Wideband Spectrum Sensing From Compressed Measurements Using Spectral Prior Information

Daniel Romero, *Student Member, IEEE*, and Geert Leus, *Fellow, IEEE*

***Abstract*—Wideband spectrum sensing (WSS) encompasses a collection of techniques intended to estimate or to decide over the occupancy parameters of a wide frequency band. However, broad bands require expensive acquisition systems, thus motivating the use of compressive schemes. In this context, previous works in compressive WSS have already realized that great compression rates can be achieved if only second-order statistics are of interest in spectrum sensing. In this paper, we go a step further by exploiting spectral prior information that is typically available in practice in order to reduce the sampling rate even more. The signal model assumes that the acquisition is done by means of an analog-to-information converter (A2I). The input signal is the linear combination of a number of signals whose second-order statistics are known and the goal is to estimate/decide over the coefficients of this combination. The problem is thus a particular instance of the well-known structured covariance estimation problem. Unfortunately, the algorithms used in this area are extremely complex for inexpensive spectrum sensors so that alternative techniques need to be devised. Exploiting the fact that the basis matrices are Toeplitz, we use the asymptotic theory of circulant matrices to propose a dimensionality reduction technique that simplifies existing structured covariance estimation algorithms, achieving a similar performance at a much lower computational cost.**

***Index Terms*—Analog-to-information converters, compressed sensing, covariance matching, wideband spectrum sensing.**

I. Introduction

SPECTRUM sensing [1] encompasses a collection of procedures intended to determine the occupancy state of a particular frequency band, and it is of critical importance in certain applications such as those employing dynamic spectrum access (DSA) [2] (sometimes known as cognitive radio [3]). When several channels need to be scanned, a simple approach is to apply a conventional narrowband spectrum sensing procedure independently on a channel by channel basis. However, certain particularities arise from considering multiple channels at the same time, these being the subject of study of the so-called wideband spectrum sensing (WSS).

Previous wideband schemes include [4], where the throughput is maximized with an interference constraint on the primary network. The drawback though is that noise power knowledge is assumed, which makes the scheme sensitive to the noise uncertainty problem [5]. This problem is solved in [6], where this parameter is estimated using idle channels. However, in practice we may not guarantee the existence of such channels. This difficulty is overcome in [7] by assuming that the power spectral density (PSD) of the individual channels is known up to a scaling factor. Unfortunately, this scheme requires acquiring the wideband signal at the Nyquist rate, which drastically limits the maximum sensed bandwidth.

The sampling rate clearly represents a bottleneck for WSS. This observation has led to the concept of *compressive* WSS (see [8] and the references therein), where the sensing is done based on samples taken below the Nyquist rate. However, most schemes employ *lossless* compression, i.e., they aim to exactly reconstruct the original signal, so that assumptions about sparsity in the frequency domain are required. An alternative view that implements *lossy* compression stems from the observation that only the second-order statistics of the received signal are of interest, not the time-domain waveform itself [9], [10]. As a convenient byproduct, sparsity is typically not needed anymore. This standpoint is also used in [11], [12], where the signal is acquired in the spatial domain to estimate the direction of arrival (DoA), and in [13], where a distributed scheme to estimate the power spectrum from lossy observations is proposed.

The scheme in [9] provides a general spectrum estimation method based on compressed measurements, where the spectrum of the uncompressed signal is retrieved by inverting the linear relationship between the input and the output of the acquisition system. Interestingly, this setting allows two important refinements on the grounds that certain prior information about the power spectrum of the individual transmissions is typically available in practice. First, this information allows a more detailed interpretation than the raw spectral estimate of [9], enabling us to obtain the occupancy parameters more directly: we are only interested in estimating the power of each channel. Second, it suggests that drastic reductions in the sampling rates can be achieved since only a few parameters need to be estimated.

Manuscript received March 01, 2013; revised June 24, 2013 and August 22, 2013; accepted September 11, 2013. Date of publication September 25, 2013; date of current version November 13, 2013. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Emmanuel Candes. This work has been supported by the European Regional Development Fund (ERDF) and the Spanish Government (TEC2010-21245-C02-02/TCM DYNACS, CONSOLIDERINGENIO 2010 CSD2008-00010 COMONSENS), FPU Grant AP2010-0149, and the Galician Regional Government (CN 2012/260 AtlantTIC). This work is further supported by NWO-STW under the VICI program (project 10382). Parts of this work have been presented at the IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, Canada, 2013.

D. Romero is with the Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, 2628 CD Delft, The Netherlands, and also with the Department of Signal Theory and Communications, University of Vigo, 36310 Vigo, Spain (e-mail: dromero@gts.uvigo.es).

G. Leus is with the Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: g.j.t.leus@tudelft.nl).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2013.2283473

Typical prior information available about the signals being transmitted at a particular frequency band is often very detailed since many transmission systems today obey public standards. Among them we can mention, for example, the case of WiFi, Bluetooth, cellular telephony, digital television, etc., where the transmitted waveforms are strongly determined by the standard. This information, together with the specifications dictated by spectrum regulatory bodies, determine spectral masks, carrier frequencies, bandwidths, etc. Thus, it is reasonable to assume that the PSD of the individual transmissions is known up to a scaling factor. Consider, for instance, the case of a linear modulation such as QAM, where the PSD is determined by the roll-off factor of the square root raised cosine pulse and the bandwidth, which are typically known in practice. Another example is given by multi-carrier modulations, where the PSD can be inferred from the number of active carriers, the positions and power of the pilots, etc., which are also standardized in many transmission systems. Moreover, in those scenarios where different transmissions have disjoint frequency supports, we expect any reasonable sensing algorithm to tolerate moderate modeling errors in the PSD.

In the considered model, a frequency band of large bandwidth is acquired by means of an analog to information converter (A2I) [14]–[16], which provides linear measurements of the analog input signal at rates below the Nyquist rate. This signal is assumed to be the sum of a certain number of component signals whose second-order moments are known up to a scaling factor, one of them will typically be noise/interference. The goal is to estimate the power of these components based on the compressed observations provided by the A2I. These estimates may be used afterwards by some opportunistic node to decide over the occupancy of a channel in order to identify transmission opportunities, or they may be used by some device responsible for monitoring the correct usage of the spectrum.

In order to define estimation criteria, we start by assuming that the signal is Gaussian distributed, which is motivated in communication scenarios for several reasons: first, the Gaussian distribution is the distribution that achieves the capacity of an additive white Gaussian noise channel, so that many high-rate transmission signals are designed to follow an approximately Gaussian distribution. Second, multi-carrier communications, which are widespread nowadays, are approximately Gaussian distributed since they are generated as linear combinations of many independent (or nearly independent) subcarriers. Third, it is a working assumption very common in signal processing and statistics since it leads to tractable models and since many methods designed for Gaussian distributions also work even when there exist considerable departures from Gaussianity [17].

Using this assumption, we find that the algorithms for classical structured covariance estimation may be applied. However, the complexity of these algorithms limits their applicability to real-world scenarios, where the spectrum sensors are typically low-end devices that must minimize the sensing and processing time, both for properly exploiting the transmission opportunities and for saving battery [1], [2]. A dimensionality reduction technique is therefore presented in Section II based on two key observations: first, the input signals are assumed stationary so that their covariance matrices are redundant. Second, most information in practical signals is concentrated on the first few lags of the autocorrelation. In Section III, we review and modify some well-known structured covariance estimation algorithms to work in our setting, and we exploit this dimensionality reduction technique to propose efficient approximations that achieve a similar performance at a much lower computational cost. Later, the Gaussian assumption is released and a couple of algorithms are proposed for non-Gaussian signals. In this context, asymptotic considerations lead to an efficient algorithm that has been recently proposed in the context of array processing [18]. After discussing these algorithms for estimation in the time domain in Section III; we explore estimation in the frequency domain in Section IV. Finally, we asses the estimation and detection performance in Section V and highlight the main conclusions in Section VI.

Notation: Throughout, $\otimes$ denotes the Kronecker product, $\|\cdot\|$ the Euclidean norm, $\|\cdot\|_F$ the Frobenius norm, $|\cdot|$ the determinant, Tr the trace, $\mathrm{E}\{\cdot\}$ the expectation and $\boldsymbol{I}_P$ the $P \times P$ identity matrix. The superscripts $\cdot^H$ and $\cdot^\dagger$ represent, respectively, the conjugate transpose and the pseudo-inverse matrices. The notation vec means column-wise vectorization, and $\mathrm{diag}\{\boldsymbol{v}\}$ is a diagonal matrix with the components of $\boldsymbol{v}$ along its main diagonal. Finally, we denote as $[\boldsymbol{A}]_{i,j}$ the (i,j)-th element of the matrix $\boldsymbol{A}$ and $[\boldsymbol{b}]_i$ the i-th component of the vector $\boldsymbol{b}$.

II. Wideband Spectral Estimation and Detection

In this section, we formulate the spectrum sensing problem, which can be formalized as an estimation/detection problem. First, the observation model is presented. Next, we introduce the maximum likelihood (ML) framework and the dimensionality reduction techniques that are used in the following sections.

A. Observation Model

Suppose that a sensor receives a signal $x(t)$ that is a linear combination with unknown coefficients of I Gaussian component signals $x_i(t), i = 0, 1, \ldots, I-1$, which may be present or not, i.e., $x(t) = \sum_i \theta_i^{1/2} x_i(t)$, where some of the θ_i's may be zero. For concreteness, we assume that these signals are normalized such that $\mathrm{E}\{|x_i(t)|^2\} = 1$, so that the non-negative coefficients θ_i actually represent the power of each component. Estimating and deciding over these unknown coefficients θ_i is the goal of the statistical procedures developed in the rest of the paper. In a typical scenario, the signal $x_i(t)$ will correspond to the waveform being transmitted by the i-th user operating in the band. We also reserve one or more signals $x_i(t)$ to model noise and interference.

The conversion to the digital domain is carried out by an A2I that produces the M outputs $y_m[k], m = 0, 1, \ldots, M-1$, with k representing the discrete time index. Since no loss of information is incurred by sampling a signal at the Nyquist rate, a common abstraction when dealing with A2Is is to replace the analog input signal $x(t)$ by its Nyquist-sampled version $x[n]$ and to assume that the outputs $y_m[k]$ are the result of some linear manipulation of these input samples. Extending this idea to the components $x_i(t)$ it is possible to write $x[n] = \sum_i \theta_i^{1/2} x_i[n]$, where $x[n] = x(nT)$ and $x_i[n] = x_i(nT)$ are

the corresponding sampled sequences at a period T lower than the Nyquist period.[1]

The input samples $x[n]$ are arranged in groups of N, each one giving rise to one sample at every output $y_m[k]$. Specifically, we have that $y_m[k] = \boldsymbol{\phi}_m^H \boldsymbol{x}[k]$, where $\boldsymbol{\phi}_m \in \mathbb{C}^N$ is the measurement vector corresponding to the m-th output sequence and $\boldsymbol{x}[k] = [x[kN], x[kN+1], \dots x[kN+(N-1)]]^T$. By stacking these M outputs in the vector $\boldsymbol{y}[k] = [y_0[k], y_1[k], \dots y_{M-1}[k]]^T$ it is possible to write the more compact form $\boldsymbol{y}[k] = \boldsymbol{\Phi}\boldsymbol{x}[k]$, where $\boldsymbol{\Phi} = [\boldsymbol{\phi}_0, \boldsymbol{\phi}_1, \dots \boldsymbol{\phi}_{M-1}]^H$ is the sampling matrix. Further, if we arrange all observations together, we can form the vector[2] $\boldsymbol{y} = [\boldsymbol{y}^T[0], \boldsymbol{y}^T[1], \dots, \boldsymbol{y}^T[K/N-1]]^T$, with KT the observation time, and write $\boldsymbol{y} = \bar{\boldsymbol{\Phi}}\boldsymbol{x}$, where $\bar{\boldsymbol{\Phi}} = \boldsymbol{I}_{K/N} \otimes \boldsymbol{\Phi}$ and $\boldsymbol{x} = [\boldsymbol{x}^T[0], \boldsymbol{x}^T[1], \dots, \boldsymbol{x}^T[K/N-1]]^T$.

The input signals $x_i[n]$ are assumed to be wide-sense stationary random processes, independent of each other, with zero mean and known autocorrelation function $r_i[n] = \mathrm{E}\{x_i[\nu+n]x_i^*[\nu]\}, \forall \nu$. The Fourier transform of the autocorrelation sequence $r_i[n]$ is assumed to exist and it is the PSD of the process $x_i[n]$. The statistics $r_i[n]$ can be arranged as the elements of the Hermitian Toeplitz covariance matrices $\boldsymbol{\Sigma}_i = \mathrm{E}\{\boldsymbol{x}_i\boldsymbol{x}_i^H\}$, where $\boldsymbol{x}_i = [x_i[0], x_i[1], \dots, x_i[K-1]]^T$. The set of matrices $\mathcal{S} = \{\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_{I-1}\}$ is assumed $\mathbb{R}$-linearly independent, in the sense that no two different linear combinations of these matrices taking real coefficients can result in the same matrix, i.e.,

$$\sum_i \alpha_i \boldsymbol{\Sigma}_i = \sum_i \beta_i \boldsymbol{\Sigma}_i \quad \Rightarrow \quad \alpha_i = \beta_i \ \forall i. \tag{1}$$

The vector $\boldsymbol{x}$ containing the received signal, now written as $\boldsymbol{x} = \sum_i \theta_i^{1/2}\boldsymbol{x}_i$, will have a covariance matrix given by $\boldsymbol{\Sigma} = \mathrm{E}\{\boldsymbol{x}\boldsymbol{x}^H\} = \sum_i \theta_i \boldsymbol{\Sigma}_i$ since $\boldsymbol{x}_i$ was assumed independent of (and thus uncorrelated with) $\boldsymbol{x}_j$ for $i \neq j$. Clearly, $\boldsymbol{\Sigma}$ is Hermitian Toeplitz since $\boldsymbol{x}$ is stationary. For convenience we also consider its decomposition into $N \times N$ blocks:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}[0] & \boldsymbol{\Sigma}^H[1] & \dots & \boldsymbol{\Sigma}^H\left[\frac{K}{N}-1\right] \\ \boldsymbol{\Sigma}[1] & \boldsymbol{\Sigma}[0] & \dots & \boldsymbol{\Sigma}^H\left[\frac{K}{N}-2\right] \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}\left[\frac{K}{N}-1\right] & \boldsymbol{\Sigma}\left[\frac{K}{N}-2\right] & \dots & \boldsymbol{\Sigma}[0] \end{bmatrix}$$

where the blocks $\boldsymbol{\Sigma}[k]$ are given by $\boldsymbol{\Sigma}[k] = \mathrm{E}\{\boldsymbol{x}[\kappa+k]\boldsymbol{x}^H[\kappa]\}$. A similar decomposition applied over $\boldsymbol{\Sigma}_i$ results in blocks

$$\boldsymbol{\Sigma}_i[k] = \mathrm{E}\left\{\boldsymbol{x}_i[\kappa+k]\boldsymbol{x}_i^H[\kappa]\right\},$$

where

$$\boldsymbol{x}_i[k] = [x_i[kN], x_i[kN+1], \dots x_i[kN+(N-1)]]^T.$$

The covariance matrix of $\boldsymbol{y}$ is clearly

$$\bar{\boldsymbol{\Sigma}} = \mathrm{E}\{\boldsymbol{y}\boldsymbol{y}^H\} = \bar{\boldsymbol{\Phi}}\boldsymbol{\Sigma}\bar{\boldsymbol{\Phi}}^H = \sum_i \theta_i \bar{\boldsymbol{\Sigma}}_i, \tag{2}$$

where $\bar{\boldsymbol{\Sigma}}_i = \bar{\boldsymbol{\Phi}}\boldsymbol{\Sigma}_i\bar{\boldsymbol{\Phi}}^H$. Note that expression (2) decomposes $\bar{\boldsymbol{\Sigma}}$ in terms of the covariance matrices of the *compressed domain* basis $\bar{\mathcal{S}} = \{\bar{\boldsymbol{\Sigma}}_0, \bar{\boldsymbol{\Sigma}}_1, \dots, \bar{\boldsymbol{\Sigma}}_{I-1}\}$. These matrices are similarly decomposed in blocks $\bar{\boldsymbol{\Sigma}}[k] = \boldsymbol{\Phi}\boldsymbol{\Sigma}[k]\boldsymbol{\Phi}^H$ and $\bar{\boldsymbol{\Sigma}}_i[k] = \boldsymbol{\Phi}\boldsymbol{\Sigma}_i[k]\boldsymbol{\Phi}^H$, in this case of size $M \times M$. Although these blocks are not Toeplitz, the matrices $\bar{\boldsymbol{\Sigma}}$ and $\bar{\boldsymbol{\Sigma}}_i$ are block-wise Toeplitz, which means that the processes $y_m[k]$ and $y_{m'}[k]$ are jointly stationary.

B. Estimation and Detection

As mentioned in Section I, we start by considering that the amplitude of the signals involved is zero-mean Gaussian distributed. The statistical characterization of the observations $\boldsymbol{y}$ is thus completely determined by the second-order statistics introduced in the previous section. The probability density function (PDF) of the observations can be written as:

$$p(\boldsymbol{y};\boldsymbol{\theta}) = \frac{\exp\{-\boldsymbol{y}^H\bar{\boldsymbol{\Sigma}}^{-1}\boldsymbol{y}\}}{\pi^{\frac{MK}{N}}|\bar{\boldsymbol{\Sigma}}|}, \tag{3}$$

where $\boldsymbol{\theta} = [\theta_0, \theta_1, \dots, \theta_{I-1}]^T$ is the vector of unknown parameters. Although $\bar{\boldsymbol{\Sigma}}$ also depends on this vector, we dismiss the notation $\bar{\boldsymbol{\Sigma}}(\boldsymbol{\theta})$ in favor of clarity, but this dependence should be kept in mind throughout.

The ML estimate [19] of $\boldsymbol{\theta}$ given $\boldsymbol{y}$ is the maximizer of the PDF in (3) seen as a likelihood function, that is, $\boldsymbol{\theta}_{ML} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{y};\boldsymbol{\theta})$, where the feasible set for $\boldsymbol{\theta}$ is the non-negative orthant, i.e., $\{\boldsymbol{\theta} \in \mathbb{R}^I : \theta_i \geq 0\}$. This optimization problem has been widely analyzed and, up to now, no analytical solution is known. It is therefore necessary to deal with numerical methods capable of maximizing (3), which will be the subject of Sections III and IV.

We are also interested in testing for the presence of the signals $x_i[n]$, which naturally leads to an involved multiple hypothesis test with 2^I hypotheses. Since in applications like DSA we are typically interested in some particular signal, say $x_j[n]$, a more reasonable option is to perform a binary hypothesis test based on this signal only. This is also motivated by simplicity and because the binary hypothesis testing framework is a more mature and developed field in statistics than multiple hypothesis testing. Without any loss of generality, we can assume that $j = 0$, so that the problem can be formally stated as that of deciding between the following two hypotheses:

$$\mathcal{H}_0 : \theta_0 = 0, \quad \mathcal{H}_1 : \theta_0 > 0$$

The presence of the unknown parameters $\theta_i, i = 1, 2, \dots I-1$, means that no test will exist, in general, that is optimal in the Neyman-Pearson sense [20]. A sensible option arising from ML criteria is to make the decision based on the generalized likelihood ratio (GLR) statistic $\mathcal{L}(\boldsymbol{y})$, which leads to the GLR test (GLRT) [17], [20], [21]:

$$\mathcal{L}(\boldsymbol{y}) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \eta. \tag{4}$$

Here, η is some predefined threshold, set to satisfy certain probability of detection/false alarm requirements. The notation in (4)

[1]Note that this sampling process is a virtual process that does not actually take place; it is only introduced here to simplify the presentation of the A2I. Further note that the considered samples $x[n] = x(nT)$ and $x_i[n] = x_i(nT)$ could similarly be modeled by means of more realistic (virtual) samplers applied to the analog signals.

[2]Throughout it will be assumed that K is an integer multiple of N.

means that we decide that $\mathcal{H}_1$ is true for high values of $\mathcal{L}(\boldsymbol{y})$ and $\mathcal{H}_0$ otherwise.

The GLR statistic is defined as follows:

$$\mathcal{L}(\boldsymbol{y}) = \frac{p\left(\boldsymbol{y};\hat{\boldsymbol{\theta}}_{\mathcal{H}_1}\right)}{p\left(\boldsymbol{y};\hat{\boldsymbol{\theta}}_{\mathcal{H}_0}\right)} = \frac{\left|\bar{\boldsymbol{\Sigma}}\left(\hat{\boldsymbol{\theta}}_{\mathcal{H}_0}\right)\right|}{\left|\bar{\boldsymbol{\Sigma}}\left(\hat{\boldsymbol{\theta}}_{\mathcal{H}_1}\right)\right|} \frac{\exp\left\{-\boldsymbol{y}^H\bar{\boldsymbol{\Sigma}}^{-1}\left(\hat{\boldsymbol{\theta}}_{\mathcal{H}_1}\right)\boldsymbol{y}\right\}}{\exp\left\{-\boldsymbol{y}^H\bar{\boldsymbol{\Sigma}}^{-1}\left(\hat{\boldsymbol{\theta}}_{\mathcal{H}_0}\right)\boldsymbol{y}\right\}} \tag{5}$$

where the vectors $\hat{\boldsymbol{\theta}}_{\mathcal{H}_0}$ and $\hat{\boldsymbol{\theta}}_{\mathcal{H}_1}$ denote, respectively, the ML estimates of $\boldsymbol{\theta}$ under hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$. In particular, $\hat{\boldsymbol{\theta}}_{\mathcal{H}_0}$ is the maximizer of (3) subject to[3] $\theta_0 = 0$ and $\theta_i \geq 0$, $i = 1,\ldots,I-1$, whereas $\hat{\boldsymbol{\theta}}_{\mathcal{H}_1}$ is the maximizer of (3) subject to $\theta_0 > 0$ and $\theta_i \geq 0$, $i = 1,\ldots,I-1$. For concreteness, but without loss of generality, we only discuss the estimation of $\boldsymbol{\theta}$ subject to the constraints that all θ_i are non-negative. Note that both the estimation of $\hat{\boldsymbol{\theta}}_{\mathcal{H}_0}$ and $\hat{\boldsymbol{\theta}}_{\mathcal{H}_1}$ can be formulated as particular instances of this problem: in the first case, it is easy to see that $\hat{\boldsymbol{\theta}}_{\mathcal{H}_0}$ can be obtained in this way just by removing $\bar{\boldsymbol{\Sigma}}_0$ from $\bar{\mathcal{S}}$. In the second case, it is readily seen that both problems are equivalent since the likelihood function is continuous in θ_0 so that using the constraint $\theta_0 > 0$ amounts to using the constraint $\theta_0 \geq 0$.

C. Complexity of the Exact ML Solution

It can be easily seen that the complexity of obtaining the exact ML solution is prohibitive to a conventional sensor even in cases where the size of the compressed data record MK/N is of the order of a few hundreds. This is due to the fact that MK/N is also the size of the matrices involved and typical algorithms need to evaluate their inverses and determinants at each iteration. Moreover, certain numerical instabilities are expected in these algorithms, requiring many checks that slow down the execution even more. To see that, let us look at the problem from an alternative perspective.

Observe that expression (3) can also be written in terms of the sample covariance matrix (SCM) $\boldsymbol{y}\boldsymbol{y}^H$ as

$$p(\boldsymbol{y};\boldsymbol{\theta}) = \frac{\exp\left\{-\mathrm{Tr}(\bar{\boldsymbol{\Sigma}}^{-1}\boldsymbol{y}\boldsymbol{y}^H)\right\}}{\pi^{\frac{MK}{N}}|\bar{\boldsymbol{\Sigma}}|}. \tag{6}$$

The problem is to find a matrix $\bar{\boldsymbol{\Sigma}}$ maximizing the metric in (6), which is a measure of fit between $\boldsymbol{y}\boldsymbol{y}^H$ and $\bar{\boldsymbol{\Sigma}}$. This matrix has to be sought in the feasible set, which is the intersection of the subspace of $MK/N \times MK/N$ matrices spanned by $\bar{\mathcal{S}}$ with non-negative coordinates and the cone of positive definite matrices. Note that this condition is necessary in order to evaluate (6), since if we allow semi-definite matrices the determinant may vanish and the inverse may not exist. A conventional iterative algorithm will move across the feasible set ensuring that the iterate stays far enough from the boundary of the cone (i.e., the set of singular positive semi-definite matrices) since otherwise the algorithm will be unstable. This goal is particularly hampered by the fact that $\boldsymbol{y}\boldsymbol{y}^H$ is exactly on that border: note that this matrix has one positive eigenvalue with multiplicity 1 and a null eigenvalue with multiplicity $MK/N - 1$.

[3]In this paper, we assume that all matrices in $\bar{\mathcal{S}}$ are positive definite, so that $\bar{\boldsymbol{\Sigma}}$ is positive definite if $\theta_i \geq 0$, $i = 0,\ldots,I-1$.

D. Dimensionality Reduction

As opposed to many statistical problems where the number of samples increases the cost of the solution linearly, here a higher amount of samples increases the dimension of the problem, forcing us to work with larger matrices, whose determinants and inverses are more difficult to compute and ill-conditioned. Moreover, the problem described in the previous section would still be present. Intuition suggests attempting to approximate the ML solution by replacing the raw SCM with a modified version satisfying two properties:

- **Fixed dimension**: the size of the modified SCM does not increase with the number of samples. It is kept low enough in order to guarantee easy evaluation of the likelihood function.
- **Non-singular**: the modified SCM should be full rank whenever the number of samples is high enough. This would overcome the problem described in the previous section.

Moreover, observe that the raw covariance matrix $\boldsymbol{y}\boldsymbol{y}^H$ does not even have the block structure of $\bar{\boldsymbol{\Sigma}}$, which also suggests looking for an SCM of that particular form. The asymptotic theory of Toeplitz matrices will give us some clue about how to accomplish this search.

1) Averaging the SCM: As we know [22], given a sequence of $n \times n$ Toeplitz covariance matrices $\{\boldsymbol{T}_n\}$ satisfying the assumption that the Fourier transform of the associated correlation sequences[4] exists, it is possible to find a sequence of $n \times n$ circulant matrices $\{\boldsymbol{C}_n\}$ which is asymptotically equivalent. The concept of asymptotic equivalence means that the products $\mathrm{Tr}(\boldsymbol{T}_n^{-1}\boldsymbol{A}_n)$ and $\mathrm{Tr}(\boldsymbol{C}_n^{-1}\boldsymbol{A}_n)$ will converge to the same value provided that $\{\boldsymbol{A}_n\}$ is a sequence of matrices bounded in some norm. The sequences of determinants $|\boldsymbol{T}_n|$ and $|\boldsymbol{C}_n|$ will also have the same limit as long as these matrices are non-singular. Since these are the only operations we are going to perform with these matrices, we will allow ourselves to say that the sequence of matrices $\{\boldsymbol{T}_n\}$ is *asymptotically circulant* or, for a finite n, *approximately circulant*. Likewise, since any circulant matrix can be diagonalized by the vectors in the Fourier basis, we can also say that these are, asymptotically (or approximately), the eigenvectors of $\boldsymbol{T}_n$.

With this in mind, we may say that, as MK/N becomes larger, $\bar{\boldsymbol{\Sigma}}$ is *approximately block circulant* and, consequently, it will remain approximately the same after a circular rotation of the block-rows and block-columns. More formally, this can be expressed as $\bar{\boldsymbol{R}}_k\bar{\boldsymbol{\Sigma}}\bar{\boldsymbol{R}}_k^T \approx \bar{\boldsymbol{\Sigma}}$, where $\bar{\boldsymbol{R}}_k$ denotes the matrix performing a block-row circular rotation of the k-th order, i.e., $\bar{\boldsymbol{R}}_k = \boldsymbol{R}_k \otimes \boldsymbol{I}_M$ where $\boldsymbol{R}_k$ circularly shifts the columns of the identity matrix $\boldsymbol{I}_{K/N}$ k positions to the bottom. This enables us to approximate the density of the observations as

$$\begin{aligned} p(\boldsymbol{y};\boldsymbol{\theta}) &= \frac{\exp\left\{-\mathrm{Tr}\left(\bar{\boldsymbol{\Sigma}}^{-1}\boldsymbol{y}\boldsymbol{y}^H\right)\right\}}{\pi^{\frac{MK}{N}}|\bar{\boldsymbol{\Sigma}}|} \\ &\approx \frac{\exp\left\{-\mathrm{Tr}\left(\left(\bar{\boldsymbol{R}}_k\bar{\boldsymbol{\Sigma}}\bar{\boldsymbol{R}}_k^T\right)^{-1}\boldsymbol{y}\boldsymbol{y}^H\right)\right\}}{\pi^{\frac{MK}{N}}|\bar{\boldsymbol{\Sigma}}|}, \end{aligned}$$

[4]The correlation sequence associated with the Toeplitz matrix $\boldsymbol{T}_n$ is made up of the coefficients in the first row and column of $\boldsymbol{T}_n$ set in the proper order.

where we have made use of the fact that the squared determinant $|\bar{\boldsymbol{R}}_k|^2$ is one. Noting that

$$\mathrm{Tr}\left(\left(\bar{\boldsymbol{R}}_k\bar{\boldsymbol{\Sigma}}\bar{\boldsymbol{R}}_k^T\right)^{-1}\boldsymbol{y}\boldsymbol{y}^H\right)=\mathrm{Tr}\left(\bar{\boldsymbol{\Sigma}}^{-1}\bar{\boldsymbol{R}}_k^T\boldsymbol{y}\boldsymbol{y}^H\bar{\boldsymbol{R}}_k\right),$$

for all $k=0,1\ldots K/N-1$ shows that $\mathrm{Tr}(\bar{\boldsymbol{\Sigma}}^{-1}\boldsymbol{y}\boldsymbol{y}^H)$ can actually be approximated as

$$\begin{aligned}\mathrm{Tr}\left(\bar{\boldsymbol{\Sigma}}^{-1}\boldsymbol{y}\boldsymbol{y}^H\right)&\approx\frac{1}{K/N}\sum_{k=0}^{K/N-1}\mathrm{Tr}\left(\bar{\boldsymbol{\Sigma}}^{-1}\bar{\boldsymbol{R}}_k^T\boldsymbol{y}\boldsymbol{y}^H\bar{\boldsymbol{R}}_k\right)\\&=\mathrm{Tr}(\bar{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{S}})\end{aligned}$$

where we have defined the averaged SCM as $\hat{\boldsymbol{S}}=(K/N)^{-1}\sum_k\bar{\boldsymbol{R}}_k^T\boldsymbol{y}\boldsymbol{y}^H\bar{\boldsymbol{R}}_k$. In other words, we have that

$$p(\boldsymbol{y};\boldsymbol{\theta})\approx\frac{\exp\left\{-\mathrm{Tr}(\bar{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{S}})\right\}}{\pi^{\frac{MK}{N}}|\bar{\boldsymbol{\Sigma}}|}.\tag{7}$$

This means that[5] averaging along the *modular* block-diagonals of the SCM has a small influence on the likelihood function. However, in our case $\bar{\boldsymbol{\Sigma}}$ is not block-circulant but block-Toeplitz, so that we may prefer to simply average along the actual diagonals, i.e., to take

$$\hat{\boldsymbol{S}}=\begin{bmatrix}\hat{\boldsymbol{S}}[0]&\hat{\boldsymbol{S}}^H[1]&\ldots&\hat{\boldsymbol{S}}^H\left[\frac{K}{N}-1\right]\\\hat{\boldsymbol{S}}[1]&\hat{\boldsymbol{S}}[0]&\ldots&\hat{\boldsymbol{S}}^H\left[\frac{K}{N}-2\right]\\\vdots&\vdots&\ddots&\vdots\\\hat{\boldsymbol{S}}\left[\frac{K}{N}-1\right]&\hat{\boldsymbol{S}}\left[\frac{K}{N}-2\right]&\ldots&\hat{\boldsymbol{S}}[0]\end{bmatrix},\tag{8}$$

where $\hat{\boldsymbol{S}}[k]=\frac{1}{K_k}\sum_\kappa\boldsymbol{y}[\kappa+k]\boldsymbol{y}^H[\kappa]$ and K_k is a constant depending on k. If $K_k=\frac{K}{N}$, then $\hat{\boldsymbol{S}}$ is actually composed of the traditional biased estimates [23] of the autocorrelation/crosscorrelation of the processes $y_m[k]$. On the other hand, if we set $K_k=\frac{K}{N}-k$, then we are using the unbiased estimate, which may be more justified in the general case since then K_k also represents the length of the k-th block-diagonal of $\hat{\boldsymbol{S}}$, i.e., the number of terms averaged. As we know [24], the use of either one of these estimates will guarantee that $\hat{\boldsymbol{S}}[k]$ is a consistent estimator for $\boldsymbol{\Sigma}[k]$ for a fixed k as we take $K\to\infty$. Note that the simpler estimate $\boldsymbol{y}\boldsymbol{y}^H$ lacks this property. In virtue of the interpretation of Section II-C, consistency means that this modification of the SCM becomes *closer*, as K increases, to the feasible set where $\bar{\boldsymbol{\Sigma}}$ must be sought, and intuition suggests that this is numerically convenient. Moreover, the averaged SCM $\hat{\boldsymbol{S}}$ has the further advantage of presenting the same block structure as the true covariance matrix $\bar{\boldsymbol{\Sigma}}$.

2) Cropping the Covariance Matrix: Now that the averaging has been motivated by arguing that the likelihood function should not be considerably affected after this procedure, it seems reasonable to consider reducing the dimension of the problem by exploiting the redundancy present in both the covariance matrix and the SCM.

Two observations apply: first, due to the block-Toeplitz structure of the covariance and averaged SCM, a large part of the coefficients is replicated several times. Note that this redundancy is reduced as we move further away from the main diagonal. Second, the coefficients of $\hat{\boldsymbol{S}}$ that are closer to the diagonal have a lower estimation variance and, consequently, are more *reliable*. If $\hat{\boldsymbol{S}}$ has the block structure in (8), these remarks suggest truncating both the covariance matrix and the SCM, disregarding the blocks $\bar{\boldsymbol{\Sigma}}[k]$ and $\hat{\boldsymbol{S}}[k]$ for high k, for example retaining only the blocks corresponding to lags not greater than L, i.e., we only keep the blocks $\hat{\boldsymbol{S}}[0],\hat{\boldsymbol{S}}[1],\ldots,\hat{\boldsymbol{S}}[L]$ of the modified SCM in (8). This parameter L was already used in [9], [25], [26] where, like here, it is regarded as a design parameter. Observe that after applying this selection, the matrices $\bar{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{S}}$ have dimension $M(L+1)\times M(L+1)$. For example, the *averaged and cropped* SCM is given by

$$\hat{\boldsymbol{S}}=\begin{bmatrix}\hat{\boldsymbol{S}}[0]&\hat{\boldsymbol{S}}^H[1]&\ldots&\hat{\boldsymbol{S}}^H[L]\\\hat{\boldsymbol{S}}[1]&\hat{\boldsymbol{S}}[0]&\ldots&\hat{\boldsymbol{S}}^H[L-1]\\\vdots&\vdots&\ddots&\vdots\\\hat{\boldsymbol{S}}[L]&\hat{\boldsymbol{S}}[L-1]&\ldots&\hat{\boldsymbol{S}}[0]\end{bmatrix}.\tag{9}$$

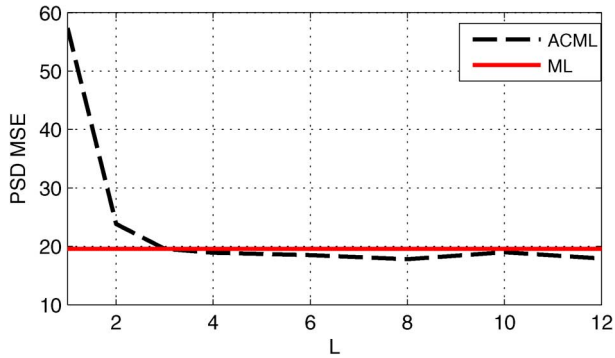


Fig. 1. Comparison between the true ML solution and that incorporating averaging and cropping (ACML). $K=325$, $\boldsymbol{\theta}=[4,9,4,9,4,9,1]^T$, Gaussian A2I, $\sigma^2_{\mathrm{A2I}}=1$, $N=5$, $M=2$.

There are further arguments supporting averaging and cropping, but the ultimate reason we may take into account is performance. In Fig. 1, we compare the estimation mean squared error (MSE) of the PSD obtained by the true ML solution (full red line) with the MSE of the scheme incorporating averaging and cropping (dashed black line). The signal $x[n]$ is composed of six bandpass signals with non-overlapping spectra of power [4,9,4,9,4,9], and white noise of power 1. Gaussian compression is carried out from $N=5$ samples to $M=2$ samples. More details about these computations can be found in Section V. We observe that a value of $L=3$ is enough to preserve most second-order information. From that point on, the MSE of the ML and the one based on using the modified SCM is roughly the same.

3) Properties of the Averaged and Cropped SCM: In the algorithms presented in the subsequent sections, a couple of properties of the modified SCM will be required. In particular, in some cases it is required to be positive semi-definite since a matrix square root must be computed. In other cases, the inverse must exist or the determinant must be different from zero. These properties are affected by the selection of the design parameters K, M, N and L. The following theorem, whose proof can be found in Appendix A, summarizes which combinations

[5]Note that we have unintentionally derived the exact ML estimate of $\bar{\boldsymbol{\Sigma}}$ if the matrices in $\bar{\mathcal{S}}$ were block-circulant.

of these parameters result in positive semi-definite matrices or invertible matrices.

Theorem 1: Under a continuous probability model for $\boldsymbol{y}$ such as that in (7), let $\hat{\boldsymbol{S}}$ be given by (9). Then, we have the following:

1) If the values of K_k correspond to the biased estimator of the autocorrelation, then $\hat{\boldsymbol{S}}$ is positive semi-definite.
2) If the values of K_k correspond to the biased estimator of the autocorrelation, we have that
$$\operatorname{rank}(\hat{\boldsymbol{S}}) = \min\left(\frac{K}{N} + L, M(L+1)\right)$$
with probability one.
3) If the values of K_k correspond to the unbiased estimator of the autocorrelation, we have that
$$\begin{aligned}\operatorname{rank}(\hat{\boldsymbol{S}}) &= \min\left(\frac{K}{N}(L+1), M(L+1)\right)\\ &= (L+1)\min\left(\frac{K}{N}, M\right)\end{aligned}$$
with probability one.

The hypothesis in this theorem assumes that $\boldsymbol{y}$ is distributed according to a continuous *non-degenerate* probability distribution, in the sense that no set of zero Lebesgue measure can have a non-null probability[6] (see e.g. [27]). In our case, this condition will hold provided that the covariance matrix $\bar{\boldsymbol{\Sigma}}$ is full rank. The reason for such an assumption is that we can take advantage of the fact that $\boldsymbol{y}[k_1]$ and $\boldsymbol{y}[k_2]$ are then linearly independent with probability 1 if $k_1 \neq k_2$.

The first part of the theorem guarantees that $\hat{\boldsymbol{S}}$ is positive semi-definite in case that the biased estimate is used, but the same cannot be said when we use the unbiased estimate, whose eigenvalues can be either positive or negative. Intuitively, this is because the coefficients far from the main diagonal can take very high values since the averaging needed for their computation is performed over just a few samples.

From the second part of Theorem 1 we conclude that $K/N + L$ has to be greater than or equal to $M(L+1)$ in order to impose that $\hat{\boldsymbol{S}}$ is invertible in the biased case. In other words, the invertibility of $\hat{\boldsymbol{S}}$ imposes the following inequality over the number of samples required:

$$K \geq N\left[M(L+1) - L\right]. \tag{10}$$

On the other hand, when the unbiased estimator is used, the third part of Theorem 1 says that the number of samples K must be greater than or equal to the product MN. Interestingly, this relation does not depend on L, which can be adjusted with complete freedom. Moreover, it is readily seen that the minimum K required for $\hat{\boldsymbol{S}}$ to be full rank is much lower, in general, for the unbiased estimator, thus motivating its use.

Finally, note that there is a further restriction on the values these design parameters can take on. In particular, the estimation problem must allow a solution, i.e., the coefficients θ_i must remain identifiable [24] after the A2I. This topic requires extensive analysis since it depends on the specific architecture of the acquisition system. For the case of periodic non-uniform sampling, this analysis can be found in [28]. Further results in this context will be the object of subsequent publications.

[6]In this case, it is usually said that the distribution of $\boldsymbol{y}$ is *absolutely continuous* with respect to Lebesgue measure.

III. Estimation in the Time Domain

As we saw in Section II-B, the detection problem reduces to computing two vector estimates, one under each hypothesis. The result of this computation is directly plugged into (5), where no simplification is possible since the estimates are computed numerically. For this reason, we only discuss estimation in this section as well as in Section IV, leaving the analysis of the detection problem to Section V. We first assume Gaussianity and discuss in detail how to approximate the ML estimate either in a constrained or unconstrained way. Based on the same ideas, we next discuss estimators that can work even in the non-Gaussian case.

A. The True ML Estimate

In order to find the ML estimate of $\boldsymbol{\theta}$ one must maximize (3) with respect to $\boldsymbol{\theta}$ given the observation $\boldsymbol{y}$ or, equivalently, minimize the negative of the log-likelihood function:

$$L(\boldsymbol{\theta}) = \log|\bar{\boldsymbol{\Sigma}}| + \operatorname{Tr}(\bar{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{R}}) \tag{11}$$

where $\hat{\boldsymbol{R}} = \boldsymbol{y}\boldsymbol{y}^H$. This minimization must be carried out by taking into account the constraints $\theta_i \geq 0, \forall i$.

In order to minimize (11), one must resort to numerical methods such as LIKES (likelihood-based estimation of sparse parameters). LIKES is an algorithm proposed in [29] for rank-1 SCMs in the context of sparse-parameter estimation and subsequently extended to the full-rank case in [30]. In both cases the basis covariance matrices $\bar{\boldsymbol{\Sigma}}_i$ are assumed rank-1, but a simple extension to the general case is possible as discussed next. We next briefly summarize the derivation of LIKES, with application to our problem.

The algorithm is based on the minimization-majorization (MM) principle, which locally majorizes the cost by a convex function at each iteration and minimizes this function to obtain the next estimate. Suppose that the current iterate is given by $\underline{\boldsymbol{\theta}}$. Since the first term in (11) is concave whereas the second one is convex, it suffices to majorize the first one by its tangent plane at $\boldsymbol{\theta} = \underline{\boldsymbol{\theta}}$:

$$\begin{aligned}\log|\bar{\boldsymbol{\Sigma}}| &\leq \log|\underline{\bar{\boldsymbol{\Sigma}}}| + \sum_{i=0}^{I-1}\operatorname{Tr}(\underline{\bar{\boldsymbol{\Sigma}}}^{-1}\bar{\boldsymbol{\Sigma}}_i)(\theta_i - \underline{\theta}_i)\\ &= \log|\underline{\bar{\boldsymbol{\Sigma}}}| - M(L+1) + \sum_{i=0}^{I-1}\operatorname{Tr}(\underline{\bar{\boldsymbol{\Sigma}}}^{-1}\bar{\boldsymbol{\Sigma}}_i)\theta_i\end{aligned} \tag{12}$$

where the parameters using the *under-bar* notation are those associated with $\underline{\boldsymbol{\theta}}$. Therefore the cost in (11) can be majorized as

$$L(\boldsymbol{\theta}) \leq \log|\underline{\bar{\boldsymbol{\Sigma}}}| - M(L+1) + \boldsymbol{y}^H\bar{\boldsymbol{\Sigma}}^{-1}\boldsymbol{y} + \sum_{i=0}^{I-1}\operatorname{Tr}(\underline{\bar{\boldsymbol{\Sigma}}}^{-1}\bar{\boldsymbol{\Sigma}}_i)\theta_i \tag{13}$$

so that we must solve the program

$$\underset{\boldsymbol{\theta}\succeq\mathbf{0}}{\text{minimize}} \quad \boldsymbol{y}^H\bar{\boldsymbol{\Sigma}}^{-1}\boldsymbol{y} + \sum_{i=0}^{I-1}\operatorname{Tr}(\underline{\bar{\boldsymbol{\Sigma}}}^{-1}\bar{\boldsymbol{\Sigma}}_i)\theta_i \tag{14}$$

at every iteration, where the notation $\boldsymbol{\theta} \succeq \mathbf{0}$ means that the elements of $\boldsymbol{\theta}$ are non-negative. The cost function in (14) is actually the SPICE cost function if we make $\beta_i = \operatorname{Tr}(\underline{\bar{\boldsymbol{\Sigma}}}^{-1}\bar{\boldsymbol{\Sigma}}_i)$, as we will see in Section III-D. LIKES is an algorithm that iteratively

uses SPICE to refine its current estimate. At every iteration, the weight vector $\boldsymbol{\beta} = [\beta_0, \ldots, \beta_{I-1}]^T$ is computed and SPICE executed. The algorithm is summarized below:

Algorithm $\boldsymbol{\theta} = \mathbf{LIKES}(\boldsymbol{y})$

Initialize $\boldsymbol{\theta} = [\theta_0, \ldots, \theta_{I-1}]^T$

while stopping_criterion == FALSE **do**
- $\bar{\boldsymbol{\Sigma}} \leftarrow \sum_i \theta_i \bar{\boldsymbol{\Sigma}}_i$
- $\beta_i \leftarrow \mathrm{Tr}(\bar{\boldsymbol{\Sigma}}^{-1}\bar{\boldsymbol{\Sigma}}_i)$
- $\boldsymbol{\theta} \leftarrow SPICE\_ITERATION(\boldsymbol{\beta}, \boldsymbol{y})$

end while

Unfortunately, both the matrix inversion needed for the computation of $\boldsymbol{\beta}$ and the execution of SPICE are considerably slow operations for observation vectors $\boldsymbol{y}$ of moderate sizes. Following the guidelines in Section II, we propose to replace the $MK/N \times MK/N$ SCM $\hat{\boldsymbol{R}}$ in (11) by its averaged and cropped version $\hat{\boldsymbol{S}}$, which is just $M(L+1) \times M(L+1)$. It is understood that the matrices $\bar{\boldsymbol{\Sigma}}$ and $\bar{\boldsymbol{\Sigma}}_i$ to be used must be cropped accordingly so that their dimensions are also $M(L+1) \times M(L+1)$.

The algorithm described in the previous paragraph will be referred to as simplified-LIKES (SLIKES). Although the result of executing this algorithm is an approximation of the ML estimate, the computational time can be reduced meaningfully depending on the value of L. It can be shown that the expressions for SLIKES are the same as those presented earlier if we replace $\boldsymbol{y}$ by $\hat{\boldsymbol{S}}^{1/2}$ and $\boldsymbol{y}^H\bar{\boldsymbol{\Sigma}}^{-1}\boldsymbol{y}$ by $\mathrm{Tr}\{\hat{\boldsymbol{S}}^{H/2}\bar{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{S}}^{1/2}\}$, where $\hat{\boldsymbol{S}}^{1/2}$ is an $M(L+1) \times M(L+1)$ matrix such that $\hat{\boldsymbol{S}}^{1/2}\hat{\boldsymbol{S}}^{H/2} = \hat{\boldsymbol{S}}$.

Unfortunately, the fact that the square root $\hat{\boldsymbol{S}}^{1/2}$ must exist requires $\hat{\boldsymbol{S}}$ to be positive semi-definite. This means that, according to what was explained in Section II-D-3, the unbiased estimate of the auto-correlation cannot be used to construct a modified SCM. Since using the biased estimate is expected to increase the bias of the estimation algorithm for finite data records, a problem that naturally arises is that of devising an algorithm that can use the unbiased estimate and is still ML-based. The next section provides such an algorithm based on the idea of relaxing the positivity constraints $\theta_i \geq 0$. It will also be seen that the resulting algorithm requires fewer samples to operate.

B. Unconstrained ML Estimation

Since $L(\boldsymbol{\theta})$ in (11) is twice differentiable, relaxing the constraint $\boldsymbol{\theta} \succeq \mathbf{0}$ enables us to find a minimum just by setting the gradient equal to zero. This idea is exploited in the current section to derive a simple approximation of the ML estimate.

By noting that $\bar{\boldsymbol{\Sigma}} = \sum_i \theta_i \bar{\boldsymbol{\Sigma}}_i$, it is possible to write the derivative of $L(\boldsymbol{\theta})$ with respect to θ_i as

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \theta_i} = \mathrm{Tr}(\bar{\boldsymbol{\Sigma}}^{-1}\bar{\boldsymbol{\Sigma}}_i) - \mathrm{Tr}(\bar{\boldsymbol{\Sigma}}^{-1}\bar{\boldsymbol{\Sigma}}_i\bar{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{R}}). \tag{15}$$

Due to the regularity of this function, the gradient vanishes when a minimum of $L(\boldsymbol{\theta})$ is attained, that is, the right hand side of (15) should be zero for all i. After some algebraic manipulations, this condition can be rewritten as

$$\mathrm{Tr}\left((\bar{\boldsymbol{\Sigma}}^{-1} - \bar{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{R}}\bar{\boldsymbol{\Sigma}}^{-1})\bar{\boldsymbol{\Sigma}}_i\right) = 0 \quad \forall i. \tag{16}$$

To the best of our knowledge, no analytical solution has been found yet for this non-linear system of equations. In fact, even for the simpler case where the matrices $\bar{\boldsymbol{\Sigma}}_i$ are Toeplitz, one must resort to numerical computations [31]. An example of an algorithm to solve (16) is the inverse iteration algorithm (IIA) proposed in [32], but, similarly to LIKES, the complexity is high, especially for large K. Moreover, the numerical stability is low, and intense efforts must be made in order to stay in the positive definite region (recall the discussion at the end of Section II-C).

Observe that the condition in (16) can be rewritten as

$$\mathrm{Tr}\left((\bar{\boldsymbol{\Sigma}} - \hat{\boldsymbol{R}})\bar{\boldsymbol{\Sigma}}^{-1}\bar{\boldsymbol{\Sigma}}_i\bar{\boldsymbol{\Sigma}}^{-1}\right) = 0 \quad \forall i. \tag{17}$$

A possibility to refine an estimate $\underline{\boldsymbol{\theta}}$ is to take $\bar{\boldsymbol{\Sigma}}^{-1} = \underline{\bar{\boldsymbol{\Sigma}}}^{-1}$ and compute $\boldsymbol{\theta}$ so that

$$\mathrm{Tr}\left((\bar{\boldsymbol{\Sigma}} - \hat{\boldsymbol{R}})\underline{\bar{\boldsymbol{\Sigma}}}^{-1}\bar{\boldsymbol{\Sigma}}_i\underline{\bar{\boldsymbol{\Sigma}}}^{-1}\right) = 0 \quad \forall i, \tag{18}$$

This expression can be rewritten as $\boldsymbol{A}\boldsymbol{\theta} = \boldsymbol{b}$, where

$$[\boldsymbol{A}]_{i,j} = \mathrm{Tr}(\underline{\bar{\boldsymbol{\Sigma}}}^{-1}\bar{\boldsymbol{\Sigma}}_i\underline{\bar{\boldsymbol{\Sigma}}}^{-1}\bar{\boldsymbol{\Sigma}}_j) \tag{19}$$

$$[\boldsymbol{b}]_i = \mathrm{Tr}(\underline{\bar{\boldsymbol{\Sigma}}}^{-1}\bar{\boldsymbol{\Sigma}}_i\underline{\bar{\boldsymbol{\Sigma}}}^{-1}\hat{\boldsymbol{R}}) \tag{20}$$

for $i, j = 0, \ldots, I-1$. The IIA algorithm from [32] iteratively uses this system of equations to refine the previous estimate $\underline{\boldsymbol{\theta}}$. The updating rule does not take directly the solution $\boldsymbol{\theta}_* = \boldsymbol{A}^{-1}\boldsymbol{b}$ since it can be out of the feasible region. Contrarily, the new iterate is taken to be $\underline{\boldsymbol{\theta}} + \alpha\boldsymbol{d}$, where $\boldsymbol{d} = \boldsymbol{\theta}_* - \underline{\boldsymbol{\theta}}$ and α is a factor that simultaneously ensures that $\underline{\boldsymbol{\theta}} + \alpha\boldsymbol{d}$ is in the feasible region of the unconstrained problem[7] and the cost function is smaller than that for $\underline{\boldsymbol{\theta}}$.

Although this algorithm converges to a local minimum, it is seen from (19) that the $MK/N \times MK/N$ matrix $\underline{\bar{\boldsymbol{\Sigma}}}$ must be inverted at each iteration and its determinant must be evaluated. This may be prohibitive even for moderate values of K so that a sensible approximation is to substitute $\hat{\boldsymbol{R}}$ by the averaged and cropped version $\hat{\boldsymbol{S}}$. The resulting algorithm will be thus referred to as the simplified IIA (SIIA) and, depending on the value of L chosen, may achieve considerable reductions in the computational cost. The resulting algorithm is summarized as Algorithm 2:

Algorithm 2 $\boldsymbol{\theta} = \mathbf{SIIA}(\hat{\boldsymbol{S}})$

Initialize $\boldsymbol{\theta} = [\theta_0, \ldots, \theta_{I-1}]^T$

while stopping_criterion == FALSE **do**
- $\bar{\boldsymbol{\Sigma}} \leftarrow \sum_i \theta_i \bar{\boldsymbol{\Sigma}}_i$
- $[\boldsymbol{A}]_{i,j} \leftarrow \mathrm{Tr}(\bar{\boldsymbol{\Sigma}}^{-1}\bar{\boldsymbol{\Sigma}}_i\bar{\boldsymbol{\Sigma}}^{-1}\bar{\boldsymbol{\Sigma}}_j)$
- $[\boldsymbol{b}]_i \leftarrow \mathrm{Tr}(\bar{\boldsymbol{\Sigma}}^{-1}\bar{\boldsymbol{\Sigma}}_i\bar{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{S}})$
- $\boldsymbol{\theta}_* \leftarrow \boldsymbol{A}^{-1}\boldsymbol{b}$
- $\boldsymbol{d} \leftarrow \boldsymbol{\theta}_* - \boldsymbol{\theta}$
- Choose α as explained above
- $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha\boldsymbol{d}$

end while

[7]Note that in the unconstrained problem it is necessary to check that $\bar{\boldsymbol{\Sigma}}$ is positive definite. That was not the case in the constrained case since the positive definiteness of $\bar{\boldsymbol{\Sigma}}$ is guaranteed by the fact that $\theta_i \geq 0 \forall i$ and the fact that the matrices in $\bar{\mathcal{S}}$ are positive definite.

As an initialization we propose to approximate $\bar{\boldsymbol{\Sigma}}^{-1} \approx \hat{\boldsymbol{S}}^{-1}$ in (17), so that the initial value for $\boldsymbol{\theta}$ can be chosen as the solution of the following linear system of equations:

$$\mathrm{Tr}\left((\bar{\boldsymbol{\Sigma}} - \hat{\boldsymbol{S}})\hat{\boldsymbol{S}}^{-1}\bar{\boldsymbol{\Sigma}}_i\hat{\boldsymbol{S}}^{-1}\right) = 0 \quad \forall i. \tag{21}$$

or, equivalently

$$\sum_j \theta_j \mathrm{Tr}(\hat{\boldsymbol{S}}^{-1}\bar{\boldsymbol{\Sigma}}_i\hat{\boldsymbol{S}}^{-1}\bar{\boldsymbol{\Sigma}}_j) = \mathrm{Tr}(\hat{\boldsymbol{S}}^{-1}\bar{\boldsymbol{\Sigma}}_i) \quad \forall i.$$

Finally note that, since this algorithm is designed to solve the unconstrained ML problem, we shall not expect that all the components of the resulting $\boldsymbol{\theta}$ are non-negative. A discussion on the implications of this fact is deferred to Section III-E. Note as well that even though the SIIA can be several times faster than the IIA, the computations involved can still be too burdensome in some applications. The next section provides two simpler algorithms relying on a least squares approximation.

C. Least Squares Estimation

We note that the condition in (16) can be rewritten as

$$\mathrm{Tr}\left(\bar{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{I} - \hat{\boldsymbol{R}}\bar{\boldsymbol{\Sigma}}^{-1})\bar{\boldsymbol{\Sigma}}_i\right) = 0 \quad \forall i,$$

which would immediately hold if one had $\hat{\boldsymbol{R}}\bar{\boldsymbol{\Sigma}}^{-1} = \boldsymbol{I}$. This suggests approximating $\bar{\boldsymbol{\Sigma}} \approx \hat{\boldsymbol{R}}$ somehow, for example in the least squares sense, i.e., we may try finding the vector $\boldsymbol{\theta}$ that minimizes the Frobenius distance $\|\hat{\boldsymbol{R}} - \bar{\boldsymbol{\Sigma}}\|_F^2$. However, the approximation $\hat{\boldsymbol{R}}\bar{\boldsymbol{\Sigma}}^{-1} \approx \boldsymbol{I}$ makes little sense in our case since $\hat{\boldsymbol{R}} = \boldsymbol{y}\boldsymbol{y}^H$ is rank one. Thus, it is not surprising that better results are obtained, in general, when we consider the averaged and cropped version of the SCM $\hat{\boldsymbol{S}}$ which, apart from being of higher rank, presents the same block structure as $\bar{\boldsymbol{\Sigma}}_i$. This clearly improves the consistency of the approximation and allows to consider only one representative from each block-diagonal. By defining

$$\hat{\boldsymbol{s}} = \mathrm{vec}\begin{bmatrix} L_L\hat{\boldsymbol{S}}^H[L] \\ \vdots \\ L_1\hat{\boldsymbol{S}}^H[1] \\ L_0\hat{\boldsymbol{S}}[0] \\ L_1\hat{\boldsymbol{S}}[1] \\ \vdots \\ L_L\hat{\boldsymbol{S}}[L] \end{bmatrix} \quad \text{and} \quad \boldsymbol{v}_i = \mathrm{vec}\begin{bmatrix} L_L\bar{\boldsymbol{\Sigma}}_i^H[L] \\ \vdots \\ L_1\bar{\boldsymbol{\Sigma}}_i^H[1] \\ L_0\bar{\boldsymbol{\Sigma}}_i[0] \\ L_1\bar{\boldsymbol{\Sigma}}_i[1] \\ \vdots \\ L_L\bar{\boldsymbol{\Sigma}}_i[L] \end{bmatrix},$$

where $L_k = L+1-k$ accounts for the number of times the k-th block is present in $\hat{\boldsymbol{S}}$ or $\bar{\boldsymbol{\Sigma}}_i$, the problem of minimizing $\|\hat{\boldsymbol{S}} - \bar{\boldsymbol{\Sigma}}\|_F^2$ can be stated as a least squares program:

$$\min_{\boldsymbol{\theta}} \|\hat{\boldsymbol{s}} - \boldsymbol{V}\boldsymbol{\theta}\|^2 \tag{22}$$

where $\boldsymbol{V} = [\boldsymbol{v}_0, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_{I-1}]$. Considering the blocks $\hat{\boldsymbol{S}}[n]$ along with their Hermitian versions $\hat{\boldsymbol{S}}[n]^H$ naturally imposes that $\theta_i \in \mathbb{R}$. Although this makes sense in theory, a more efficient choice when computing this solution is to separately consider real and imaginary parts.

For simplicity and efficiency, we may think of solving (22) without enforcing positivity on the θ_i's. In that case, the solution is given by $\boldsymbol{\theta}_{WLS} = \boldsymbol{V}^{\dagger}\boldsymbol{s}$, where WLS stands for weighted least squares.[8] The modifier *weighted* stems from the scaling factors L_k. Note that what this algorithm actually does is just to project the modified SCM $\hat{\boldsymbol{S}}$ onto the space spanned by the I basis matrices in $\bar{\boldsymbol{\Sigma}}_i$. On the other hand, if the constraints $\theta_i \geq 0$ are taken into account, the resulting algorithm is termed constrained WLS (CWLS).

Finally note that the approximation suggested in this section is equally well motivated even if the underlying distribution is not Gaussian: we may argue that taking $\bar{\boldsymbol{\Sigma}} \approx \hat{\boldsymbol{S}}$ is reasonable whenever $\hat{\boldsymbol{S}}$ is a consistent estimator for $\bar{\boldsymbol{\Sigma}}$.

D. Estimation Using SPICE

SPICE is a collection of algorithms for sparse parameter estimation which have in common that they all formulate the estimation problem as a program of the form

$$\text{minimize} \quad \mathrm{Tr}(\boldsymbol{Z}^H\bar{\boldsymbol{\Sigma}}^{-1}\boldsymbol{Z}) + \boldsymbol{\beta}^T\boldsymbol{\theta}, \tag{23}$$

where $\boldsymbol{Z}$ is either a matrix with the same dimensions as $\bar{\boldsymbol{\Sigma}}$ or a column vector and $\boldsymbol{\beta} \succ 0$ is a vector with the same dimension as $\boldsymbol{\theta}$. The program (23) is then solved as a second-order cone program (SOCP) or in an fixed-point fashion. It is derived in [18], [29], [30], [33], [34] under different conditions, showing that the problem is convex since it can be reformulated as either a SOCP or a semi-definite program (SDP). Although, to the best of our knowledge, the core SPICE iteration that solves (23) for the case where the basis matrices $\bar{\boldsymbol{\Sigma}}_i$ have a rank greater than one has not been derived in the literature, the extension is simple and it is summarized as Algorithm 3:

Algorithm 3 $\boldsymbol{\theta} = \mathbf{SPICE\_ITERATION}(\boldsymbol{\beta}, \boldsymbol{Z})$

Initialize $\boldsymbol{\theta} = [\theta_0, \ldots, \theta_{I-1}]^T$

while stopping_criterion == FALSE **do**

- $\bar{\boldsymbol{\Sigma}} \leftarrow \sum_i \theta_i\bar{\boldsymbol{\Sigma}}_i$
- $\boldsymbol{Q}_i \leftarrow \theta_i\bar{\boldsymbol{\Sigma}}_i^{H/2}\bar{\boldsymbol{\Sigma}}^{-1}\boldsymbol{Z}$, $i = 0, \ldots, I-1$
- $\alpha_i \leftarrow \mathrm{Tr}(\boldsymbol{Q}_i^H\boldsymbol{Q}_i)$, $i = 0, \ldots, I-1$
- $\theta_i \leftarrow \sqrt{\alpha_i/\beta_i}$, $i = 0, \ldots, I-1$

end while

Note that Algorithm 3 contains the operations to execute in the instruction of Algorithm 1 labeled as SPICE_ITERATION. In the case of LIKES we take $\boldsymbol{Z} = \boldsymbol{y}$ whereas in the case of SLIKES $\boldsymbol{Z}$ is set to $\hat{\boldsymbol{S}}^{1/2}$. In the remainder of this section we apply SPICE to our problem, first considering the raw SCM $\hat{\boldsymbol{R}} = \boldsymbol{y}\boldsymbol{y}^H$ and next the averaged and cropped version $\hat{\boldsymbol{S}}$. The first algorithm will be simply referred to as SPICE whereas the second one will be termed simplified SPICE (SSPICE).

In [18], [33], the so-called extended invariance principle [35] is invoked to simplify the problem of structured covariance estimation, achieving an estimate that asymptotically matches

[8]It was shown by simulations that, in most cases, WLS works better than the pure LS, which is the result of taking $L_k = 1 \forall k$. We omit the discussion of this algorithm due to space limitations.

the ML solution. It is also seen to work in the non-Gaussian case [29]. This estimate is the minimizer of the following criterion [33]

$$\left\|\bar{\boldsymbol{\Sigma}}^{-1/2}(\hat{\boldsymbol{R}}-\bar{\boldsymbol{\Sigma}})\right\|_F^2, \tag{24}$$

with $\hat{\boldsymbol{R}}=\boldsymbol{y}\boldsymbol{y}^H$, for the case where the observations are composed of exactly one realization $\boldsymbol{y}$, or [18]

$$\left\|\bar{\boldsymbol{\Sigma}}^{-1/2}(\hat{\boldsymbol{R}}-\bar{\boldsymbol{\Sigma}})\hat{\boldsymbol{R}}^{-1/2}\right\|_F^2, \tag{25}$$

for the case where several realizations are available so that $\hat{\boldsymbol{R}}$ is a full-rank SCM. In this second case, we will use the averaged and cropped $\hat{\boldsymbol{S}}$ in place of $\hat{\boldsymbol{R}}$ since only one single realization is available in our problem.

1) Standard SPICE: The criterion in expression (24) can be expanded as

$$\mathrm{Tr}(\bar{\boldsymbol{\Sigma}})+\|\boldsymbol{y}\|^2\cdot\boldsymbol{y}^H\bar{\boldsymbol{\Sigma}}^{-1}\bar{\boldsymbol{y}}-2\|\boldsymbol{y}\|^2 \tag{26}$$

so that the problem is to compute

$$\underset{\boldsymbol{\theta}\succeq 0}{\text{minimize}}\quad \boldsymbol{y}^H\bar{\boldsymbol{\Sigma}}^{-1}\boldsymbol{y}+\frac{\mathrm{Tr}(\bar{\boldsymbol{\Sigma}})}{\|\boldsymbol{y}\|^2}. \tag{27}$$

Clearly, (27) is a particular case of (23) where $\boldsymbol{Z}=\boldsymbol{y}$ and $\beta_i=\mathrm{Tr}(\bar{\boldsymbol{\Sigma}}_i)/\|\boldsymbol{y}\|^2$ and it can be solved using Algorithm 3.

2) Simplified SPICE: We propose in this section to use SPICE over an averaged and cropped SCM $\hat{\boldsymbol{S}}$. In this case, we prefer the criterion in (25) since its statistical motivation is stronger than that of the criterion in (24) (see [18]). Using simple algebra, it can be seen that the minimization of (25), with $\hat{\mathbf{R}}$ replaced by $\hat{\mathbf{S}}$, can be rewritten as the problem

$$\underset{\boldsymbol{\theta}\succeq 0}{\text{minimize}}\,\mathrm{Tr}\left(\hat{\boldsymbol{S}}^{\frac{H}{2}}\bar{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{S}}^{\frac{1}{2}}\right)+\mathrm{Tr}(\hat{\boldsymbol{S}}^{-1}\bar{\boldsymbol{\Sigma}}). \tag{28}$$

This is clearly a particular case of (23) with $\boldsymbol{Z}=\hat{\boldsymbol{S}}^{1/2}$ and $\beta_i=\mathrm{Tr}(\hat{\boldsymbol{S}}^{-1}\bar{\boldsymbol{\Sigma}}_i)$ so that it can be efficiently solved using Algorithm 3. We will refer to this estimator as simplified SPICE (SSPICE).

Although (standard) SPICE is an efficient method in the context of array processing and spectral analysis for line-spectrum/direction of arrival estimation [18], [29], [30], [34], the need to invert an $NK/M\times NK/M$ matrix at each iteration makes its application to spectrum sensing difficult since practical conditions involve low signal-to-noise ratios and, consequently, long observation times. For this reason, the size of $\boldsymbol{y}$, and therefore the size of $\bar{\boldsymbol{\Sigma}}$, may be prohibitively high for SPICE. For concreteness, and since we have observed in our experiments that the rank-1 criterion in (24) does not lead to a good estimation performance in the context of compressive WSS, we will only consider SSPICE in the rest of the paper. Note that this fact agrees with the statement given in [33] that (24) does not provide good estimates of the coefficients θ_i.

E. General Considerations

In this section, we have proposed several algorithms that rely on the averaged and cropped SCM. This idea potentially enables a huge computational cost reduction at the expense of obtaining an approximation of the ML solution in some cases, or an approximation of an asymptotic criterion such as the SPICE cost in other cases. Fortunately, as we will see in Section V, the performance degradation is small. On the other hand, the computational load of the approximated algorithms may be several orders of magnitude below that of the exact algorithms depending on the choice for L.

If the algorithms that provide unconstrained estimates are used, i.e. SIIA and WLS, some of the elements of the estimated $\boldsymbol{\theta}$ may be negative. In those cases, we may set the negative entries equal to zero in order to reconstruct the spectrum or compute the GLR statistic.

Finally, note that among all the algorithms employing the averaged and cropped SCM, some of them impose certain requirements over this matrix, namely SLIKES, SIIA and SSPICE. In particular, it can be easily seen that SLIKES requires the SCM to be positive semi-definite, whereas SIIA and SSPICE need it to be invertible and positive definite, respectively.

1) SCM Selection: According to what was discussed above and in Section II-D-3, we may establish a simple guideline to select a proper averaging method for the computation of $\hat{\boldsymbol{S}}$. The main decision criterion depends on whether negative eigenvalues are allowed or not. In particular, if $\hat{\boldsymbol{S}}$ is required to be positive definite or positive semi-definite, then the traditional unbiased estimate of the autocorrelation cannot be used since it does not guarantee that all the eigenvalues of $\hat{\boldsymbol{S}}$ are non-negative. A simple alternative is to employ the biased estimate of the autocorrelation, which, according to Theorem 1, guarantees that all the eigenvalues of $\hat{\boldsymbol{S}}$ are non-negative. The disadvantage of this method is, however, that the bias of the estimate will be, in general, increased. A different alternative is to apply the unbiased estimate but set the negative eigenvalues equal to zero. Unfortunately, this method has the disadvantage of reducing the rank of $\hat{\boldsymbol{S}}$ and cannot be used in those algorithms requiring the SCM to be non-singular.

2) Parameter Selection: The use of the algorithms above and, in general, any algorithm that uses a modified SCM requires a choice of the parameters M, N, K and L. We next provide a set of guidelines about how to accomplish that choice:

- M and N are hardware-related parameters that affect the acquisition stage. Since the influence of these parameters strongly determines the cost of the A2I, here we may consider them as given. For a more in-depth discussion on this topic see [28].
- K is a parameter that mainly affects the acquisition time, but does not meaningfully affect the complexity of the algorithm since its only influence is in the computation of $\hat{\boldsymbol{S}}$. In typical spectrum sensing applications, the acquisition time is determined by the target performance of the sensing algorithm; for example when both the probability of detection and false alarm are specified in a detection problem. In other situations, the goal may be to minimize the sensing time so that K may be chosen equal to its minimum given by the requirements of the algorithm used. For example, if $\hat{\boldsymbol{S}}$ needs to be full-rank, this minimum value is

$N[M(L+1)-L]$ for the biased estimate and NM for the unbiased estimate, as seen in Section II-D-3.

- Although the choice of L strongly determines the computational time of the algorithm used since the matrices to be manipulated are of size $M(L+1)\times M(L+1)$, the performance is not meaningfully affected for reasonable values of L since in typical spectrum sensing applications most relevant second-order information is concentrated around the main diagonal of the covariance matrices (see discussion about Fig. 1).
  There is a lower bound imposed by the linear independence of the matrices in $\bar{\mathcal{S}}$ and $\mathcal{S}$ (see expression (1)) which, of course, depends on I. On the other hand, an upper bound may be imposed by the requirements of Section II-D-3 if K is given, depending on the algorithm used. According to our experience, good values for L are those that make the product ML a few times higher than I. In any case, the best choice for L will depend on the matrices in $\bar{\mathcal{S}}$ and it is the subject of future work (see also [28]).

3) Convergence and Consistency: From an optimization point of view, all algorithms described above enjoy interesting convergence properties. For instance, LIKES guarantees convergence to a local minimum of the negative log-likelihood function since the cost is always decreased at each iteration [29]. A similar argument establishes local convergence for SLIKES. This property is also satisfied by the IIA, where each iteration provides an *improving direction* [32] and the stepsize should be selected so that the cost is decreased. Moreover, since the arguments in [32] assume a general positive definite SCM, it is clear that they remain valid if the modified SCM is used, thus showing local convergence for SIIA.

Regarding non-ML algorithms, WLS is a convex cost whose optimal point can be expressed in closed form, whereas CWLS is a convex cost which can be minimized using any conventional convex solver, so that both algorithms are globally convergent. The SPICE cost is also convex and the global convergence of the alternating algorithm is established in [33]. Global convergence for SSPICE follows from the arguments in [18].

In order to consider asymptotic statistical performance, let us assume, for simplicity, that $\boldsymbol{\theta}\succ\mathbf{0}$. In that case and according to the arguments in [19], the ML-based algorithms LIKES and IIA are expected to be asymptotically efficient and unbiased. Since the Cramér-Rao bound (CRB) is decreasing with the number of samples (See Section V for more information) this in turn means that these ML algorithms will be consistent [24].

Regarding the algorithms employing the modified SCM, that is, SLIKES, SIIA, WLS, CWLS and SSPICE, consistency follows from a different argument: the fact that the modified SCM is a consistent estimator for the corresponding part of the true covariance matrix means that $\boldsymbol{S}$ will converge in probability[9] to $\bar{\boldsymbol{\Sigma}}$. Intuitively, consistency follows by noting that all the criteria of these algorithms are minimized by making $\bar{\boldsymbol{\Sigma}}=\hat{\boldsymbol{S}}$ provided that $\hat{\boldsymbol{S}}$ is in the feasible set, i.e., provided that $\hat{\boldsymbol{S}}$ is in the span of $\bar{\mathcal{S}}$ with non-negative coefficients, but $\hat{\boldsymbol{S}}$ becomes closer and closer to this region as $K\to\infty$. A more rigorous proof of this effect is out of the scope of this paper.

[9]Formally, this means that $\mathbf{P}\{\|\hat{\boldsymbol{S}}-\bar{\boldsymbol{\Sigma}}\|>\alpha\}\to 0$ for all $\alpha>0$ as $K\to\infty$, where $\|\cdot\|$ denotes some norm, e.g., the Frobenius norm.

IV. Estimation in the Frequency Domain

It is interesting to see that the same reasoning used in [7] to motivate the use of least squares in the frequency domain also applies here just by replacing scalars by blocks. Indeed, expression (16) can be easily written in terms of PSDs by exploiting the asymptotic properties of Toeplitz covariance matrices that were already used in Section II-D-1. The idea is to diagonalize the cross-covariance matrices associated with the processes that are jointly stationary.

In order not to overload the notation, let us apply the superposition principle to compute the asymptotic approximation of $\bar{\boldsymbol{\Sigma}}$. First consider that all the entries in the vector $\boldsymbol{\theta}$ are set to zero except for the one corresponding to the i-th signal component $x_i[n]$. Later we just need to add together the contributions for all values of i to get the desired covariance.

If $\bar{\boldsymbol{\Sigma}}_i^{(m,m')}$ denotes the cross-covariance matrix of $y_m[k]$ and $y_{m'}[k]$, then it is easy to see that under our observation model

$$\bar{\boldsymbol{\Sigma}}_i^{(m,m')}=\left(\boldsymbol{I}_{\frac{K}{N}}\otimes\boldsymbol{\phi}_m^H\right)\boldsymbol{\Sigma}_i\left(\boldsymbol{I}_{\frac{K}{N}}\otimes\boldsymbol{\phi}_{m'}\right),$$

which is Toeplitz. This enables us to approximate $\bar{\boldsymbol{\Sigma}}_i^{(m,m')}\approx\boldsymbol{U}\bar{\boldsymbol{P}}_i^{(m,m')}\boldsymbol{U}^H$, where $\bar{\boldsymbol{P}}_i^{(m,m')}$ is a diagonal matrix containing the cross-PSD of the processes $y_m[k]$ and $y_{m'}[k]$ when only $x_i[n]$ is present at the input; and $\boldsymbol{U}$ is the unitary IDFT matrix. The covariance of the observations can now be composed by noting that $\bar{\boldsymbol{\Sigma}}_i=\sum_{m,m'}\bar{\boldsymbol{\Sigma}}_i^{(m,m')}\otimes\boldsymbol{E}_{m,m'}$, where $\boldsymbol{E}_{m,m'}$ is a matrix with a one in the position (m,m') and zeros elsewhere, so that $\bar{\boldsymbol{\Sigma}}_i$ can be approximated as $\bar{\boldsymbol{\Sigma}}_i\approx(\boldsymbol{U}\otimes\boldsymbol{I}_M)\bar{\boldsymbol{P}}_i(\boldsymbol{U}^H\otimes\boldsymbol{I}_M)$, where $\bar{\boldsymbol{P}}_i=\sum_{m,m'}\bar{\boldsymbol{P}}_i^{(m,m')}\otimes\boldsymbol{E}_{m,m'}$ is a block diagonal matrix containing the cross PSDs:

$$\bar{\boldsymbol{P}}_i=\begin{bmatrix}\bar{\boldsymbol{P}}_i[0] & & \dots & \mathbf{0}\\ & \bar{\boldsymbol{P}}_i[1] & & \vdots\\ \vdots & & \ddots & \\ \mathbf{0} & \dots & & \bar{\boldsymbol{P}}_i[K/N-1]\end{bmatrix}$$

Finally, considering all contributions for $i=0,1,\dots,I-1$ yields

$$\bar{\boldsymbol{\Sigma}}\approx(\boldsymbol{U}\otimes\boldsymbol{I}_M)\bar{\boldsymbol{P}}(\boldsymbol{U}^H\otimes\boldsymbol{I}_M),\tag{29}$$

where $\bar{\boldsymbol{P}}=\sum_i\theta_i\bar{\boldsymbol{P}}_i$. If we now substitute this in (16) we obtain the ML condition in the frequency domain:

$$\operatorname{Tr}\left((\bar{\boldsymbol{P}}^{-1}-\bar{\boldsymbol{P}}^{-1}\tilde{\boldsymbol{R}}\bar{\boldsymbol{P}}^{-1})\bar{\boldsymbol{P}}_i\right)=0\quad\forall i.\tag{30}$$

Here, we have defined $\tilde{\boldsymbol{R}}=(\boldsymbol{U}^H\otimes\boldsymbol{I}_M)\hat{\boldsymbol{R}}(\boldsymbol{U}\otimes\boldsymbol{I}_M)$ which is the sample estimate of the PSD. If the raw estimate $\hat{\boldsymbol{R}}=\boldsymbol{y}\boldsymbol{y}^H$ is used, then $\tilde{\boldsymbol{R}}=\tilde{\boldsymbol{y}}\tilde{\boldsymbol{y}}^H$, where $\tilde{\boldsymbol{y}}=(\boldsymbol{U}^H\otimes\boldsymbol{I}_M)\boldsymbol{y}$ is the vector with the DFT of the observations, is a generalized

periodogram. To see this, note that we can decompose $\tilde{\boldsymbol{y}}$ as $\tilde{\boldsymbol{y}} = [\tilde{\boldsymbol{y}}^T[0], \tilde{\boldsymbol{y}}^T[1], \ldots, \tilde{\boldsymbol{y}}^T[K/N-1]]^T$ where

$$\tilde{\boldsymbol{y}}[k] = \frac{1}{\sqrt{K/N}} \sum_n \boldsymbol{y}[n] e^{-j\frac{2\pi}{K/N}kn}$$

Solving the equations in (30) is conceptually the same as solving (16), although one of these conditions can be numerically more convenient than the other depending on the particular scenario. The IIA algorithm can be extended easily to the frequency domain. However, if we are not willing to accept the computational cost of finding the exact ML solution, or if the Gaussian assumption cannot be applied, we can follow the same guidelines as in Section III-C by rewriting (30) as

$$\mathrm{Tr}\left((\boldsymbol{I} - \bar{\boldsymbol{P}}^{-1}\tilde{\boldsymbol{R}})\bar{\boldsymbol{P}}^{-1}\bar{\boldsymbol{P}}_m\right) = 0 \quad \forall m, \tag{31}$$

which leads us to consider making $\bar{\boldsymbol{P}}$ and $\tilde{\boldsymbol{R}}$ close in the LS sense or, in other words, minimizing the metric $\|\bar{\boldsymbol{P}} - \tilde{\boldsymbol{R}}\|_F^2$. Since $\bar{\boldsymbol{P}}$ is block diagonal, this simplifies just to minimizing $\sum_k \|\bar{\boldsymbol{P}}[k] - \tilde{\boldsymbol{R}}[k]\|_F^2$, where $\tilde{\boldsymbol{R}}[k]$ represents the k-th block on the diagonal of $\tilde{\boldsymbol{R}}$.

This particular view of the problem is especially attractive for interpretations in terms of guard bands, like those in [7]. It can also be seen that the frequency criterion is equivalent to the LS methods in the time domain: just note that the Frobenius norm is invariant to unitary transformations so that the metric $\|\bar{\boldsymbol{P}} - \tilde{\boldsymbol{R}}\|_F^2$ is also

$$\left\|(\boldsymbol{U}\otimes\boldsymbol{I}_M)\bar{\boldsymbol{P}}(\boldsymbol{U}^H\otimes\boldsymbol{I}_M) - (\boldsymbol{U}\otimes\boldsymbol{I}_M)\tilde{\boldsymbol{R}}(\boldsymbol{U}^H\otimes\boldsymbol{I}_M)\right\|_F^2$$

which, in turn, equals $\|\bar{\boldsymbol{\Sigma}} - \hat{\boldsymbol{R}}\|_F^2$. Although the algorithms developed for the time domain in Section III could also be extended to the frequency domain, we do not consider them since the complexity is expected higher with respect to the time domain, and the performance is expected slightly worse due to the asymptotic approximation used in (29).

V. Simulations

In virtue of the discussion of Section III-E-3, it is expected that the algorithms presented in the previous sections perform asymptotically well. In particular, all the estimation algorithms are consistent [24] as $K \to \infty$. Unfortunately, a performance assessment when K takes finite values is not so immediate, especially for detection, where an analytical evaluation seems to be a formidable task. For this reason, we resort to Monte Carlo (MC) simulation in this section.

All the proposed algorithms use an averaged version of the SCM except for LIKES and IIA, which are sometimes presented for the sake of comparison. In particular, SLIKES and SSPICE are implemented here based on the biased SCM whereas the rest of the algorithms (SIIA, WLS and CWLS) operate on the unbiased estimate. Cropping is also implemented for these algorithms unless the parameter L takes on the maximum admissible value, i.e., that satisfying $N(L+1) = K$.

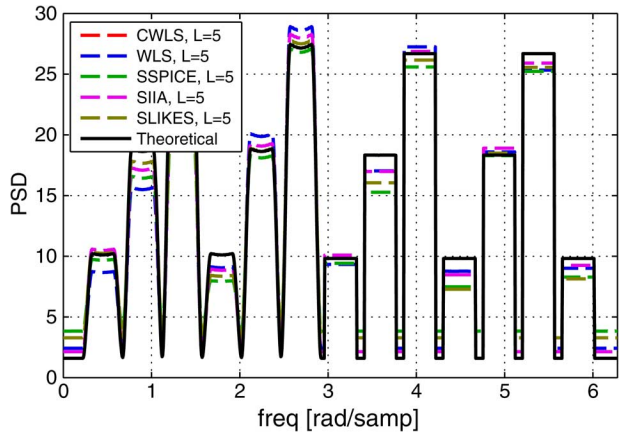


Fig. 2. PSD estimates for (from left to right) six QAM signals and seven OFDM signals. Note that many curves overlap since their estimates are close to each other. $K = 10240$, Gaussian A2I, $\sigma^2_{\mathrm{A2I}} = 1$, $N = 10$, $M = 5$.

The A2Is used are of two kinds. The first one is generated by drawing an independent zero-mean complex Gaussian random variable with variance σ^2_{A2I} for each of the MN components of $\boldsymbol{\Phi}$. This procedure is repeated at every MC iteration in order to average over this family of sampling matrices, thus making our results independent of the particular choice of the matrix. The A2I of the second kind is a multi-coset A2I with associated matrix

$$\boldsymbol{\phi} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{32}$$

Observe that summing all rows results in a minimal sparse ruler of length $N = 10$, which is important for identifiability reasons [9], [28].

The general setting is depicted in Fig. 2, where the true PSD is shown along with the reconstructed PSD, which corresponds to the frequency domain version of the covariance estimate $\hat{\Sigma} = \sum_i \hat{\theta}_i \boldsymbol{\Sigma}_i$, with $\hat{\theta}_i$ the outputs of the algorithms. In this case we illustrate the use of these algorithms in a realistic scenario. The sensed band contains six 4-QAM signals, seven OFDM signals and white noise. The power of the signals is given by $\theta_i = 3 \times (i \bmod 3 + 1)$, $i = 0, 1 \ldots, 12$, whereas the noise power is set to $\theta_{13} = 10$. Regarding the modulation parameters, a root raised cosine pulse with a roll-off factor of 0.3 is used in the QAM signal. The OFDM signal uses 512 subcarriers, although the last and the first 50 are set to zero in order to ease the design of the interpolation filter. Each subcarrier is also modulated by a 4-QAM constellation and a cyclic prefix with length 1/4 of the symbol period is used. We observe that although the distribution of these signals is not exactly Gaussian (especially in the case of the QAM signals) all the schemes are still able to estimate the power of each channel with acceptable accuracy.

However, in order not to confine ourselves to specific modulations, in the remaining simulations we generate the I signals $x_i[n]$ by passing white Gaussian noise of power θ_i through an energy-normalized prototype FIR filter with 31 coefficients and

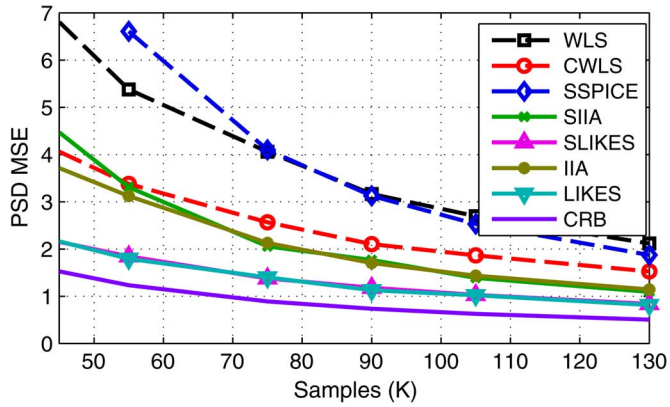


Fig. 3. Comparison of the estimation performance of the different algorithms presented in the paper. $\boldsymbol{\theta} = [4, 9, 1]^T$, Gaussian A2I, $\sigma^2_{\text{A2I}} = 1$, $N = 5$, $M = 2$, $L = 8$.

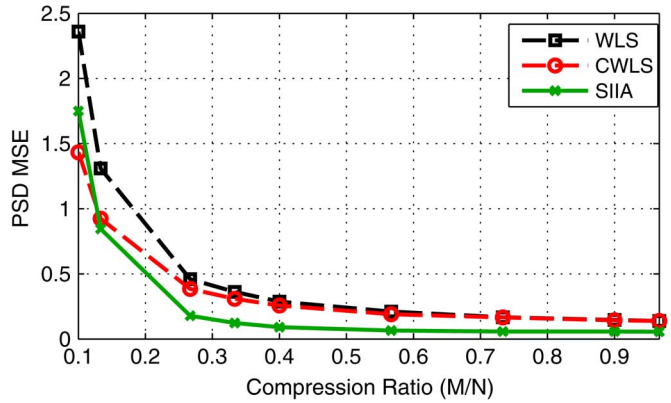


Fig. 4. A large portion of the samples can be discarded without a meaningful performance loss. $K = 900, \boldsymbol{\theta} = [4, 9, 1]^T$ Gaussian A2I, $\sigma^2_{\text{A2I}} = 1, N = 30$, $L = 15$.

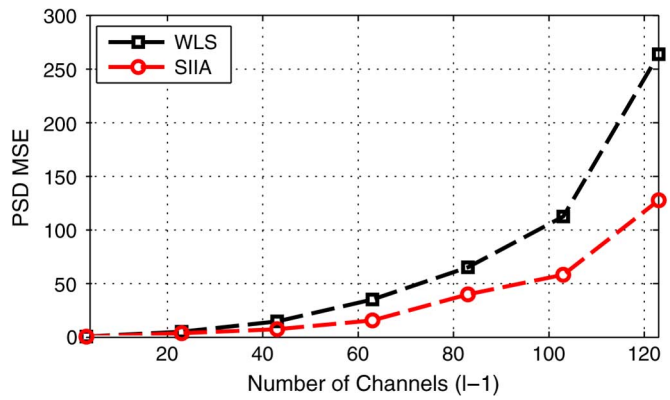


Fig. 5. Influence of the number of channels in the estimation performance. $\theta_0 = 1$ for $i = 1, 2 \ldots, I-2$ and $\theta_{I-1} = 10$. $K = 5120$, Gaussian A2I, $\sigma^2_{\text{A2I}} = 1$, $N = 10$, $M = 5$, $L = 10$.

passband bandwidth 0.4π rad/samp, except when many channels need to be considered, where the order is increased and the bandwidth reduced to avoid frequency overlap.[10] Every signal is then frequency shifted according to equally spaced carrier frequencies. The last signal $x_{I-1}[n]$ is in all cases white Gaussian noise.

A. Estimation Performance

The first experiment of this section illustrates the fact that the proposed algorithms are consistent, in the sense that the estimates of $\boldsymbol{\theta}$ converge to the true values as K becomes larger. In Fig. 3, the MSE is represented vs. K along with the CRB, which is obtained from combining the bound for $\boldsymbol{\theta}$ in [23], [36] with a linear transformation to obtain the bound in the PSD domain. See [19, Sec. 3.8] for more details about how to accomplish this computation. Note that $\boldsymbol{\theta}$ is constrained to satisfy $\boldsymbol{\theta} \succeq \mathbf{0}$, which means that the bound to be used is actually the constrained CRB. However, it is shown in [37] that this bound coincides with the unconstrained CRB at those points where $\boldsymbol{\theta} \succ \mathbf{0}$. For simplicity, in this experiment we only consider points satisfying this condition. The MSE is measured by summing the squares of the point-by-point difference between the reconstructed PSD and the theoretical one. The result is normalized by the number of points considered so that the MSE is measured in squared units of power per radian.

We also show in this figure, for comparison purposes, the PSD MSE of the constrained and unconstrained ML estimators obtained with the LIKES and IIA algorithms. Note that SSPICE is not defined for the leftmost point since K is too small for the biased SCM to be non-singular. We observe that SIIA and SLIKES achieve an estimation performance which is almost identical to the performance of IIA or LIKES, yet with a much lower computational complexity. The advantage of using a constrained estimate is also noticed.

The influence of the compression ratio on the MSE is investigated in Fig. 4. Since K is set to a fixed value, it is clear that a worse performance must be expected whenever M/N is smaller since fewer samples are processed, that is, the MSE is always decreasing in M/N. Interestingly, it is actually observed that only a small performance loss is entailed even when the sampling rate is considerably decreased. Although both SLIKES and SSPICE were omitted for simplicity, a similar behavior can be expected for these algorithms.

To close this section, we analyze in Fig. 5 the influence of the number of channels I in the estimation MSE. The variance vector is set to $\boldsymbol{\theta} = [1, 1, \ldots, 1, 10]^T$. As intuition predicts, the MSE increases with I since the number of parameters increases for a fixed number of observations [19].

B. Detection Performance

Now we compare the algorithms presented above in the detection setting. The interest here is to decide over the first component θ_0. According to what was explained in Section II-B, after estimating $\hat{\boldsymbol{\theta}}$ under both hypotheses, the results are substituted in (5) and $\mathcal{L}(\boldsymbol{y})$ is compared against a threshold η. If any algorithm returns a negative θ_i, it is set to zero in order to evaluate (5) since otherwise $\bar{\boldsymbol{\Sigma}}$ may not be positive definite. The results are not averaged over a family of matrices since the distribution of $\mathcal{L}(\boldsymbol{y})$ under $\mathcal{H}_0$ is required to remain the same from one MC iteration to another in order to adjust the threshold η. We thus fix the sampling matrix to that in (32).

A simple receiver operating characteristic (ROC) [20], i.e., the representation of the probability of detection (P_D) vs. the probability of false alarm (P_{FA}), is shown in Fig. 6, where we can observe that the probability of detection of LIKES, SLIKES,

[10]Note that the scheme is still able to work even if there is overlap.

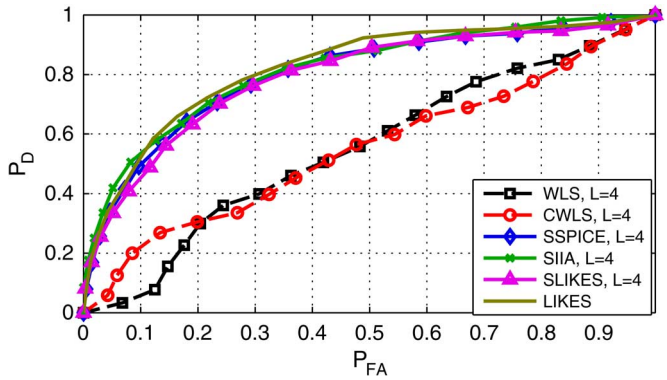


Fig. 6. ROC comparing the algorithms of this paper. $K = 300$, $\boldsymbol{\theta} = [0.2, 9, 1]^T$ Multi-Coset A2I, $N = 10$, $M = 5$.

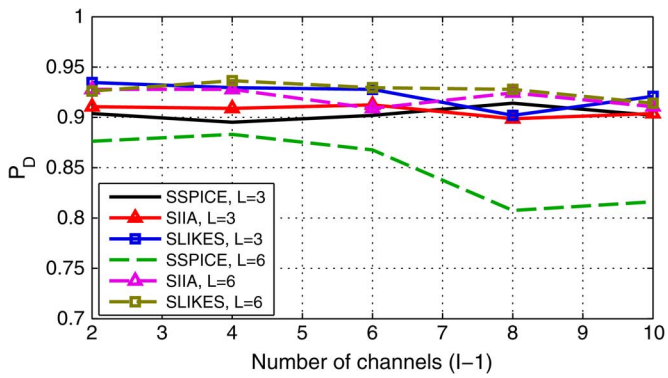


Fig. 7. Influence of the number of channels on the detection performance. $\theta_0 = 0.4$, $\theta_1 = 4$, $\theta_{I-1} = 10$ and $\theta_i = 0$ for $i = 2, 3, \ldots, I-2$. $K = 300$, Multi-Coset A2I, $N = 10$, $M = 5$.

SIIA and SSPICE is roughly the same. On the other hand, the LS criterion is seen not to result in good detection rules. In particular, we observe that the ROC curves corresponding to WLS and CWLS are not even concave, meaning that their performance can be improved by using randomized versions of these tests [21].

Finally, a comparison of the detection power for fixed false alarm rate $P_{FA} = 0.1$ is presented vs. the number of channels in Fig. 7. The LS detectors are omitted in favor of clarity since their detection performance was not seen to be good in the previous figure. It is observed that the influence of I in P_D is not as important as in the MSE: only small variations are noticed. It is also seen that this influence is more important if we increase the parameter L in SSPICE.

VI. Conclusions

The problem of wideband spectrum sensing with prior information was formulated for the case when a frequency band is observed through an A2I. For Gaussian signals, ML estimation was seen to result in computational intensive solutions, which motivated us to find low-complexity approximations. For non-Gaussian signals, the same principles were used to find alternative estimation rules. All of them stem from considering an averaged and cropped version of the SCM which is closely connected to the standard estimates of the autocorrelation for ergodic processes.

The proposed algorithms (SLIKES, SIIA, WLS, CWLS and SSPICE), some of which are based on adapting existing algorithms to operate over the modified SCM, are consistent in the number of samples and trade performance for complexity. They may be classified as constrained and unconstrained estimation procedures, the first ones yielding the best performance at the expense of a higher computational complexity. For finite data records, a comparison was carried out by means of Monte Carlo simulations, revealing interesting properties such as the fact that extremely low compression rates can be achieved.

Appendix A
Proof of Theorem 1

For convenience, let us define $r = \frac{K}{N}$, $\boldsymbol{Y} = [\boldsymbol{y}[0], \boldsymbol{y}[1], \ldots, \boldsymbol{y}[r-1]]$ and $\bar{\boldsymbol{Y}} = [\boldsymbol{Y}, \boldsymbol{0}_{M\times L}]$. Since the columns of $\boldsymbol{Y}$ are drawn from a continuous (non-degenerate) probabilistic model, they are linearly independent with probability one so that we have $\operatorname{rank}\{\boldsymbol{Y}\} = \min\{r, M\}$ in that case.

Let us also consider the group of $(r+L)\times(r+L)$ circular rotation matrices $\boldsymbol{R}_k$, where $\boldsymbol{R}_k$ is defined as the k-th circular shift of the columns of the identity matrix to the left. Except for the dimensions, the definition of $\boldsymbol{R}_k$ here is the same as that in Section II-D-1. Note that $\boldsymbol{R}_{-k} = \boldsymbol{R}_k^H$ and $\boldsymbol{R}_k\boldsymbol{R}_p = \boldsymbol{R}_{k+p}$.

With this notation, it is possible to rewrite $\hat{\boldsymbol{S}}[k]$ in a more convenient form as $\hat{\boldsymbol{S}}[k] = \alpha_k \bar{\boldsymbol{Y}}\boldsymbol{R}_k\bar{\boldsymbol{Y}}^H$, with $\alpha_k = 1/K_k$. This means that $\hat{\boldsymbol{S}}$ can in turn be written as $\hat{\boldsymbol{S}} = \bar{\boldsymbol{Y}}_d\boldsymbol{Z}\bar{\boldsymbol{Y}}_d^H$, where $\bar{\boldsymbol{Y}}_d = \boldsymbol{I}_{L+1}\otimes\bar{\boldsymbol{Y}}$ and

$$\boldsymbol{Z} = \begin{bmatrix} \alpha_0\boldsymbol{R}_0 & \alpha_1\boldsymbol{R}_{-1} & \cdots & \alpha_L\boldsymbol{R}_{-L} \\ \alpha_1\boldsymbol{R}_1 & \alpha_0\boldsymbol{R}_0 & \cdots & \alpha_{L-1}\boldsymbol{R}_{-L+1} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_L\boldsymbol{R}_L & \alpha_{L-1}\boldsymbol{R}_{L-1} & \cdots & \alpha_0\boldsymbol{R}_0 \end{bmatrix}. \quad (33)$$

Interestingly, $\boldsymbol{Z}$ is also $\boldsymbol{Z} = \bar{\boldsymbol{R}}_d(\boldsymbol{Z}_0\otimes\boldsymbol{I}_{r+L})\bar{\boldsymbol{R}}_d^H$, where

$$\boldsymbol{Z}_0 = \begin{bmatrix} \alpha_0 & \alpha_1 & \cdots & \alpha_L \\ \alpha_1 & \alpha_0 & \cdots & \alpha_{L-1} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_L & \alpha_{L-1} & \cdots & \alpha_0 \end{bmatrix} \quad (34)$$

and where $\bar{\boldsymbol{R}}_d = \operatorname{diag}\{\boldsymbol{R}_0, \boldsymbol{R}_1, \ldots, \boldsymbol{R}_L\}$ is a block diagonal matrix with the blocks $\boldsymbol{R}_k$ on its diagonal.

In the biased case, we can go a step further by factoring $\boldsymbol{Z}_0\otimes\boldsymbol{I}_{r+L}$ as $\boldsymbol{Z}_0\otimes\boldsymbol{I}_{r+L} = \alpha_0(\boldsymbol{1}_{L+1}\otimes\boldsymbol{I}_{r+L})(\boldsymbol{1}_{L+1}\otimes\boldsymbol{I}_{r+L})^H$ where $\boldsymbol{1}_n$ denotes the $n\times 1$ vector of all ones. Then, it is possible to write $\hat{\boldsymbol{S}} = \alpha_0\boldsymbol{A}_{\hat{\boldsymbol{S}}}\boldsymbol{A}_{\hat{\boldsymbol{S}}}^H$ by defining $\boldsymbol{A}_{\hat{\boldsymbol{S}}} = \bar{\boldsymbol{Y}}_d\bar{\boldsymbol{R}}_d(\boldsymbol{1}_{L+1}\otimes\boldsymbol{I}_{r+L})$. Noting that $\hat{\boldsymbol{S}}$ is positive semi-definite since $\alpha_0 > 0$ concludes the proof of the first part.

The same factorization also shows that [38, Chap. 2]

$$\begin{aligned}\operatorname{rank}\{\hat{\boldsymbol{S}}\} &= \operatorname{rank}\left\{\alpha_0\boldsymbol{A}_{\hat{\boldsymbol{S}}}\boldsymbol{A}_{\hat{\boldsymbol{S}}}^H\right\} \\ &= \operatorname{rank}\{\boldsymbol{A}_{\hat{\boldsymbol{S}}}\}\end{aligned}$$

Since $\boldsymbol{A}_{\hat{\boldsymbol{S}}}$ has the special form of

$$\boldsymbol{A}_{\hat{\boldsymbol{S}}} = \bar{\boldsymbol{Y}}_d\bar{\boldsymbol{R}}_d(\boldsymbol{1}_{L+1}\otimes\boldsymbol{I}_{r+L}) = \begin{bmatrix} \bar{\boldsymbol{Y}}\boldsymbol{R}_0 \\ \bar{\boldsymbol{Y}}\boldsymbol{R}_1 \\ \vdots \\ \bar{\boldsymbol{Y}}\boldsymbol{R}_L \end{bmatrix},$$

it is possible to compute the rank by counting the number of linearly independent columns. If $r \le M(L+1)$, where $M(L+1)$ is the number of rows, we obtain r columns from the non-null

columns of $\bar{\boldsymbol{Y}}\boldsymbol{R}_0$ and one extra column for each $\bar{\boldsymbol{Y}}\boldsymbol{R}_k$ with $k = 1, 2, \ldots, L$. Therefore there are $\min(r + L, M(L + 1))$ independent columns. If $r > M(L + 1)$, the number of independent columns is also $M(L + 1)$. Consequently, we have established that $\operatorname{rank}\hat{\boldsymbol{S}} = \min(r+L, M(L+1))$, which concludes the proof of the second part.

Proving the third part is a little more involved since the factorization of the biased estimate does not apply to its unbiased counterpart. We start by considering the case $r \leq M$ and then move to the case $r > M$.

Since all entries in $\boldsymbol{Y}$ are different from zero with probability one, there exists an $M \times M$ invertible matrix of elementary row operations $\boldsymbol{F}$ such that

$$\boldsymbol{F}\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{I}_r \\ \boldsymbol{0} \end{bmatrix}$$

or, alternatively, $\boldsymbol{F}\bar{\boldsymbol{Y}} = \boldsymbol{E}$, with

$$\boldsymbol{E} = \begin{bmatrix} \boldsymbol{I}_r & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}. \tag{35}$$

The block diagonal matrix $\boldsymbol{F}_d = \boldsymbol{I}_{L+1} \otimes \boldsymbol{F}$ is clearly invertible and, consequently,

$$\begin{aligned} \operatorname{rank}\{\hat{\boldsymbol{S}}\} &= \operatorname{rank}\left\{\boldsymbol{F}_d \hat{\boldsymbol{S}} \boldsymbol{F}_d^H\right\} \\ &= \operatorname{rank}\left\{\boldsymbol{E}_d \bar{\boldsymbol{R}}_d (\boldsymbol{Z}_0 \otimes \boldsymbol{I}_{r+L}) \bar{\boldsymbol{R}}_d^H \boldsymbol{E}_d^H\right\} \\ &= \operatorname{rank}\left\{\boldsymbol{E}_d \bar{\boldsymbol{R}}_d (\boldsymbol{Z}_0 \otimes \boldsymbol{I}_{r+L}) \bar{\boldsymbol{R}}_d^H\right\} \\ &= \operatorname{rank}\left\{\boldsymbol{E}_d \bar{\boldsymbol{R}}_d (\boldsymbol{Z}_0 \otimes \boldsymbol{I}_{r+L})\right\} \end{aligned} \tag{36}$$

where $\boldsymbol{E}_d = \boldsymbol{I}_{L+1} \otimes \boldsymbol{E}$. The third equality is shown in Appendix B, whereas the forth one is a consequence of the fact that $\bar{\boldsymbol{R}}_d$ is invertible. In order to prove the third part of the theorem, we note that in the unbiased case $\boldsymbol{Z}_0$ in (34) is full rank [38, Chap. 2], i.e., $\operatorname{rank}\{\boldsymbol{Z}_0\} = L + 1$. Consequently, $\operatorname{rank}\{\boldsymbol{Z}_0 \otimes \boldsymbol{I}_{r+L}\} = (L + 1)(r + L)$, which is exactly its dimension so that $\boldsymbol{Z}_0 \otimes \boldsymbol{I}_{r+L}$ is also invertible. As a result,

$$\operatorname{rank}\hat{\boldsymbol{S}} = \operatorname{rank}\{\boldsymbol{E}_d \bar{\boldsymbol{R}}_d\} = (L + 1)\operatorname{rank}\{\boldsymbol{E}\} = (L + 1) \cdot r$$

This concludes the proof for the case $r \leq M$. In particular, for $r = M$ we have that $\hat{\boldsymbol{S}}$ is maximum rank. Thus, for $r > M$ it is easy to see that the rank of $\hat{\boldsymbol{S}}$ will also be maximum with probability one, since the resulting matrix is obtained by adding further random contributions to the averaged SCM that only considers the first r vectors.

Appendix B
Proof of Expression (36)

The purpose of this section is to show that

$$\operatorname{rank}\left\{\boldsymbol{E}_d \boldsymbol{Z} \boldsymbol{E}_d^H\right\} = \operatorname{rank}\{\boldsymbol{E}_d \boldsymbol{Z}\}. \tag{37}$$

Although in Appendix A we made use of this expression from left to right, in this section we will proceed from right to left. This approach follows from noting that left-multiplying $\boldsymbol{Z}$ by $\boldsymbol{E}_d$ amounts to a row selection where we take the first r rows of every block of $L + r$ rows. The rows that are not selected are set to zero but, as far as rank is concerned, this is equivalent to removing those rows. Right multiplying by $\boldsymbol{E}_d^H$ performs the analogous operation with the columns. Although the rank of $\boldsymbol{E}_d\boldsymbol{Z}$ can be less than or equal to the rank of $\boldsymbol{Z}$, we will see that once we have left-multiplied by $\boldsymbol{E}_d$, the right multiplication by $\boldsymbol{E}_d^H$ entails no further rank reduction.

In other words, we must show that the columns of $\boldsymbol{E}_d\boldsymbol{Z}$ that $\boldsymbol{E}_d^H$ sets to zero in $\boldsymbol{E}_d\boldsymbol{Z}\boldsymbol{E}_d^H$ are linearly dependent on the columns that are not set to zero. To do so, we must examine the structure of $\boldsymbol{Z}$. From (33) we observe that $\boldsymbol{Z}$ is a sparse matrix since the matrices $\alpha_k \boldsymbol{R}_k$ have exactly one non-null element in each row and column. As it is seen next, to establish (37) it suffices to analyze the position of the non-null elements of $\boldsymbol{Z}$. To simplify the explanation, let us define $\bar{\boldsymbol{Z}}$ as a matrix which is zero where $\boldsymbol{Z}$ is zero and one where $\boldsymbol{Z}$ is different from zero.

Let $r = \frac{K}{N}$ and let $\boldsymbol{p}_k, k = 0, 1, \ldots, (L+1)(r+L)-1$, denote the k-th column of $\boldsymbol{Z}$. We define $\mathcal{P}_k \subset \{0, 1, \ldots, (L + 1)(r + L) - 1\}$ as the set with the positions of the non-null elements of $\boldsymbol{p}_k$. It is easy to see that all the columns of $\bar{\boldsymbol{Z}}$ are present in the first block-column since the rest of the block-columns are just cyclic shifts of the first one. In particular we have that, for $k = 0, 1, \ldots, s - 1$,

$$\mathcal{P}_k = \mathcal{P}_{s+(k+1)_s} = \mathcal{P}_{2s+(k+2)_s} = \ldots = \mathcal{P}_{Ls+(k+L)_s} \tag{38}$$

where $s = r + L$ and $(\cdot)_s$ means remainder of integer division by s. Clearly, each one of these sets has $L + 1$ elements, those being

$$\begin{aligned} \mathcal{P}_k &= \{k, s + (k + 1)_s, 2s + (k + 2)_s, \ldots, Ls + (k + L)_s\} \\ &= \{ns + (k + n)_s, n = 0, 1, \ldots, L\} \end{aligned} \tag{39}$$

for $k = 0, 1, \ldots, s - 1$. It is not surprising that the elements in $\mathcal{P}_k$ are the same as those in the subscripts of (38) since $\bar{\boldsymbol{Z}}$ is symmetric.

Now denote as $\bar{\boldsymbol{p}}_k, k = 0, 1, \ldots, (L + 1)(r + L) - 1$, the k-th column in $\boldsymbol{E}_d\boldsymbol{Z}$. As discussed above, the left multiplication by this matrix amounts to setting to zero the rows of $\boldsymbol{Z}$ whose indices are not present in the set (see (35))

$$\mathcal{I} = \{n(L + r) + k : n = 0, 1, \ldots, L; k = 0, 1, \ldots, r - 1\}.$$

In other words, only the rows with indices in $\mathcal{I}$ are respected. Thus, if $\bar{\mathcal{P}}_k$ denotes the set of non-null indices in $\bar{\boldsymbol{p}}_k$, it is clear that $\bar{\mathcal{P}}_k = \mathcal{P}_k \cap \mathcal{I}$ and also that the same relationship in (38) still holds if we replace $\mathcal{P}$ by $\bar{\mathcal{P}}$, i.e.,

$$\bar{\mathcal{P}}_k = \bar{\mathcal{P}}_{s+(k+1)_s} = \bar{\mathcal{P}}_{2s+(k+2)_s} = \ldots = \bar{\mathcal{P}}_{Ls+(k+L)_s}. \tag{40}$$

This expression allows us to arrange the columns of $\boldsymbol{E}_d\boldsymbol{Z}$ in s equivalence classes. We say that the column $\bar{\boldsymbol{p}}_q$ is in the class k if q is contained in the set

$$\begin{aligned} \mathcal{S}_k &= \{k, s + (k + 1)_s, 2s + (k + 2)_s, \ldots, Ls + (k + L)_s\} \\ &= \{ns + (k + n)_s, n = 0, 1, \ldots, L\} \end{aligned} \tag{41}$$

for $k = 0, 1, \ldots, s - 1$. Note that $\bar{\mathcal{P}}_k \cap \bar{\mathcal{P}}_{k'} = \varnothing$ for $k \neq k'$ and that columns in class k are linearly independent from those in class k'. Also, if $|\mathcal{A}|$ denotes the number of elements in the set $\mathcal{A}$, then a collection of $|\bar{\mathcal{P}}_k|$ or fewer columns of class k is necessarily independent, since the coefficients of the non-null entries in $\bar{\boldsymbol{p}}_k$ are taken from the matrix $\boldsymbol{Z}_0$. Conversely, a

collection of more than $|\bar{\mathcal{P}}_k|$ elements of class k is necessarily a dependent set of vectors.

With these observations in mind it is clear that the right-multiplication of $\boldsymbol{E}_d\boldsymbol{Z}$ by $\boldsymbol{E}_d^H$ is not going to change the rank provided that this column selection respects at least $|\bar{\mathcal{P}}_k|$ elements of class k, for all k. The number of elements surviving in class k is clearly given by $|\mathcal{S}_k \cap \mathcal{I}|$. However, as we observe from (39) and (41), $\mathcal{S}_k = \mathcal{P}_k$ so that $|\mathcal{S}_k \cap \mathcal{I}| = |\mathcal{P}_k \cap \mathcal{I}| = |\bar{\mathcal{P}}_k|$, which is exactly the minimum number of elements needed from each class.

REFERENCES

[1] E. Axell, G. Leus, E. G. Larsson, and H. V. Poor, "Spectrum sensing for cognitive radio: State-of-the-art and recent advances," *IEEE Signal Process. Mag.*, vol. 29, no. 3, pp. 101–116, May 2012.

[2] Q. Zhao and B. M. Sadler, "A survey of dynamic spectrum access," *IEEE Signal Process. Mag.*, vol. 24, no. 3, pp. 79–89, 2007.

[3] J. Mitola, III and G. Q. Maguire, Jr, "Cognitive radio: Making software radios more personal," *IEEE Pers. Commun.*, vol. 6, no. 4, pp. 13–18, 1999.

[4] Z. Quan, S. Cui, A. H. Sayed, and H. V. Poor, "Optimal multiband joint detection for spectrum sensing in cognitive radio networks," *IEEE Trans. Signal Process.*, vol. 57, no. 3, pp. 1128–1140, 2009.

[5] R. Tandra and A. Sahai, "SNR walls for signal detection," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 1, pp. 4–17, 2008.

[6] A. Taherpour, S. Gazor, and M. Nasiri-Kenari, "Wideband spectrum sensing in unknown white Gaussian noise," *IET Commun.*, vol. 2, no. 6, pp. 763–771, Jul. 2008.

[7] G. Vázquez-Vilar and R. López-Valcarce, "Spectrum sensing exploiting guard bands and weak channels," *IEEE Trans. Signal Process.*, vol. 59, no. 12, pp. 6045–6057, 2011.

[8] M. Mishali and Y. C. Eldar, "Blind multiband signal reconstruction: Compressed sensing for analog signals," *IEEE Trans. Signal Process.*, vol. 57, no. 3, pp. 993–1009, Mar. 2009.

[9] D. D. Ariananda and G. Leus, "Compressive wideband power spectrum estimation," *IEEE Trans. Signal Process.*, vol. 60, no. 9, pp. 4775–4789, 2012.

[10] M. A. Lexa, M. E. Davies, J. S. Thompson, and J. Nikolic, "Compressive power spectral density estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2011, pp. 3884–3887.

[11] P. Pal and P. P. Vaidyanathan, "Coprime sampling and the MUSIC algorithm," in *Proc. IEEE Digit. Signal Process. Workshop Signal Process. Educ. Workshop (DSP/SPE)*, Jan. 2011, pp. 289–294.

[12] 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.

[13] O. Mehanna and N. Sidiropoulos, "Frugal sensing: Wideband power spectrum sensing from few bits," *IEEE Trans. Signal Process.*, vol. 61, no. 10, pp. 2693–2703, 2013.

[14] M. Wakin, S. Becker, E. Nakamura, M. Grant, E. Sovero, D. Ching, Y. Juhwan, J. Romberg, A. Emami-Neyestanak, and E. Candès, "A nonuniform sampler for wideband spectrally-sparse environments," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 2, no. 3, pp. 516–529, 2012.

[15] M. Mishali and Y. C. Eldar, "Sub-nyquist sampling: Bridging theory and practice," *IEEE Signal Process. Mag.*, vol. 28, no. 6, pp. 98–124, 2011.

[16] J. N. Laska, S. Kirolos, M. F. Duarte, T. S. Ragheb, R. G. Baraniuk, and Y. Massoud, "Theory and implementation of an analog-to-information converter using random demodulation," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2007, pp. 1959–1962.

[17] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*. New York, NY, USA: Wiley, 1958, vol. 2.

[18] 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, Feb. 2011.

[19] S. M. Kay, *Fundamentals of Statistical Signal Process*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993, vol. I, Estimation Theory.

[20] S. M. Kay, *Fundamentals of Statistical Signal Processing*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1998, vol. II, Detection Theory.

[21] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*. New York, NY, USA: Springer, 2005.

[22] R. M. Gray, *Toeplitz and Circulant Matrices: A Review*. Hanover, MA, USA: Now Publishers, 2006.

[23] P. Stoica and R. L. Moses, *Spectral Analysis of Signals*. Englewood Cliffs, NJ, USA: Pearson/Prentice Hall, 2005.

[24] E. L. Lehmann and G. Casella, *Theory of Point Estimation*. New York, NY, USA: Springer, 1998, vol. 31.

[25] Y. Zeng and Y. C. Liang, "Eigenvalue-based spectrum sensing algorithms for cognitive radio," *IEEE Trans. Commun.*, vol. 57, no. 6, pp. 1784–1793, 2009.

[26] Y. Zeng and Y. C. Liang, "Spectrum-sensing algorithms for cognitive radio based on statistical covariances," *IEEE Trans. Veh. Technol.*, vol. 58, no. 4, pp. 1804–1815, 2009.

[27] P. Billingsley, *Probability and Measure*. New York, NY, USA: Wiley, 1995.

[28] D. Romero and G. Leus, "Compressive covariance sampling," in *Proc. Inf. Theory Appl. Workshop*, Feb. 2013, pp. 1–8.

[29] P. Stoica and P. Babu, "SPICE and LIKES: Two hyperparameter-free methods for sparse-parameter estimation," *Signal Process.*, vol. 92, no. 7, pp. 1580–1590, 2012.

[30] P. Babu and P. Stoica, "Sparse spectral-line estimation for nonuniformly sampled multivariate time series: SPICE, LIKES and MSBL," in *Proc. 20th Eur. Signal Process. Conf. (EUSIPCO)*, 2012, pp. 445–449.

[31] D. R. Fuhrmann, "Application of Toeplitz covariance estimation to adaptive beamforming and detection," *IEEE Trans. Signal Process.*, vol. 39, no. 10, pp. 2194–2198, Oct. 1991.

[32] J. P. Burg, D. G. Luenberger, and D. L. Wenger, "Estimation of structured covariance matrices," *Proc. IEEE*, vol. 70, no. 9, pp. 963–974, Sep. 1982.

[33] P. Stoica, P. Babu, and J. Li, "New method of sparse parameter estimation in separable models and its use for spectral analysis of irregularly sampled data," *IEEE Trans. Signal Process.*, vol. 59, no. 1, pp. 35–47, Jan. 2011.

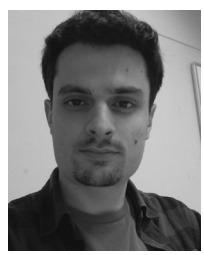
[34] P. Stoica, P. Babu, and J. Li, "A sparse covariance-based method for direction of arrival estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2011, pp. 2844–2847.

[35] 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, 1998.

[36] P. Stoica and P. Babu, "Maximum-likelihood nonparametric estimation of smooth spectra from irregularly sampled data," *IEEE Trans. Signal Process.*, vol. 59, no. 12, pp. 5746–5758, Dec. 2011.

[37] J. D. Gorman and A. O. Hero, "Lower bounds for parametric estimation with constraints," *IEEE Trans. Inf. Theory*, vol. 36, no. 6, pp. 1285–1301, 1990.

[38] D. S. Bernstein, *Matrix Mathematics: Theory, Facts, Formulas*. Princeton, NJ, USA: Princeton Univ. Press, 2011.



Daniel Romero (S'11) received the telecommunication engineering and the M.S. degrees from University of Vigo (Spain) in 2009 and 2011, respectively. He is currently working towards the Ph.D. degree at the same university. In 2012, he visited the Circuits and Systems group at Delft University of Technology. His main research interests are compressed sensing, signal processing for communications and statistical signal processing, in particular detection theory for dynamic spectrum access.

Geert Leus (M'01–SM'05–F'12) received the electrical engineering degree and the PhD degree in applied sciences from the Katholieke Universiteit Leuven, Belgium, in June 1996 and May 2000, respectively. Currently, Geert Leus is an "Antoni van Leeuwenhoek" Full Professor at the Faculty of Electrical Engineering, Mathematics and Computer Science of the Delft University of Technology, The Netherlands. His research interests are in the area of signal processing for communications. Geert Leus received a 2002 IEEE Signal Processing Society Young Author Best Paper Award and a 2005 IEEE Signal Processing Society Best Paper Award. He is a Fellow of the IEEE. Geert Leus was the Chair of the IEEE Signal Processing for Communications and Networking Technical Committee, and an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE SIGNAL PROCESSING LETTERS, and the EURASIP Journal on Advances in Signal Processing. Currently, he is a member of the IEEE Sensor Array and Multichannel Technical Committee and serves as the Editor in Chief of the EURASIP Journal on Advances in Signal Processing.