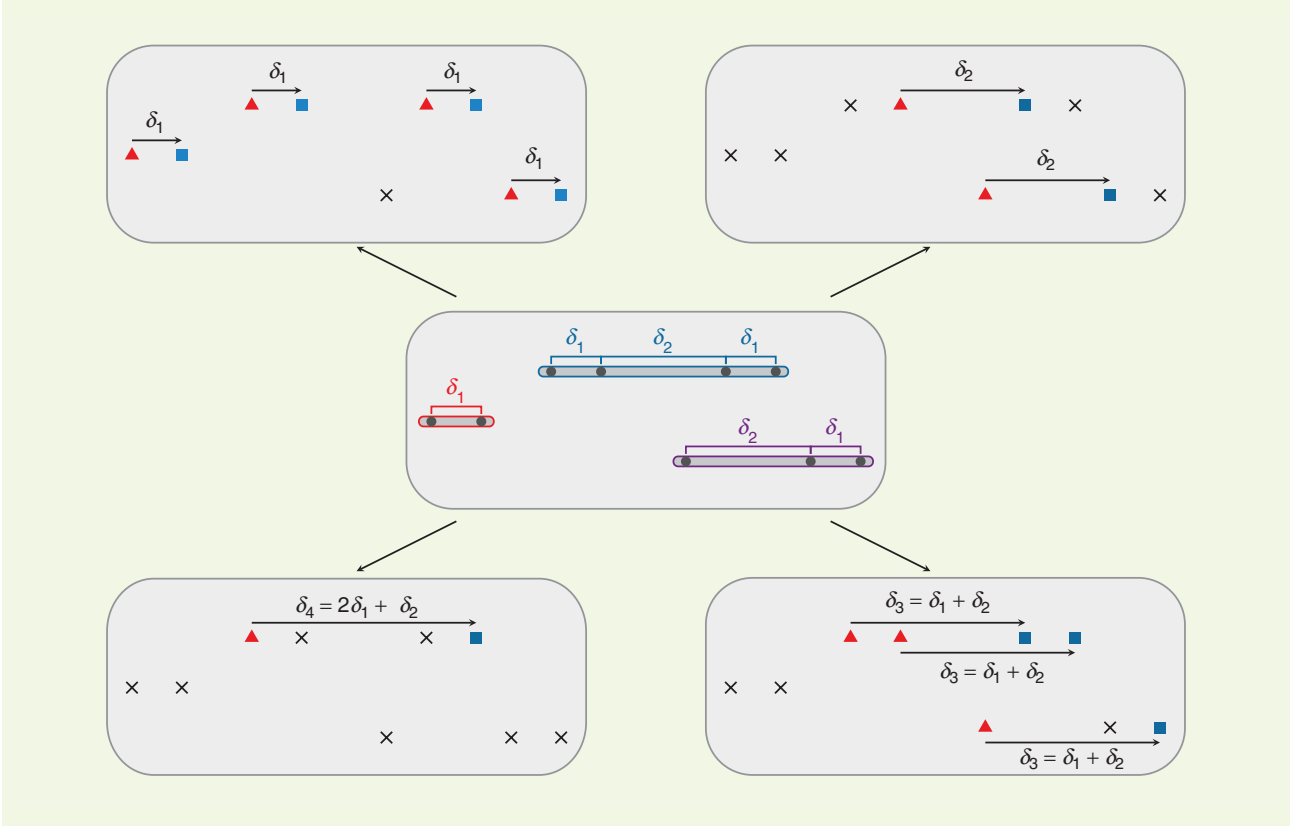

FIGURE 7. The equivalence between the partly calibrated array setup and multiple shift-invariant setup. As depicted in the center, an exemplary partly calibrated array setup comprises three linear subarrays with unknown intersubarray displacements. The displacements between sensors in one subarray are, however, a priori known. From the known intrasubarray displacement, multiple shift-invariant structures between sensor pairs are exploited while formulating the optimization problem. Such exploitation allows reinterpretation of the RARE algorithm [15], [16] as a generalized multiple shift ESPRIT [31].

Due to the diagonal structure of $\boldsymbol{D}(\boldsymbol{z}^{\delta_k})$, the objective function in (11) is separable into N identical terms, one for each source. Hence the subspace fitting problem reduces to finding the N distinct minima of the rank reduction estimator (RARE) [15], [16] function $f_{\mathrm{RARE}}(\theta) = \min_{\|\boldsymbol{t}\|=1} \boldsymbol{t}^{\mathsf{H}} \boldsymbol{M}(\theta) \boldsymbol{t}$ with respect to the DoAs $\theta \in \Theta$ where

$$\boldsymbol{M}(\theta) = \sum_{k=1}^{K} \Big(-\underline{\hat{\boldsymbol{U}}}_{\mathrm{s},k}^{\mathsf{H}} \overline{\hat{\boldsymbol{U}}}_{\mathrm{s},k} e^{\mathrm{j}\pi\delta_k \cos(\theta_n)} + \Big(\underline{\hat{\boldsymbol{U}}}_{\mathrm{s},k}^{\mathsf{H}} \underline{\hat{\boldsymbol{U}}}_{\mathrm{s},k} + \overline{\hat{\boldsymbol{U}}}_{\mathrm{s},k}^{\mathsf{H}} \overline{\hat{\boldsymbol{U}}}_{\mathrm{s},k} \Big) - \overline{\hat{\boldsymbol{U}}}_{\mathrm{s},k}^{\mathsf{H}} \underline{\hat{\boldsymbol{U}}}_{\mathrm{s},k} e^{-\mathrm{j}\pi\delta_k \cos(\theta_n)} \Big) \tag{12}$$

and $\boldsymbol{t}$ represents a particular column of $\boldsymbol{T}$. The unit-norm constraint in the RARE problem is introduced to ensure that the zero solution $\boldsymbol{t} = \boldsymbol{0}$ is excluded since the zero solution $\boldsymbol{t} = \boldsymbol{0}$ violates the constraint that $\boldsymbol{T}$ and $\boldsymbol{V}$ are nonsingular. The minimization of the RARE cost function w.r.t. to the vector $\boldsymbol{t}$ admits the minor eigenvector of $\boldsymbol{M}(\theta)$ as a minimizer. As a result, the concentrated RARE cost function is given by $f_{\mathrm{RARE}}(\theta) = \lambda_{\min}(\boldsymbol{M}(\theta))$. Thus, similar to the single-source approximation approach, the DoAs are determined from the N-deepest minima of the RARE function. This ensures that the corresponding transformation matrices $\boldsymbol{T}$ and $\boldsymbol{V}$ are nonsingular. We remark that the RARE estimator has originally been derived from a relaxation of the MUSIC function, and the minimum eigenvalue function can equivalently be replaced by the determinant of the matrix $\boldsymbol{M}(\theta)$. The latter is, e.g., useful for developing a search-free variant of the spectral RARE algorithm based on matrix polynomial rooting in the case that the shifts $\delta_1, \ldots, \delta_K$ are integer multiples of a common baseline.

Exploitation of array configuration

As mentioned in the "Signal Model" section, the number of signals N that can be uniquely recovered from DoA estimation methods with second-order statistics is strictly upper bounded by the Kruskal rank of the oversampled ($K \geq M$) steering matrix $\tilde{\boldsymbol{A}} \in \mathcal{A}_K$, which, e.g., for ULA geometries is equal to the number of sensors. A direct consequence is that, using the conventional signal model in the "Signal Model" section, the number of uniquely identifiable sources N must be less than the number of sensors M. If further information on the source signals is available, e.g., that the source signals are uncorrelated, then the number of uniquely identifiable source signals can be improved.

This claim can be explained by comparing the number of equations and the number of unknowns, which are implied from the covariance model $\boldsymbol{R} = \boldsymbol{A}(\boldsymbol{\theta}) \boldsymbol{D}(\boldsymbol{p}) \boldsymbol{A}^{\mathsf{H}}(\boldsymbol{\theta}) + \nu \boldsymbol{I}_M$, with $\boldsymbol{D}(\boldsymbol{p}) = \mathrm{diag}(p_1, \ldots, p_N)$. We assume that the number of snapshots T is sufficiently high such that the covariance matrix $\boldsymbol{R}$ can be estimated with high accuracy. In addition, we remark that the structure of the covariance matrix depends on the geometry of the sensor array. For example, for a ULA, the covariance matrix is a Hermitian Toeplitz matrix. Conversely, if we assume that the sensor array does not exhibit any particular geometry, e.g., no ULA structure, then there is generally no relation between the elements in the covariance matrix. Consequently, the covariance matrix $\boldsymbol{R}$ is parameterized by maximally $M^2 - M + 1$. independent real-valued variables (note that the diagonal entries are all identical). Thus, the number of independent equations from the covariance model is also $M^2 - M + 1$.

On the other hand, in the uncorrelated source case, the model on the right-hand side $\boldsymbol{A}(\boldsymbol{\theta}) \boldsymbol{D}(\boldsymbol{p}) \boldsymbol{A}^{\mathsf{H}}(\boldsymbol{\theta}) + \nu \boldsymbol{I}_M$ contains only $2N + 1$ unknowns (N DoA parameters in vector $\boldsymbol{\theta}$, N-source powers in $\boldsymbol{p} = [p_1, \ldots, p_N]^{\mathsf{T}}$, and the noise variance ν). This observation suggests that it is possible to significantly increase the number of uniquely identifiable sources in an array from $\mathcal{O}(M)$ to $\mathcal{O}(M^2)$ if the number of redundant entries in the covariance matrix is reduced. Therefore, from the viewpoint of improving the number of detectable sources for a fixed number of sensors, we should deviate from the conventional ULA array structure. The reason is that the covariance matrix in the case of a ULA and uncorrelated source signals is a Hermitian Toeplitz matrix, which contains only $(2M - 1)$ real-valued independent entries.

In fact, the covariance matching approach in Table 1 combined with the concepts of sparse reconstruction in DoA estimation and positive definite Toeplitz matrix low-rank factorization has inspired an interesting line of research on nested and coprime arrays that aims at designing favorable nonredundant spatial sampling patterns [19]. These types of arrays include the class of minimum redundancy and augmentable arrays whose design approaches rely on thinned ULA geometries. One example is the sparse nonuniform arrays with intersensor spacings being integer multiples of a common baseline δ. These geometries have the benefit over arrays with arbitrary noninteger spacings that they allow the use of search-free DoA estimation methods (compare the gridless sparse methods introduced in the previous section) and spatial smoothing techniques to build subspace estimates of the required rank. More precisely, given the stochastic signal model in the uncorrelated source case with source powers $\boldsymbol{p}$, hence, $\boldsymbol{R} = \boldsymbol{A}(\boldsymbol{\theta}) \boldsymbol{D}(\boldsymbol{p}) \boldsymbol{A}^{\mathsf{H}}(\boldsymbol{\theta}) + \nu \boldsymbol{I}_M$, an equivalent single-snapshot model is obtained from vectorization.

Defining $\boldsymbol{C}(\boldsymbol{\theta}) = \boldsymbol{A}^{*}(\boldsymbol{\theta}) \odot \boldsymbol{A}(\boldsymbol{\theta})$ as the steering matrix of a so-called virtual difference coarray, where $\odot$ stands for the Khatri-Rao product, i.e., column-wise Kronecker product, the vectorized covariance model reads $\boldsymbol{r} = \mathrm{vec}(\boldsymbol{R}) = \boldsymbol{C}(\boldsymbol{\theta}) \boldsymbol{p} + \nu \, \mathrm{vec}(\boldsymbol{I}_M)$ [19]. For this model, the $\ell_{2,1}$-norm regularized LS approach in (8) is a suitable candidate for DoA estimation. An interesting alternative approach that does not rely on sparsity but on the nonnegativity property of the source power vector $\boldsymbol{p}$ is proposed in [33].

In the vectorized covariance model, the number of identifiable sources is fundamentally limited by the Kruskal rank of the difference coarray steering matrix $\boldsymbol{C}(\boldsymbol{\theta})$. Hence, the design objective for the physical array is to place the sensors such that, in the difference coarray, redundant rows of the difference coarray steering matrix $\boldsymbol{C}(\boldsymbol{\theta})$ are avoided and the number of contiguous lags is maximized. Avoiding redundant rows is equivalent to maximizing the diversity of the coarray, i.e., the number of different lags in the coarray. Maximizing the number of contiguous lags in the coarray corresponds to

maximizing the size of the largest "hole-free" ULA partition of the coarray, which in turn is directly related to the Kruskal rank of the ULA partition (and therefore also the Kruskal rank of the entire difference coarray) due to the Vandermonde property of the ULA steering matrix.

Because of the Khatri-Rao structure of the difference coarray steering matrix $\boldsymbol{C}(\boldsymbol{\theta})$, the Kruskal rank and, thus, the number of uniquely identifiable sources, grows quadratically rather than linearly with the number of physical sensors M. While coarray designs allow one to significantly increase the number of detectable sources for a given number of physical sensors using standard DoA estimation algorithms, recent theoretical performance results reveal that, in the regime where the number of sources N exceeds the number of sensors M, the mean-square estimation error of the MUSIC algorithm applied to the coarray data does not vanish asymptotically with SNR [34].

In the minimum redundancy array design, the number of contiguous lags in the difference coarray, and hence the size of the largest ULA partition, is maximized by definition, which generally requires a computationally extensive combinatorial search over all possible spatial sampling patterns. The nested and coprime array designs, in contrast, represent systematic design approaches associated with computationally efficient analytic array design procedures [19]. Nested array and coprime arrays are composed of two uniform linear subarrays with different baselines. In the nested array structure, each subarray is composed of M_1 and M_2 sensors with baselines δ and $(M_1+1)\delta$, respectively, where the first sensor of the first subarray lies at the origin and the first sensor of the second subarray is displaced by M_1. It can be shown that, with an equal split ($M_1 = M_2$ and $M_1 = M_2 + 1$ for even and odd M, respectively), the difference coarray becomes a ULA with $2M_2(M_1+1)-1$ elements.

Coprime arrays represent more general array structures and comprise, as the name suggests, two uniform linear subarrays with M_1 and $M_2 - 1$ sensors, respectively, with M_1 and M_2 being coprime numbers. The first and the second subarray have baselines $L_1\delta$ and $L_2\delta$, respectively, where $L_1 = M_2$ and $L_2 = M_1/F$ are coprime numbers and integer F is a given array compression factor in the range $1 \le F \le M_1$. Furthermore, the subarrays are displaced by integer multiples of the baseline δ.

In Figure 8, the nested and coprime array structures and their respective virtual coarrays are illustrated for the case of three sensors in each subarray. While the nested array structure yields a coarray with a maximum number of contiguous lags, the coprime array structure may often be preferable in practice as it can achieve not only a larger number of unique lags, i.e., degrees of freedom of up to $(M/2)^2 + M/2$, but also a larger virtual coarray aperture as well as a larger minimum interelement spacing of the physical array to reduce, e.g., mutual coupling effects.

To further increase the estimation performance of DoA estimators in a coprime array structure, low-rank Toeplitz and Hankel matrix completion approaches have been proposed to fill the "holes" and augment the data in sparse virtual coarrays to the corresponding full virtual ULA [35]. This concept has been successfully applied in [36] in the context of bistatic automotive radar to improve the angular resolution without increasing the hardware costs. Similarly, in [37], matrix completion for data interpolation in coprime virtual arrays has been used for subspace estimation in hybrid analog and digital precoding with a reduced number of analog-to-digital converters and radio frequency chains in the hardware receives. Conditions under which, in the noise-free case, the completion from a single temporal snapshot is exact have been derived in [38].

Sparsity-based methods have their merit in difficult scenarios with low sample size or highly correlated and even coherent source signals.

Conclusions and future directions

In this review article, we revisit important developments in area sensor array processing for DoA estimation in the past three decades from a modern optimization and structure exploitation perspective. From several illustrative examples, we show how novel concepts and algorithms that have advanced the research field in the last decades are proposed to solve, in some way or the other, the same notoriously challenging multisource optimization problems, such as the well-known classical DML problem. Advances in convex optimization research and the development of efficient interior point solvers for semidefinite programs made it possible to compute close-to-optimal approximate solutions to these problems with significantly reduced effort.

In addition, we also show how particular structure in the measurement model has been efficiently exploited to make the problems computationally tractable, both in terms of an affordable computational complexity as well as in terms of well posedness of the problem for identifying the parameters of interest. Nevertheless, we remark that our coverage of the sensor array processing

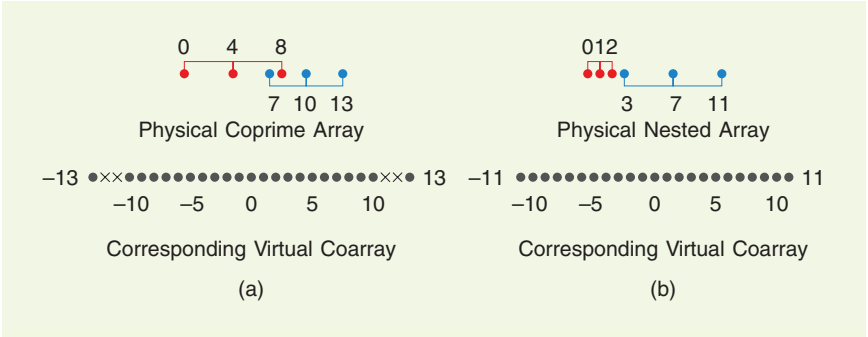

FIGURE 8. Examples of the coprime and nested array structure consisting of two subarrays, each with three sensors. The baseline of each physical subarray is a multiple of the baseline of the virtual coarray.

research of the past three decades is by no means meant to be exhaustive. Given the long history of array signal processing, by now, this field of research can certainly be considered mature. Despite all the progress that has been made over the past decades, many important and fundamental research problems in this area have not yet been solved completely and require new ideas and concepts, and some of these are outlined next.

Harmonic retrieval in large dimensional datasets

One example is the extension of the parameter estimation in 1D spaces, such as in conventional DoA estimation, to higher dimensional spaces, e.g., as required in the aforementioned parametric MIMO channel estimation problem. With the trend to massive sensing systems and high dimensional datasets, the harmonic retrieval problem in extremely large dimensions gains significant interest. Due to the phenomenon known as the *curse of dimensionality*, where the computation workload increases exponentially with the number of dimensions, the extension of 1D DoA estimation methods to higher dimensions is not straightforward. Existing works on multidimensional harmonic retrieval either consider rather low dimensions or rely on dimensionality reduction approaches, i.e., projecting the multidimensional datasets into lower dimensions. This is, however, associated with a significant performance degradation if sources are not well separated in the projected domain.

In modern applications, the received signals as well as the waveforms often exhibit additional properties that can be exploited while designing novel DoA estimators.

Incorporation of signal properties as prior information

The use of particular structures in the array manifold, which is considered in this review article, is only one form of incorporating additional prior information into the estimation problem. As more information is exploited, the parameter estimation task can be correspondingly simplified, and the estimation performance is enhanced. Theoretical investigations on the general use of additional side information incorporated in the DoA estimation problem as well as its estimation performance bound are addressed in [39]. In modern applications, the received signals as well as the waveforms often exhibit additional properties that can be exploited while designing novel DoA estimators. For example, constant modulus properties of the transmitted signals or signal waveforms with temporal dependence as, e.g., in radar chirp signals, enable coherent processing across multiple snapshots and dramatically enhance the resolution capabilities. Another example of signal exploitation is the DoA estimation with quantized or one-bit measurements, which has been studied in [40].

Robust sensor array processing

In many real-world applications, the classical array signal model may be oversimplistic. This can lead to a severe performance degradation of conventional high-resolution DoA estimation methods, which are known to be very sensitive to even small model mismatches. In recent years, significant efforts have been made to design DoA estimation methods that are robust to various model mismatches, including array imperfections due to miscalibration, impairments of the receiver front ends, mutual coupling between antennas, waveform decorrelation across the sensor array in inhomogeneous media, and multipath environments as well as impulsive and heavy-tailed noise [41].

Combining model-based with data-driven DoA estimation

Recently, data-driven machine learning approaches have been successfully introduced in many areas of signal processing to overcome the existing limitations of traditional model-based approaches. Data-driven algorithms have the benefit that they naturally generalize to various statistics of the training data and, thus, are flexible to adapt to time-varying estimation scenarios. As such, data-driven algorithms are potential candidates to overcome the aforementioned challenges in DoA estimation. However, typical off-the-shelf data-driven algorithms are known to be data hungry, which limits their practical use in many DoA estimation applications. Recently, hybrid model-and-data-driven methods were proposed in the context of deep algorithm unfolding, which combine the benefits of both approaches. The hybrid algorithms inherit the structure of existing model-based algorithms in their learning architecture to reduce the number of learning parameters and therefore speed up the learning and improve the generalization capability of the algorithms [42].

Acknowlegment

The work of Marius Pesavento was supported by the BMBF project Open6GHub under Grant 16KISK014 and the DFG PRIDE Project PE 2080/2-1 under Project No. 423747006.

Authors

Marius Pesavento (pesavento@nt.tu-darmstadt.de) received his Dr.-Ing. degree in electrical engineering from Ruhr-University Bochum, Germany, in 2005. He is a professor at the Department of Electrical Engineering and Information Technology, TU Darmstadt, 64289 Darmstadt, Germany. He is the recipient of the 2005 Young Author Best Paper Award of *IEEE Transactions on Signal Processing* and chair of the technical area committee Signal Processing for Multisensor Systems of the EURASIP. His research interests are in statistical and array signal processing, multiuser communications, optimization, and learning.

Minh Trinh-Hoang (thminh@nt.tu-darmstadt.de) received his Dr. -Ing. in electrical engineering from TU Darmstadt, Germany in 2020. From 2020 to 2021 he was a postdoctoral researcher at the Department of Electrical Engineering and Information Technology at TU Darmstadt. He is currently a research and development engineer at Rohde & Schwarz, 81614 Munich, Germany. He is a recipient of the Best PhD

Thesis Award of the Association of Friends of TU Darmstadt. His research interests include statistical signal processing, array processing, parallel hardware architecture, numerical computation, and large-scale optimization.

Mats Viberg (mats.viberg@bth.se) received his Ph.D. degree from Linköping University, Sweden. During 1993–2018, he was a professor of signal processing at the Chalmers University of Technology, Sweden, and since 2018, he has served as the vice-chancellor at the Blekinge Institute of Technology, 371 79 Karlskrona, Sweden. He has served in various capacities in the IEEE Signal Processing Society, including chair of two technical committees, and he has received two paper awards. His research interests are in statistical signal processing and its various applications. He is a Fellow of IEEE and a member of the Royal Swedish Academy of Sciences (KVA).

References

[1] F. Gao, Z. Tian, E. G. Larsson, M. Pesavento, and S. Jin, "Introduction to the special issue on array signal processing for angular models in massive MIMO communications," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 5, pp. 882–885, Sep. 2019, doi: 10.1109/JSTSP.2019.2938880.

[2] H. L. Van Trees, *Optimum Array Processing*. New York, NY, USA: Wiley, 2002.

[3] H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach," *IEEE Signal Process. Mag.*, vol. 13, no. 4, pp. 67–94, Jul. 1996, doi: 10.1109/79.526899.

[4] M. Trinh-Hoang, M. Viberg, and M. Pesavento, "Partial relaxation approach: An eigenvalue-based DOA estimator framework," *IEEE Trans. Signal Process.*, vol. 66, no. 23, pp. 6190–6203, Dec. 2018, doi: 10.1109/TSP.2018.2875853.

[5] J.-J. Fuchs, "Detection and estimation of superimposed signals," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 1998, vol. 3, pp. 1649–1652, doi: 10.1109/ICASSP.1998.681771.

[6] J.-J. Fuchs, "On sparse representations in arbitrary redundant bases," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 1341–1344, Jun. 2004, doi: 10.1109/TIT.2004.828141.

[7] J. H. Ender, "On compressive sensing applied to radar," *Signal Process.*, vol. 90, no. 5, pp. 1402–1414, May 2010, doi: 10.1016/j.sigpro.2009.11.009.

[8] P. Stoica, P. Babu, and J. Li, "SPICE: A sparse covariance-based estimation method for array processing," *IEEE Trans. Signal Process.*, vol. 59, no. 2, pp. 629–638, Mar. 2011, doi: 10.1109/TSP.2010.2090525.

[9] Z. Yang, J. Li, P. Stoica, and L. Xie, "Chapter 11 - Sparse methods for direction-of-arrival estimation," in *Academic Press Library in Signal Processing*, vol. 7, R. Chellappa and S. Theodoridis, Eds. New York, NY, USA: Academic, 2018, pp. 509–581.

[10] D. Malioutov, M. Cetin, and A. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 3010–3022, Aug. 2005, doi: 10.1109/TSP.2005.850882.

[11] C. Steffens, M. Pesavento, and M. E. Pfetsch, "A compact formulation for the $\ell_{2,1}$ mixed-norm minimization problem," *IEEE Trans. Signal Process.*, vol. 66, no. 6, pp. 1483–1497, Mar. 2018, doi: 10.1109/TSP.2017.2788431.

[12] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht, "Compressed sensing off the grid," *IEEE Trans. Inf. Theory*, vol. 59, no. 11, pp. 7465–7490, Nov. 2013, doi: 10.1109/TIT.2013.2277451.

[13] Z. Yang, L. Xie, and P. Stoica, "Vandermonde decomposition of multilevel Toeplitz matrices with application to multidimensional super-resolution," *IEEE Trans. Inf. Theory*, vol. 62, no. 6, pp. 3685–3701, Jun. 2016, doi: 10.1109/TIT.2016.2553041.

[14] A. Scaglione, R. Pagliari, and H. Krim, "The decentralized estimation of the sample covariance," in *Proc. 42nd Asilomar Conf. Signals, Syst. Comput.*, 2008, pp. 1722–1726, doi: 10.1109/ACSSC.2008.5074720.

[15] M. Pesavento, A. Gershman, and K. Wong, "Direction finding in partly calibrated sensor arrays composed of multiple subarrays," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2103–2115, Oct. 2002, doi: 10.1109/TSP.2002.801929.

[16] C. See and A. Gershman, "Direction-of-arrival estimation in partly calibrated subarray-based sensor arrays," *IEEE Trans. Signal Process.*, vol. 52, no. 2, pp. 329–338, Feb. 2004, doi: 10.1109/TSP.2003.821101.

[17] A. Moffet, "Minimum-redundancy linear arrays," *IEEE Trans. Antennas Propag.*, vol. 16, no. 2, pp. 172–175, Mar. 1968, doi: 10.1109/TAP.1968.1139138.

[18] Y. Abramovich, D. Gray, A. Gorokhov, and N. Spencer, "Positive-definite Toeplitz completion in DOA estimation for nonuniform linear antenna arrays. I. Fully augmentable arrays," *IEEE Trans. Signal Process.*, vol. 46, no. 9, pp. 2458–2471, Sep. 1998, doi: 10.1109/78.709534.

[19] P. Pal and P. P. Vaidyanathan, "Nested arrays: A novel approach to array processing with enhanced degrees of freedom," *IEEE Trans. Signal Process.*, vol. 58, no. 8, pp. 4167–4181, Aug. 2010, doi: 10.1109/TSP.2010.2049264.

[20] Z. Tan, Y. C. Eldar, and A. Nehorai, "Direction of arrival estimation using coprime arrays: A super resolution viewpoint," *IEEE Trans. Signal Process.*, vol. 62, no. 21, pp. 5565–5576, Nov. 2014, doi: 10.1109/TSP.2014.2354316.

[21] S. Qin, Y. D. Zhang, and M. G. Amin, "Generalized coprime array configurations for direction-of-arrival estimation," *IEEE Trans. Signal Process.*, vol. 63, no. 6, pp. 1377–1390, Mar. 2015, doi: 10.1109/TSP.2015.2393838.

[22] M. Viberg and B. Ottersten, "Sensor array processing based on subspace fitting," *IEEE Trans. Signal Process.*, vol. 39, no. 5, pp. 1110–1121, May 1991, doi: 10.1109/78.80966.

[23] B. Ottersten, P. Stoica, and R. Roy, "Covariance matching estimation techniques for array signal processing applications," *Digit. Signal Process.*, vol. 8, no. 3, pp. 185–210, Jul. 1998, doi: 10.1006/dspr.1998.0316.

[24] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification," *Int. J. Control*, vol. 50, no. 5, pp. 1873–1896, May 2007, doi: 10.1080/00207178908953472.

[25] I. Ziskind and M. Wax, "Maximum likelihood localization of multiple sources by alternating projection," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 10, pp. 1553–1560, Oct. 1988, doi: 10.1109/29.7543.

[26] M. M. Hyder and K. Mahata, "Direction-of-arrival estimation using a mixed $\ell_{2,0}$ norm approximation," *IEEE Trans. Signal Process.*, vol. 58, no. 9, pp. 4646–4655, Sep. 2010, doi: 10.1109/TSP.2010.2050477.

[27] Z. Yang and L. Xie, "Exact joint sparse frequency recovery via optimization methods," *IEEE Trans. Signal Process.*, vol. 64, no. 19, pp. 5145–5157, Oct. 2016, doi: 10.1109/TSP.2016.2576422.

[28] A. Manikas and C. Proukakis, "Modeling and estimation of ambiguities in linear arrays," *IEEE Trans. Signal Process.*, vol. 46, no. 8, pp. 2166–2179, Aug. 1998, doi: 10.1109/78.705428.

[29] F. Matter, T. Fischer, M. Pesavento, and M. E. Pfetsch, "Ambiguities in DoA estimation with linear arrays," *IEEE Trans. Signal Process.*, vol. 70, pp. 4395–4407, Aug. 2022, doi: 10.1109/TSP.2022.3200548.

[30] M. Babaie-Zadeh and C. Jutten, "On the stable recovery of the sparsest overcomplete representations in presence of noise," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5396–5400, Oct. 2010, doi: 10.1109/TSP.2010.2052357.

[31] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, Jul. 1989, doi: 10.1109/29.32276.

[32] A. Swindlehurst, P. Stoica, and M. Jansson, "Exploiting arrays with multiple invariances using MUSIC and MODE," *IEEE Trans. Signal Process.*, vol. 49, no. 11, pp. 2511–2521, Nov. 2001, doi: 10.1109/78.960398.

[33] H. Qiao and P. Pal, "Guaranteed localization of more sources than sensors with finite snapshots in multiple measurement vector models using difference co-arrays," *IEEE Trans. Signal Process.*, vol. 67, no. 22, pp. 5715–5729, Nov. 2019, doi: 10.1109/TSP.2019.2943224.

[34] M. Wang and A. Nehorai, "Coarrays, MUSIC, and the Cramér–Rao bound," *IEEE Trans. Signal Process.*, vol. 65, no. 4, pp. 933–946, Feb. 2017, doi: 10.1109/TSP.2016.2626255.

[35] Y. Chen and Y. Chi, "Robust spectral compressed sensing via structured matrix completion," *IEEE Trans. Inf. Theory*, vol. 60, no. 10, pp. 6576–6601, Oct. 2014, doi: 10.1109/TIT.2014.2343623.

[36] S. Sun and Y. D. Zhang, "4D automotive radar sensing for autonomous vehicles: A sparsity-oriented approach," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 4, pp. 879–891, Jun. 2021, doi: 10.1109/JSTSP.2021.3079626.

[37] S. Haghighatshoar and G. Caire, "Massive MIMO channel subspace estimation from low-dimensional projections," *IEEE Trans. Signal Process.*, vol. 65, no. 2, pp. 303–318, Jan. 2017, doi: 10.1109/TSP.2016.2616336.

[38] P. Sarangi, M. C. Hücümenoğlu, and P. Pal, "Single-snapshot nested virtual array completion: Necessary and sufficient conditions," *IEEE Signal Process. Lett.*, vol. 29, pp. 2113–2117, Oct. 2022, doi: 10.1109/LSP.2022.3213140.

[39] T. J. Moore, B. M. Sadler, and R. J. Kozick, "Maximum-likelihood estimation, the Cramér–Rao bound, and the method of scoring with parameter constraints," *IEEE Trans. Signal Process.*, vol. 56, no. 3, pp. 895–908, Mar. 2008, doi: 10.1109/TSP.2007.907814.

[40] S. Sedighi, B. S. Mysore, R. M. Soltanalian, and B. Ottersten, "On the performance of one-bit DoA estimation via sparse linear arrays," *IEEE Trans. Signal Process.*, vol. 69, pp. 6165–6182, Oct. 2021, doi: 10.1109/TSP.2021.3122290.

[41] A. M. Zoubir, V. Koivunen, E. Ollila, and M. Muma, *Robustness in Sensor Array Processing*. Cambridge, U.K.: Cambridge Univ. Press, 2018, pp. 125–146.

[42] J. P. Merkofer, G. Revach, N. Shlezinger, and R. J. G. van Sloun, "Deep augmented music algorithm for data-driven DOA estimation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2022, pp. 3598–3602, doi: 10.1109/ICASSP43922.2022.9746637.

SP