Complex-valued Signal Processing: The Proper Way to Deal with Impropriety

Tülay Adalı, *Fellow, IEEE*
Department of Computer Science and Electrical Engineering
University of Maryland, Baltimore County
Baltimore, MD 21250, USA
E-mail: adali@umbc.edu

Peter J. Schreier, *Senior Member, IEEE*
Signal and System Theory Group
Universität Paderborn
33098 Paderborn, Germany
E-mail: peter.schreier@upb.de

Louis L. Scharf, *Life Fellow, IEEE*
Department of Electrical and Computer Engineering
Colorado State University
Ft. Collins, CO 80523, USA
E-mail: scharf@engr.colostate.edu

Abstract— Complex-valued signals occur in many areas of science and engineering and are thus of fundamental interest. In the past, it has often been assumed, usually implicitly, that complex random signals are proper or circular. A proper complex random variable is uncorrelated with its complex conjugate, and a circular complex random variable has a probability distribution that is invariant under rotation in the complex plane. While these assumptions are convenient because they simplify computations, there are many cases where proper and circular random signals are very poor models of the underlying physics. When taking impropriety and noncircularity into account, the right type of processing can provide significant performance gains. There are two key ingredients in the statistical signal processing of complex-valued data: (1) utilizing the complete statistical characterization of complex-valued random signals; and (2) the optimization of real-valued cost functions with respect to complex parameters. In this overview article, we review the necessary tools, among which are widely linear transformations, augmented statistical descriptions, and Wirtinger calculus. We also present some selected recent developments in the field of complex-valued signal processing, addressing the topics of model selection, filtering, and source separation.

Index Terms— Improper, noncircular, widely linear, Wirtinger calculus, CR calculus, model selection, estimation, independent component analysis.

I. Introduction

Complex-valued signals arise in many areas of science and engineering, such as communications, electromagnetics, optics, and acoustics, and are thus of fundamental interest. In the past, it has often been assumed—usually implicitly—that complex random signals are *proper* or *circular*. A *proper* complex random variable is uncorrelated with its complex conjugate, and a *circular* complex random variable has a probability distribution that is invariant under rotation in the complex plane. These assumptions are convenient because they simplify computations and, in many respects, make complex random signals look and behave like real random signals. While these assumptions can often be justified, there are also many cases where proper and circular random signals are very poor models of the underlying physics. This fact has been known and appreciated by oceanographers since the early 1970s [80], but it has only more recently started to influence the thinking of the signal processing community.

In the last two decades, there have been important advances in this area that show impropriety and noncircularity as an important characteristic of many signals of practical interest. When taking impropriety and noncircularity into account, the right type of processing can provide significant performance gains. For instance, in mobile multiuser communications, it can enable an improved tradeoff between spectral efficiency and power consumption. Important examples of digital modulation schemes that produce improper complex baseband signals are: Binary Phase Shift Keying (BPSK), Pulse Amplitude Modulation (PAM), Gaussian Minimum Shift Keying (GMSK), Offset Quaternary Phase Shift Keying (OQPSK), and *baseband* (but not passband) Orthogonal Frequency Division Multiplexing (OFDM), which is commonly called Discrete Multitone (DMT) [119]. A small sample of papers exploiting impropriety of these signals is [140], [45], [123], [62], [46], [85], [83], [138], [22], [57], [79], [31], [78], and [23]. Improper baseband communication signals can also arise due to imbalance between their in-phase and quadrature (I/Q) components. I/Q imbalance degrades the signal-to-noise-ratio and thus bit error rate performance. Some papers proposing ways of compensating I/Q imbalance in different types of communication systems include [15], [110], and [142]. Techniques for wide-band system identification when the system (*e.g.*, a wide-band wireless communication channel) is not rotationally invariant are presented in [81] and [82].

In array processing, a typical goal is to estimate the direction of arrival (DOA) of one or more signals-of-interest impinging on a sensor array. If the signals-of-interest or the interference are improper (as they would be if they originated, *e.g.*, from a BPSK transmitter), this can be exploited to achieve higher DOA resolution or higher Signal-to-Interference-plus-Noise-Ratio (SINR). Some papers addressing the adaptation of array processing algorithms to improper signals include [25], [77], [49], [36], [26], and [107].

Data-driven methods for signal processing, in particular latent variable analysis (LVA), is another area where exploiting impropriety and noncircularity has led to important advances. In LVA, much interest has centered around independent component analysis (ICA), where the multivariate data are decomposed into additive components that are as independent as possible. It has been shown that, under certain conditions, ICA can be achieved by exploiting impropriety [63], [41]. Significant performance gains are noted when algorithms explicitly take the noncircular nature of the data into account [7], [96], [87], [88], [69], [141]. Among the many applications of ICA, medical data analysis and communications have been two of the most active. In [141], noncircularity is exploited for feature extraction in electrocardiograms (ECGs). In [67], [68], noncircularity is used in the analysis of functional magnetic resonance imaging (fMRI) data, leading to improved estimation of neural activity.

There are two key ingredients in the statistical signal processing of complex-valued data: (1) utilizing the complete statistical characterization of complex-valued random signals; and (2) the optimization of real-valued cost functions with respect to complex parameters. With respect to (1), Brown and Crane [20] in 1969 were among the first to consider the complementary correlation, which is the correlation of a complex signal with its complex conjugate. They also introduced linear-conjugate linear (or widely linear) transformations that allow access to the information contained in this correlation. At first, the impact of this paper was limited. It took until the 1990s before a string of papers showed renewed interest in this topic. Many were from the French school of signal processing. Among the important early contributions were the fundamental studies by Comon, Duvaut, Amblard, and Lacoume [10], [32], [39], [60], the derivation of the widely linear minimum mean-squared error filter by Picinbono and Chevalier [103], the work on higher-order statistics by Lacoume, Gaeta, and Amblard [11], [12], [43], [61], [76], [109], the introduction of the noncircular Gaussian distribution by van den Bos [129], and the derivation of the optimum Wiener filter for cyclostationary complex processes by Gardner [44]. We should also mention the more mathematically oriented work by Vakhania and Kandelaki [125] on random vectors with values in complex Hilbert spaces.

The results concerning the optimization of real-valued functions with complex arguments are much older still, although they have gone largely unnoticed by the engineering community. It was already in 1927 that Wirtinger proposed generalized derivatives of non-holomorphic (including real-valued) functions with respect to complex arguments [137]. However, differentiating such functions separately with respect to real and imaginary parts—and using varying definitions for the derivatives—has been the common practice. This approach is tedious, and in an attempt to make the terms and analysis more manageable, simplifying assumptions are usually introduced. A common such assumption is circularity. Wirtinger calculus (also called the $\mathbb{CR}$-calculus [59]) presents an elegant alternative, which allows keeping all computations in the complex domain. In the engineering community, Wirtinger calculus was rediscovered by Brandwood in 1983 [19], and then further developed by van den Bos, albeit for gradient and Hessian formulations in $\mathbb{C}^{2N}$ [126], [128]. The extensions of gradient and Hessian expressions to $\mathbb{C}^N$ are more recent [59], [64], [65].

Our goal in this article is to review all these fundamental results, and to present some selected recent developments in the field of complex-valued signal processing. Naturally, in a field as wide as this, one faces the difficult decision of what topics to cover. We have decided to focus on three key problems, namely model selection, filtering, and source separation. Some of these results have been drawn from the first author's chapter in the edited book [5], and the second and third authors' book [115], which provides a comprehensive review of this field. We thank Wiley Interscience and Cambridge University Press for their kind permission to use some material from these books. Another recent book on this topic is [75], which has a focus on neural networks.

We also provide some critical perspectives in this paper. As is often the case when new tools are introduced, there may also be some over-excitement and abuse. We comment on several points: There are certain instances where the use of improper/noncircular models is simply unnecessary and will not improve performance. The fact that a signal is improper

or noncircular does not guarantee that this can be exploited to improve a metric such as mean-squared error. Moreover, employing improper/noncircular models can even be disadvantageous because they are more complex than proper/circular models. Performance should be defined in a broader sense that includes not only optimality with respect to a selected metric, but also other properties of the algorithm, such as robustness and convergence speed. So an improvement in mean-squared error performance may come at the price of less robust and more slowly converging iterative algorithms. Therefore, the question of whether to use proper/circular or improper/noncircular models requires careful deliberation. We show that circular models may be preferable when the degree of impropriety/noncircularity is small, the number of samples is small, or the signal-to-noise ratio is low. We provide a number of concrete examples that clarify these points.

In terms of terminology, a general question that has divided researchers is how to extend definitions from the real to the complex case. For example, since the complementary covariance has traditionally been ignored, definitions of uncorrelatedness and orthogonality are given new names when the complementary covariance is taken into account. This practice leads to some counterintuitive results such as "two uncorrelated Gaussian random vectors need not be independent" (because their complementary covariance matrix need not be diagonal). We would like to avoid such displeasing results, and therefore always adhere to the following general principle: *Definitions and conditions derived for the real and complex domains should be equivalent.* This means that complementary covariances must be considered if they are nonzero.

The rest of the paper is organized as follows: In the next section, we provide a review of the tools needed for (1) utilizing the complete statistical characterization of complex-valued random signals and (2) the optimization of real-valued cost functions with respect to complex parameters. These are widely linear transformations, augmented statistical descriptions, and Wirtinger calculus. Section III addresses the question of how to detect whether a given signal is circular or noncircular, and more generally, how to detect the number of circular and noncircular signals in a given signal subspace. Section IV discusses widely linear estimation, which takes the information in the complementary covariance into account, and Section V deals with ICA. We hope that our paper will illuminate the role impropriety and noncircularity play in signal processing, and demonstrate when and how they should be taken into account.

II. PRELIMINARIES

A. Wirtinger Calculus

In statistical signal processing, we often deal with a real-valued cost function, such as a likelihood function or a quadratic form, which is then either analytically or numerically optimized with respect to a vector or matrix of parameters. What happens when the parameters are complex-valued? That is, how do we differentiate a real-valued function with respect to a complex argument?

Consider first a complex-valued function $f(z) = u(z_r, z_i) + jv(z_r, z_i)$, where $z = z_r + jz_i$. The classical definition of

complex differentiability requires that the derivatives defined as the limit

$$f'(z_0) = \lim_{\Delta z \to 0} \frac{f(z_0 + \Delta z) - f(z_0)}{\Delta z} \qquad (1)$$

are independent of the direction in which Δz approaches 0 in the complex plane. This requires that the Cauchy-Riemann equations [3], [5]

$$\frac{\partial u}{\partial z_r} = \frac{\partial v}{\partial z_i} \quad \text{and} \quad \frac{\partial u}{\partial z_i} = -\frac{\partial v}{\partial z_r} \qquad (2)$$

be satisfied. These conditions are necessary for $f(z)$ to be complex-differentiable. If the partial derivatives of $u(z_r, z_i)$ and $v(z_r, z_i)$ are continuous, then they are sufficient as well. A function that is complex-differentiable on its entire domain is called *holomorphic* or *analytic*. Obviously, since real-valued cost functions have $v(z_r, z_i) \equiv 0$, the Cauchy-Riemann conditions do not hold, and hence cost functions are not analytic. Indeed, the Cauchy-Riemann equations impose a rigid structure on $u(z_r, z_i)$ and $v(z_r, z_i)$ and thus $f(z)$. A simple demonstration of this fact is that either $u(z_r, z_i)$ or $v(z_r, z_i)$ alone suffices to express the derivatives of an analytic function.

The usual approach to overcome this basic limitation is to evaluate separate derivatives with respect to the real and imaginary parts of a non-analytic function, which may then be stacked in a vector of twice the original dimension. In the end, this solution is converted back into the complex domain. Needless to say, this approach is cumbersome. In some cases (for example, $f(z) = |g(z)|^2$ [5]), it even requires additional assumptions (for the example, analytic $g(z)$) to make it work.

Wirtinger calculus [137]—also called the $\mathbb{CR}$ calculus [59]—provides a framework for differentiating non-analytic functions. Importantly, it allows performing all the derivations in the complex domain, in a manner very similar to the real-valued case. All the evaluations become quite straightforward making many tools and methods developed for the real case readily available for the complex case. By keeping the expressions simple, we also eliminate the need for simplifying, but often inaccurate, assumptions such as circularity.

Wirtinger calculus only requires that $f(z)$ be differentiable when expressed as a function $f : \mathbb{R}^2 \to \mathbb{R}^2$. Such a function is called real-differentiable. If $u(z_r, z_i)$ and $v(z_r, z_i)$ have continuous partial derivatives with respect to z_r and z_i, f is real-differentiable. We now define the two generalized complex derivatives

$$\frac{\partial f}{\partial z} \triangleq \frac{1}{2} \left(\frac{\partial f}{\partial z_r} - j \frac{\partial f}{\partial z_i} \right) \quad \text{and} \quad \frac{\partial f}{\partial z^*} \triangleq \frac{1}{2} \left(\frac{\partial f}{\partial z_r} + j \frac{\partial f}{\partial z_i} \right), \qquad (3)$$

which can be formally "derived" by writing $z_r = (z + z^*)/2$ and $z_i = (z - z^*)/2j$ and then using the chain rule [104]. These generalized derivatives can be formally implemented by regarding f as a bivariate function $f(z, z^*)$ and treating z and z^* as *independent variables*. That is, when applying $\partial f / \partial z$, we take the derivative with respect to z, while formally treating z^* as a constant. Similarly, $\partial f / \partial z^*$ yields the derivative with respect to z^*, formally regarding z as a constant. Thus, there is no need to develop new differentiation rules. This was shown

in [19] in 1983 without a specific reference to Wirtinger's earlier work [137]. Interestingly, many of the references that refer to [19] and use the generalized derivatives (3) evaluate them by computing derivatives with respect to z_r and z_i, instead of z and z^*. This leads to unnecessarily complicated derivations.

The Cauchy-Riemann equations can simply be stated as $\partial f/\partial z^* = 0$. In other words, an analytic function cannot depend on z^*. If f is analytic, then the usual complex derivative in (1) and $\partial f/\partial z$ in (3) coincide. So Wirtinger calculus contains standard complex calculus as a special case.

For *real-valued* $f(z)$, we have $(\partial f/\partial z)^* = \partial f/\partial z^*$, *i.e.*, the derivative and the conjugate derivative are complex conjugates of each other. Because they are related through conjugation, we only need to compute one or the other. As a consequence, a necessary and sufficient condition for real-valued f to have a stationary point is $\partial f/\partial z = 0$. An equivalent necessary and sufficient condition is $\partial f/\partial z^* = 0$ [19].

As an example for the application of Wirtinger derivatives, consider the real-valued function $f(z) = |z|^4 = z_r^4 + 2z_r^2 z_i^2 + z_i^4$. We can evaluate $\partial f/\partial z$ by differentiating separately with respect to z_r and z_i,

$$\frac{\partial f}{\partial z} = \frac{1}{2}\left(\frac{\partial f}{\partial z_r} - j\frac{\partial f}{\partial z_i}\right) = 2z_r^3 + 2z_r z_i^2 - 2j(z_r^2 z_i + z_i^3), \quad (4)$$

or we can write the function as $f(z) = f(z, z^*) = z^2(z^*)^2$ and differentiate by treating z^* as a constant,

$$\frac{\partial f}{\partial z} = 2z(z^*)^2. \quad (5)$$

The second approach is clearly simpler. It can be easily shown that the two expressions, (4) and (5), are equal. However, while the expression in (4) can easily be derived from (5), it is not quite as straightforward the other way round. Because $f(z)$ is real-valued, there is no need to compute $\partial f/\partial z^*$: it is simply the conjugate of $\partial f/\partial z$.

Evaluation of integrals, *e.g.*, for calculating probabilities, is also commonly encountered. The same observation that allows the use of a representation in the form $f(z, z^*)$ and treating z and z^* as independent variables—when obviously they are not,—can also be made when calculating integrals. When we consider $f(\cdot)$ as a function of real and imaginary parts, the definition of an integral is well understood as the integral of the function $f(z_r, z_i)$ in a region $\mathcal{R}$ defined on $\mathbb{R} \times \mathbb{R}$ as $\iint_{\mathcal{R}} f(z_r, z_i)\, dz_r\, dz_i$. The integral $\iint f(z, z^*)\, dz\, dz^*$, on the other hand, is not meaningful as we cannot vary the two variables z and z^* independently, and we cannot define the region corresponding to $\mathcal{R}$ in the complex domain. However, using Green's theorem [3] or Stokes's theorem [51], [52], we can write the real-valued integral as a contour integral in the complex domain [91]

$$\iint_{\mathcal{R}} f(z_r, z_i)\, dz_r\, dz_i = \frac{1}{2j} \oint_{C_{\mathcal{R}}} F(z, z^*)\, dz \quad (6)$$

where

$$\frac{\partial F(z, z^*)}{\partial z^*} = f(z, z^*).$$

Here, we have to assume that $f(z_r, z_i)$ is continuous through the simply connected region $\mathcal{R}$, and $C_{\mathcal{R}}$ describes its contour. By transforming the integral defined in the real domain to a contour integral in the complex domain, the formula shows the dependence on the two variables z and z^* in a natural manner. In [91], the application of the integral relationship in (6) is discussed in detail, along with examples, for evaluating probability masses when $f(z_r, z_i)$ is a probability density function. Another example is given in [7], where (6) is used to evaluate the expectation of the score function with a generalized Gaussian distribution.

Wirtinger calculus extends straightforwardly to functions $f : \mathbb{C}^N \to \mathbb{C}^M$ or $f : \mathbb{C}^{N \times M} \to \mathbb{C}$. There is no need to develop new differentiation rules for Wirtinger derivatives. All rules for taking derivatives for real functions remain valid. However, care must be taken to properly distinguish between the variables with respect to which differentiation is performed and those that are formally regarded as constants. So all the expressions from the real-valued case given, for example, in [98] can be straightforwardly applied to the complex case. For instance, for $g(\mathbf{Z}, \mathbf{Z}^*) = \mathrm{Trace}(\mathbf{Z}\mathbf{Z}^H)$, we obtain

$$\frac{\partial g}{\partial \mathbf{Z}} = \frac{\partial \mathrm{Trace}(\mathbf{Z}(\mathbf{Z}^*)^T)}{\partial \mathbf{Z}} = \mathbf{Z}^* \quad \text{and} \quad \frac{\partial g}{\partial \mathbf{Z}^*} = \mathbf{Z}.$$

There are a number of comprehensive references (e.g., [59], [5], [115], [42]) on Wirtinger calculus that deal with the chain rule for non-analytic functions, complex gradients and Hessians, and complex Taylor series expansions.

In summary, Wirtinger calculus defines a framework in which the derivatives of non-analytic functions can be computed in a manner very similar to the real-valued or the analytic case. The framework is general in that it also includes derivatives of analytic functions as a special case. It allows keeping the computations in $\mathbb{C}^N$ and eliminates the need to double the dimensionality as in the approach taken by [126]. Most of the tools developed for real-valued signals can thus be easily extended to the complex setting, see *e.g.*, [5], [65], [66]. In this overview paper, we use Wirtinger calculus for a straightforward derivation of the complex least-mean-square (LMS) algorithm in Section IV-C, and an extension of ICA, based on higher-order statistics, to complex data in Section V-B. Without Wirtinger calculus, these extensions would be significantly more difficult.

B. Widely Linear Transformations, Inner Products, and Quadratic Forms

Let us now explore the different ways that linear transformations can be described in the real and complex domains. In order to do so, we construct three closely related vectors from two real vectors $\mathbf{x}_r \in \mathbb{R}^N$ and $\mathbf{x}_i \in \mathbb{R}^N$. The first is the *real composite* $2N$-dimensional vector $\mathbf{x}_{\mathbb{R}} = [\mathbf{x}_r^T, \mathbf{x}_i^T]^T$, obtained by stacking $\mathbf{x}_r$ on top of $\mathbf{x}_i$. The second is the *complex* vector $\mathbf{x} = \mathbf{x}_r + j\mathbf{x}_i$, and the third is the *complex augmented* vector $\underline{\mathbf{x}} = [\mathbf{x}^T, \mathbf{x}^H]^T$, obtained by stacking $\mathbf{x}$ on top of its complex conjugate $\mathbf{x}^*$. The space of complex augmented vectors, whose bottom N entries are the complex conjugates of the top N entries, is denoted by $\mathbb{C}_*^{2N}$. Augmented vectors are always underlined. In much of our discussion, our focus will be on

complex-valued quantities, where we will be using $\mathbf{x}$ and its augmentation $\underline{\mathbf{x}}$.

The complex augmented vector $\underline{\mathbf{x}} \in \mathbb{C}_*^{2N}$ is related to the real composite vector $\mathbf{x}_{\mathbb{R}} \in \mathbb{R}^{2N}$ as $\underline{\mathbf{x}} = \mathbf{T}_N \mathbf{x}_{\mathbb{R}}$ and $\mathbf{x}_{\mathbb{R}} = \frac{1}{2}\mathbf{T}_N^H\underline{\mathbf{x}}$, where the real-to-complex transformation

$$\mathbf{T}_N = \begin{bmatrix} \mathbf{I} & j\mathbf{I} \\ \mathbf{I} & -j\mathbf{I} \end{bmatrix} \in \mathbb{C}^{2N \times 2N} \tag{7}$$

is unitary up to a factor of 2, *i.e.*, $\mathbf{T}_N \mathbf{T}_N^H = \mathbf{T}_N^H \mathbf{T}_N = 2\mathbf{I}$. The complex augmented vector $\underline{\mathbf{x}}$ is obviously an equivalent redundant, but convenient, representation of $\mathbf{x}_{\mathbb{R}}$. When the size of $\mathbf{T}_N$ is clear, we may drop the subscript N for economy.

If a real linear transformation $\mathbf{M} \in \mathbb{R}^{2M \times 2N}$ is applied to the real composite vector $\mathbf{x}_{\mathbb{R}} \in \mathbb{R}^{2N}$, it yields a real composite vector $\mathbf{y}_{\mathbb{R}} \in \mathbb{R}^{2M}$,

$$\mathbf{y}_{\mathbb{R}} = \begin{bmatrix} \mathbf{y}_r \\ \mathbf{y}_i \end{bmatrix} = \begin{bmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{x}_r \\ \mathbf{x}_i \end{bmatrix} = \mathbf{M}\mathbf{x}_{\mathbb{R}},$$

where $\mathbf{M}_{ij} \in \mathbb{R}^{M \times N}$. The augmented complex version of $\mathbf{y}_{\mathbb{R}}$ is

$$\underline{\mathbf{y}} = \begin{bmatrix} \mathbf{y} \\ \mathbf{y}^* \end{bmatrix} = \mathbf{T}_M \begin{bmatrix} \mathbf{y}_r \\ \mathbf{y}_i \end{bmatrix} = \left(\tfrac{1}{2}\mathbf{T}_M\mathbf{M}\mathbf{T}_N^H\right)(\mathbf{T}_N\mathbf{x}_{\mathbb{R}}) = \underline{\mathbf{H}}\,\underline{\mathbf{x}}, \tag{8}$$

with $\mathbf{y} = \mathbf{y}_r + j\mathbf{y}_i$. The matrix $\underline{\mathbf{H}} \in \mathbb{C}^{2M \times 2N}$ is called an *augmented matrix* because it satisfies a particular block pattern, where the SE block is the conjugate of the NW block, and the SW block is the conjugate of the NE block:

$$\underline{\mathbf{H}} = \tfrac{1}{2}\mathbf{T}_M\mathbf{M}\mathbf{T}_N^H = \begin{bmatrix} \mathbf{H}_1 & \mathbf{H}_2 \\ \mathbf{H}_2^* & \mathbf{H}_1^* \end{bmatrix} \tag{9}$$

$$\mathbf{H}_1 = \tfrac{1}{2}[\mathbf{M}_{11} + \mathbf{M}_{22} + j(\mathbf{M}_{21} - \mathbf{M}_{12})]$$
$$\mathbf{H}_2 = \tfrac{1}{2}[\mathbf{M}_{11} - \mathbf{M}_{22} + j(\mathbf{M}_{21} + \mathbf{M}_{12})]$$

Hence, $\underline{\mathbf{H}}$ is an augmented description of the *widely linear* [103] or *linear–conjugate-linear* transformation

$$\mathbf{y} = \mathbf{H}_1\mathbf{x} + \mathbf{H}_2\mathbf{x}^*.$$

Even though the representation in (9) contains some redundancy as the northern blocks determine the southern blocks, it proves to be a powerful tool. For instance, it enables easy concatenation of widely linear transformations.

Obviously, the set of complex linear transformations, $\mathbf{y} = \mathbf{H}_1\mathbf{x}$ with $\mathbf{H}_2 = \mathbf{0}$, is a subset of the set of widely linear transformations. A complex linear transformation (sometimes called *strictly linear* for emphasis) has the equivalent real representation

$$\begin{bmatrix} \mathbf{y}_r \\ \mathbf{y}_i \end{bmatrix} = \begin{bmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ -\mathbf{M}_{12} & \mathbf{M}_{11} \end{bmatrix} \begin{bmatrix} \mathbf{x}_r \\ \mathbf{x}_i \end{bmatrix}. \tag{10}$$

To summarize, linear transformations on $\mathbb{R}^{2N}$ are linear on $\mathbb{C}^N$ only if they have the particular structure (10). Otherwise, the equivalent operation on $\mathbb{C}^N$ is widely linear. For analysis, it is often preferable to represent $\mathbb{R}$-linear operations as $\mathbb{C}$-widely linear operations. However, from a hardware implementation point of view, $\mathbb{R}$-linear transformations are usually preferable over $\mathbb{C}$-widely linear transformations because the former require fewer real operations (additions and multiplications) than the latter.

Next, we look at the different representations of inner products and quadratic forms. Consider the two $2N$-dimensional real composite vectors $\mathbf{x}_{\mathbb{R}} = [\mathbf{x}_r^T, \mathbf{x}_i^T]^T$ and $\mathbf{y}_{\mathbb{R}} = [\mathbf{y}_r^T, \mathbf{y}_i^T]^T$, the corresponding N-dimensional complex vectors $\mathbf{x} = \mathbf{x}_r + j\mathbf{x}_i$ and $\mathbf{y} = \mathbf{y}_r + j\mathbf{y}_i$, and their complex augmented descriptions $\underline{\mathbf{x}} = \mathbf{T}\mathbf{x}_{\mathbb{R}}$ and $\underline{\mathbf{y}} = \mathbf{T}\mathbf{y}_{\mathbb{R}}$. We may now relate the inner products defined on $\mathbb{R}^{2N}$, $\mathbb{C}_*^{2N}$, and $\mathbb{C}^N$ as

$$\mathbf{x}_{\mathbb{R}}^T\mathbf{y}_{\mathbb{R}} = \tfrac{1}{2}\underline{\mathbf{x}}^H\underline{\mathbf{y}} = \operatorname{Re}\{\mathbf{x}^H\mathbf{y}\}.$$

Thus, the usual inner product $\mathbf{x}_{\mathbb{R}}^T\mathbf{y}_{\mathbb{R}}$ defined on $\mathbb{R}^{2N}$ equals (up to a factor of $1/2$) the inner product $\underline{\mathbf{x}}^H\underline{\mathbf{y}}$ defined on $\mathbb{C}_*^{2N}$, and also the real part of the usual inner product $\mathbf{x}^H\mathbf{y}$ defined on $\mathbb{C}^N$. Both the inner product on $\mathbb{C}_*^{2N}$ and the inner product on $\mathbb{C}^N$ are useful.

Another common real-valued expression is the quadratic form $\mathbf{x}_{\mathbb{R}}^T\mathbf{M}\mathbf{x}_{\mathbb{R}}$, which may be written as a (real-valued) *widely quadratic form* in $\mathbf{x}$:

$$\mathbf{x}_{\mathbb{R}}^T\mathbf{M}\mathbf{x}_{\mathbb{R}} = \tfrac{1}{2}(\mathbf{x}_{\mathbb{R}}^T\mathbf{T}^H)\left(\tfrac{1}{2}\mathbf{T}\mathbf{M}\mathbf{T}^H\right)(\mathbf{T}\mathbf{x}_{\mathbb{R}}) = \tfrac{1}{2}\underline{\mathbf{x}}^H\underline{\mathbf{H}}\,\underline{\mathbf{x}}$$

The augmented matrix $\underline{\mathbf{H}}$ and the real matrix $\mathbf{M}$ are connected as before in (9). Thus, we obtain

$$\mathbf{x}_{\mathbb{R}}^T\mathbf{M}\mathbf{x}_{\mathbb{R}} = \mathbf{x}^H\mathbf{H}_1\mathbf{x} + \operatorname{Re}\left(\mathbf{x}^H\mathbf{H}_2\mathbf{x}^*\right).$$

C. Statistics of Complex-valued Random Variables and Vectors

We define a complex random variable x as $x = x_r + jx_i$, where x_r and x_i are a pair of real random variables. Similarly, an N-dimensional complex random vector $\mathbf{x}$ is defined as $\mathbf{x} = \mathbf{x}_r + j\mathbf{x}_i$, where $\mathbf{x}_r$ and $\mathbf{x}_i$ are a pair of N-dimensional real random vectors. Note that we do not differentiate in notation between a random vector and its realization, as the meaning should be clear from the context. The probability distribution (density) of a complex random vector is interpreted as the $2N$-dimensional joint distribution (density) of its real and imaginary parts. If the probability density function (pdf) exists, we write this as

$$p_{\mathbf{x}}(\mathbf{x}) = p_{\mathbf{x}}(\mathbf{x}_r + j\mathbf{x}_i) \triangleq p_{\mathbf{x}_r\mathbf{x}_i}(\mathbf{x}_r, \mathbf{x}_i).$$

If there is no risk of confusion, the subscripts may also be dropped. For a function $g : \mathcal{D} \to \mathbb{C}^N$ whose domain $\mathcal{D}$ includes the range of $\mathbf{x}$, the expectation operator is defined as

$$E\{g(\mathbf{x})\} = E\{\operatorname{Re}[g(\mathbf{x})]\} + jE\{\operatorname{Im}[g(\mathbf{x})]\}$$
$$= \int_{\mathbb{R}^{2N}} g(\mathbf{x}_r + j\mathbf{x}_i)p(\mathbf{x}_r + j\mathbf{x}_i)\,d\mathbf{x}_r\,d\mathbf{x}_i. \tag{11}$$

This integral can also be evaluated using contour integrals as in (6) writing the function as $g(\mathbf{x}, \mathbf{x}^*)$ and the pdf as $p(\mathbf{x}, \mathbf{x}^*)$.

Unless otherwise stated, we assume that all random vectors have zero mean. In order to characterize the second-order statistical properties of $\mathbf{x} = \mathbf{x}_r + j\mathbf{x}_i$, we consider the real composite random vector $\mathbf{x}_{\mathbb{R}}$. Its covariance matrix is

$$\mathbf{C}_{x_{\mathbb{R}}x_{\mathbb{R}}} = E\{\mathbf{x}_{\mathbb{R}}\mathbf{x}_{\mathbb{R}}^T\} = \begin{bmatrix} \mathbf{C}_{x_rx_r} & \mathbf{C}_{x_rx_i} \\ \mathbf{C}_{x_rx_i}^T & \mathbf{C}_{x_ix_i} \end{bmatrix}$$

with $\mathbf{C}_{x_r x_r} = E\{\mathbf{x}_r \mathbf{x}_r^T\}$, $\mathbf{C}_{x_r x_i} = E\{\mathbf{x}_r \mathbf{x}_i^T\}$, and $\mathbf{C}_{x_i x_i} = E\{\mathbf{x}_i \mathbf{x}_i^T\}$. The *augmented covariance matrix* of $\mathbf{x}$ is [129], [101], [114]

$$\underline{\mathbf{C}}_{xx} = E\{\underline{\mathbf{x}}\,\underline{\mathbf{x}}^H\} = \mathbf{T}\mathbf{C}_{x_{\mathbb{R}} x_{\mathbb{R}}}\mathbf{T}^H = \begin{bmatrix} \mathbf{C}_{xx} & \widetilde{\mathbf{C}}_{xx} \\ \widetilde{\mathbf{C}}_{xx}^* & \mathbf{C}_{xx}^* \end{bmatrix} = \underline{\mathbf{C}}_{xx}^H. \tag{12}$$

The NW block of the augmented covariance matrix is the usual (Hermitian and positive semi-definite) covariance matrix

$$\mathbf{C}_{xx} = E\{\mathbf{x}\mathbf{x}^H\} = \mathbf{C}_{x_r x_r} + \mathbf{C}_{x_i x_i} + j(\mathbf{C}_{x_r x_i}^T - \mathbf{C}_{x_r x_i}) = \mathbf{C}_{xx}^H \tag{13}$$

and the NE block is the *complementary covariance matrix*

$$\widetilde{\mathbf{C}}_{xx} = E\{\mathbf{x}\mathbf{x}^T\} = \mathbf{C}_{x_r x_r} - \mathbf{C}_{x_i x_i} + j(\mathbf{C}_{x_r x_i}^T + \mathbf{C}_{x_r x_i}) = \widetilde{\mathbf{C}}_{xx}^T, \tag{14}$$

which uses a regular transpose rather than a Hermitian (conjugate) transpose. Other names for $\widetilde{\mathbf{C}}_{xx}$ include *pseudo-covariance matrix* [84], *conjugate covariance matrix* [44] or *relation matrix* [102]. It is important to note that, in general, *both* $\mathbf{C}_{xx}$ *and* $\widetilde{\mathbf{C}}_{xx}$ *are required for a complete second-order characterization of* $\mathbf{x}$. In the important special case where the complementary covariance matrix vanishes, $\widetilde{\mathbf{C}}_{xx} = \mathbf{0}$, $\mathbf{x}$ is called *proper*, otherwise *improper* [84].

The conditions for propriety on the covariance and cross-covariance of real and imaginary parts $\mathbf{x}_r$ and $\mathbf{x}_i$ are $\mathbf{C}_{x_r x_r} = \mathbf{C}_{x_i x_i}$ and $\mathbf{C}_{x_r x_i} = -\mathbf{C}_{x_r x_i}^T$. When $x = x_r + jx_i$ is scalar, then $C_{x_r x_i} = 0$ is necessary for propriety. If $\mathbf{x}$ is proper, its Hermitian covariance matrix is

$$\mathbf{C}_{xx} = 2\mathbf{C}_{x_r x_r} - 2j\mathbf{C}_{x_r x_i} = 2\mathbf{C}_{x_i x_i} + 2j\mathbf{C}_{x_r x_i}^T,$$

and its augmented covariance matrix $\underline{\mathbf{C}}_{xx}$ is block-diagonal. If complex x is proper and scalar, then $C_{xx} = 2C_{x_r x_r} = 2C_{x_i x_i}$. It is easy to see that propriety is preserved by strictly linear transformations, which are represented by block-diagonal augmented matrices.

If $\mathbf{C}_{xx}$ is nonsingular, the following three conditions are necessary and sufficient for $\mathbf{C}_{xx}$ and $\widetilde{\mathbf{C}}_{xx}$ to be covariance and complementary covariance matrices of a complex random vector $\mathbf{x}$ [101]: 1) The covariance matrix $\mathbf{C}_{xx}$ is Hermitian and positive semi-definite; 2) The complementary covariance matrix is symmetric, $\widetilde{\mathbf{C}}_{xx} = \widetilde{\mathbf{C}}_{xx}^T$; 3) The Schur complement of $\mathbf{C}_{xx}$ within the augmented covariance matrix, $\mathbf{C}_{xx} - \widetilde{\mathbf{C}}_{xx}\mathbf{C}_{xx}^{-*}\widetilde{\mathbf{C}}_{xx}^*$, is positive semi-definite.

Circularity: It is also possible to define a stronger version of propriety in terms of the probability distribution of a random vector. A vector is called *circular* if its probability distribution is *rotationally invariant*, i.e., $\mathbf{x}$ and $\mathbf{x}' = e^{j\alpha}\mathbf{x}$ have the same probability distribution for any given real α. Circularity does not imply any condition on the standard covariance matrix $\mathbf{C}_{xx}$ because

$$\mathbf{C}_{x'x'} = E\{\mathbf{x}'\mathbf{x}'^H\} = E\{e^{j\alpha}\mathbf{x}\mathbf{x}^H e^{-j\alpha}\} = \mathbf{C}_{xx}. \tag{15}$$

On the other hand,

$$\widetilde{\mathbf{C}}_{x'x'} = E\{\mathbf{x}'\mathbf{x}'^T\} = E\{e^{j\alpha}\mathbf{x}\mathbf{x}^T e^{j\alpha}\} = e^{j2\alpha}\widetilde{\mathbf{C}}_{xx} \tag{16}$$

is true for arbitrary α if and only if $\widetilde{\mathbf{C}}_{xx} = \mathbf{0}$. Because the Gaussian distribution is completely determined by second-order statistics, a complex Gaussian random vector $\mathbf{x}$ is circular if and only if it is zero-mean and proper [48].

Propriety requires that second-order moments be rotationally invariant, whereas circularity requires that the distribution, and thus *all* moments (if they exist), be rotationally invariant. Therefore, circularity implies zero mean and propriety, but not vice versa, and impropriety implies noncircularity, but not vice versa. By extending the reasoning of (15) and (16) to higher-order moments, we see that if $\mathbf{x}$ is circular, a pth order moment can be nonzero only if it has the same number of conjugated and nonconjugated terms [100]. In particular, all odd-order moments must be zero. This holds for arbitrary order p.

We call a vector $\mathbf{x}$ *pth order circular* [100] or *pth order proper* if the only nonzero moments up to order p have the same number of conjugated and nonconjugated terms. In particular, all odd moments up to order p must be zero. Therefore, for zero-mean random vectors, the terms *proper* and *second-order circular* are equivalent. We note that, while this terminology is most common, there is not uniform agreement in the literature. Some researchers use the terms "proper" and "circular" interchangeably. More detailed discussions of higher-order statistics, Taylor series expansions, and characteristic functions for improper random vectors are provided in [11], [42], [115].

D. Statistics of Complex-valued Random Processes

In this section, we provide a very brief introduction to complex discrete-time random processes. Complex random processes have been studied by [12], [33], [84], [100], [102], [111], [132], among others. For a comprehensive review, we refer the reader to [115].

The covariance function of a complex-valued discrete-time random process $x(k)$ is defined as

$$c(k, m) = E\{x(k+m)x^*(k)\} - E\{x(k+m)\}E\{x^*(k)\},$$

where the first term is the correlation function. To completely describe the second-order statistics, we also define the *complementary covariance function* (also called the pseudo-covariance or relation function) as

$$\tilde{c}(k, m) = E\{x(k+m)x(k)\} - E\{x(k+m)\}E\{x(k)\}.$$

If the complementary covariance function vanishes for all pairs (k, m), then $x(k)$ is called *proper*, otherwise *improper*. If and only if $x(k)$ is zero-mean and proper is it *second-order circular*. The definitions in the continuous-time case are completely analogous.

A random signal $x(k)$ is stationary if all its statistical properties are invariant to any given time-shift (translations of the origin), or equivalently, if the family of distributions that describe the random process as a collection of random variables is shift-invariant. To define wide-sense stationarity, we therefore consider the complete second-order statistical characterization including the complementary covariance function. We call a complex random process $x(k)$ *wide-sense stationary* (WSS) if and only if $E\{x(k)\} \equiv \mu_x$, $c(k, m) \equiv c(m)$, and $\tilde{c}(k, m) \equiv \tilde{c}(m)$ are all independent of k.

Since the complementary covariance function is commonly ignored, the "traditional" definition of wide-sense stationarity for complex processes only considers the covariance function,

allowing a shift-dependent complementary covariance function. Some researchers then introduce a different term (such as *second-order stationary* [100]) to refer to the condition where both the covariance and complementary covariance function are shift-invariant.

Let $x(k)$ be a second-order zero-mean WSS process, and let $\underline{\mathbf{x}}(k) = [x(k), x^*(k)]^T$ denote the corresponding augmented process. We define the *augmented power spectral density matrix* of $x(k)$ as the discrete-time Fourier transform of the covariance function of $\underline{\mathbf{x}}(k)$, which is given by

$$\underline{\mathbf{C}}(\theta) \triangleq \mathcal{F}\{E\{\underline{\mathbf{x}}(k+m)\underline{\mathbf{x}}^H(k)\}\} = \begin{bmatrix} C(\theta) & \widetilde{C}(\theta) \\ \widetilde{C}^*(\theta) & C(-\theta) \end{bmatrix}.$$

In this equation, $C(\theta) = \mathcal{F}\{c(m)\}$ is the *power spectral density* and $\widetilde{C}(\theta) = \mathcal{F}\{\tilde{c}(m)\}$ is called the *complementary power spectral density*. When the complementary power spectral density vanishes for all θ, $\widetilde{C}(\theta) \equiv 0$, the process $x(k)$ is proper.

The augmented power spectral density matrix is positive semi-definite. As a consequence, there exists a WSS random process $x(k)$ with power spectral density $C(\theta)$ and complementary power spectral density $\widetilde{C}(\theta)$ if and only if [102]: (1) $C(\theta) \geq 0$; (2) $\widetilde{C}(\theta) = \widetilde{C}(-\theta)$ (but $\widetilde{C}(\theta)$ is generally complex); and (3) $|\widetilde{C}(\theta)|^2 \leq C(\theta)C(-\theta)$. The third condition allows us to establish the classical result that analytic WSS signals without a DC component must be proper [100]. If $x(k)$ is analytic, then $C(\theta) = 0$ for $\theta \in [-\pi, 0)$. If it does not have a DC component, then $C(\theta) = 0$ for $\theta \in [-\pi, 0]$. Therefore, $|\widetilde{C}(\theta)|^2 \leq C(\theta)C(-\theta) \equiv 0$ implies $\widetilde{C}(\theta) \equiv 0$. There are further connections between stationarity and circularity, which are explored in detail in [100]. An algorithm for the simulation of improper WSS processes having specified covariance and complementary covariance functions is given in [108].

E. Gaussian Random Variables and Vectors

Next, we look at the the ever-important Gaussian distribution. In order to derive the general complex multivariate Gaussian pdf (proper/circular or improper/noncircular) for zero-mean $\mathbf{x} = \mathbf{x}_r + j\mathbf{x}_i$, we begin with the Gaussian pdf of the composite random vector of real and imaginary parts $\mathbf{x}_{\mathbb{R}} = [\mathbf{x}_r^T, \mathbf{x}_i^T]^T$:

$$p(\mathbf{x}_{\mathbb{R}}) = \frac{1}{(2\pi)^{2N/2}\sqrt{\det \mathbf{C}_{x_{\mathbb{R}}x_{\mathbb{R}}}}} \exp\left\{-\tfrac{1}{2}\mathbf{x}_{\mathbb{R}}^T \mathbf{C}_{x_{\mathbb{R}}x_{\mathbb{R}}}^{-1} \mathbf{x}_{\mathbb{R}}\right\}. \tag{17}$$

We first consider the nondegenerate case where all covariance matrices are invertible. Using $\underline{\mathbf{x}} = \mathbf{T}\mathbf{x}_{\mathbb{R}}$, $\mathbf{C}_{x_{\mathbb{R}}x_{\mathbb{R}}}^{-1} = \mathbf{T}^H \underline{\mathbf{C}}_{xx}^{-2N}\mathbf{T}$, and $\det \underline{\mathbf{C}}_{xx} = 2^{2N}\det \mathbf{C}_{x_{\mathbb{R}}x_{\mathbb{R}}}$, where $\mathbf{T}$ was defined in (7), we obtain the pdf of complex $\mathbf{x}$ [101], [129]:

$$p(\mathbf{x}) = \frac{1}{\pi^N \sqrt{\det \underline{\mathbf{C}}_{xx}}} \exp\left\{-\tfrac{1}{2}\underline{\mathbf{x}}^H \underline{\mathbf{C}}_{xx}^{-1} \underline{\mathbf{x}}\right\}. \tag{18}$$

This pdf depends algebraically on $\underline{\mathbf{x}}$, *i.e.*, $\mathbf{x}$ and $\mathbf{x}^*$, but is interpreted as the joint pdf of $\mathbf{x}_r$ and $\mathbf{x}_i$, and can be used for proper/circular or improper/noncircular $\mathbf{x}$. In the past, the term "complex Gaussian distribution" often implicitly

assumed circularity. Therefore, some researchers call a noncircular complex Gaussian random vector "generalized complex Gaussian." The simplification that occurs when $\widetilde{\mathbf{C}}_{xx} = \mathbf{0}$ is obvious and leads to the pdf of a complex *proper* and *circular* Gaussian random vector $\mathbf{x}$ [139], [47]:

$$p(\mathbf{x}) = \frac{1}{\pi^N \det \mathbf{C}_{xx}} \exp\left\{-\mathbf{x}^H \mathbf{C}_{xx}^{-1}\mathbf{x}\right\}$$

A generalization of the Gaussian distribution is the family of elliptical distributions. A study of *complex* elliptical distributions has been provided by [94] and in Section 2.3.4 of [115].

The scalar complex Gaussian: The case of a scalar Gaussian $x = x_r + jx_i$ provides some interesting insights [115], [92]. The real bivariate Gaussian vector $[x_r, x_i]$ can be characterized by $C_{x_r x_r} = E\{x_r^2\}$ (the variance of x_r), $C_{x_i x_i} = E\{x_i^2\}$ (the variance of x_i), and $\rho_{x_r x_i} = E\{x_r x_i\}/\sqrt{E\{x_r^2\}E\{x_i^2\}}$ (the correlation coefficient between x_r and x_i). From (18), the pdf of the complex Gaussian variable $x = x_r + jx_i$ can be expressed as

$$p(x) = \frac{1}{\pi C_{xx}\sqrt{1-|\rho|^2}} \exp\left\{-\frac{|x|^2 - \text{Re}\,(\rho x^{*2})}{C_{xx}(1-|\rho|^2)}\right\} \tag{19}$$

where the *complex correlation coefficient* ρ between x and x^* is defined as

$$\rho = \frac{\widetilde{C}_{xx}}{C_{xx}}.$$

So there are two equivalent parametrizations of a complex Gaussian x: We can use the three real parameters $C_{x_r x_r}$, $C_{x_i x_i}$, and $\rho_{x_r x_i}$ or the three real parameters C_{xx}, $\text{Re}\,\rho$, and $\text{Im}\,\rho$. We can obtain the second set of parameters from the first set as

$$C_{xx} = C_{x_r x_r} + C_{x_i x_i} \tag{20}$$

$$\rho = \frac{C_{x_r x_r} - C_{x_i x_i}}{C_{xx}} + j\frac{2\sqrt{C_{x_r x_r}}\sqrt{C_{x_i x_i}}\rho_{x_r x_i}}{C_{xx}} \tag{21}$$

and vice versa as

$$C_{x_r x_r} = \tfrac{1}{2}C_{xx}(1 + \text{Re}\,\rho) \tag{22}$$

$$C_{x_i x_i} = \tfrac{1}{2}C_{xx}(1 - \text{Re}\,\rho) \tag{23}$$

$$\rho_{x_r x_i} = \frac{\text{Im}\,\rho}{\sqrt{1-(\text{Re}\,\rho)^2}}. \tag{24}$$

The complex correlation coefficient $\rho = |\rho|e^{j\psi}$ satisfies $|\rho| \leq 1$. If $|\rho| = 1$, then $x = \widetilde{C}_{xx}C_{xx}^{-1}x^* = \rho x^* = e^{j\psi}x^*$ with probability 1. Equivalent conditions for $|\rho| = 1$ are $C_{x_r x_r} = 0$, or $C_{x_i x_i} = 0$, or $\rho_{x_r x_i} = \pm 1$. The first of these conditions makes the complex variable x purely imaginary and the second makes it real. The third condition means $x_i = \tan(\psi/2)x_r$ and $x = [1 + j\tan(\psi/2)]x_r$. All these cases with $|\rho| = 1$ are called *maximally noncircular* because the support of the pdf for the complex random variable x degenerates into a line in the complex plane.

For $|\rho| < 1$, there are basically four different cases:

1) If x_r and x_i have identical variances, $C_{x_r x_r} = C_{x_i x_i} = C_{xx}/2$, and are independent, $\rho_{x_r x_i} = 0$, then x is proper

and circular, *i.e.*, $\rho = 0$, and its pdf takes on the simple and much better-known form

$$p(x) = \frac{1}{\pi C_{xx}} \exp\left(-\frac{|x|^2}{C_{xx}}\right). \tag{25}$$

2) If x_r and x_i have different variances, $C_{x_r x_r} \neq C_{x_i x_i}$, but x_r and x_i are still independent, $\rho_{x_r x_i} = 0$, then ρ is *real*, $\rho = (C_{x_r x_r} - C_{x_i x_i})/C_{xx}$, and x is noncircular.

3) If x_r and x_i have identical variances, $C_{x_r x_r} = C_{x_i x_i} = C_{xx}/2$, but x_r and x_i are correlated, $\rho_{x_r x_i} \neq 0$, then ρ is *purely imaginary*, $\rho = j\rho_{x_r x_i}$, and x is noncircular.

4) If x_r and x_i have different variances, $C_{x_r x_r} \neq C_{x_i x_i}$, and are correlated, $\rho_{x_r x_i} \neq 0$. Then ρ is generally complex.

With $x = re^{j\theta}$ and $\rho = |\rho|e^{j\psi}$, we see that the pdf $p(x)$ in (19) is constant on the level curve $r^2[1 - |\rho|\cos(2\theta - \psi)] = K^2$. This contour is an ellipse, and r is maximum when $\cos(2\theta - \psi)$ is minimum. This establishes that the ellipse orientation (the angle between the real axis and the major ellipse axis) is $\theta = \psi/2$, which is half the angle of the complex correlation coefficient $\rho = |\rho|e^{j\psi}$. It can be shown (see [92]) that the degree of noncircularity $|\rho|$ is the square of the ellipse eccentricity.

Figure 1 shows contours of constant probability density for cases 1–4 listed above. In plot (a), we see the circular case with $\rho = 0$, which exhibits circular contour lines. All remaining plots are noncircular, with elliptical contour lines. We can make two observations: First, increasing the degree of noncircularity of the signal by increasing $|\rho|$ leads to ellipses with greater eccentricity. Secondly, the angle of the ellipse orientation is half the angle of ρ, as shown above. Plot (b) shows case 2: x_r and x_i have different variance but are still independent. In this situation, the ellipse orientation is either $0°$ or $90°$, depending on whether x_r or x_i has greater variance. Plot (c) shows case 3: x_r and x_i have the same variance but are now correlated. In this situation, the ellipse orientation is either $45°$ or $135°$. The general case 4 is depicted in plot (d). Now the ellipse can have an arbitrary orientation $\psi/2$, which is controlled by the angle of $\rho = |\rho|e^{j\psi}$.

Marginal and von-Mises distributions: With $x_r = r\cos\theta$, $x_i = r\sin\theta$, $dx_r\,dx_i = r\,dr\,d\theta$, it is possible to change variables and obtain the joint pdf for the polar coordinates (r, θ)

$$p_{r\theta}(r, \theta) = \frac{r}{\pi C_{xx}\sqrt{1 - |\rho|^2}} \exp\left\{-\frac{r^2[1 - |\rho|\cos(2\theta - \psi)]}{C_{xx}(1 - |\rho|^2)}\right\},$$

where $x = re^{j\theta}$ and $\rho = |\rho|e^{j\psi}$. The marginal pdf for r is obtained by integrating over θ,

$$p_r(r) = \frac{2r}{C_{xx}\sqrt{1 - |\rho|^2}} \exp\left\{-\frac{r^2}{C_{xx}(1 - |\rho|^2)}\right\} I_0\left\{\frac{r^2|\rho|}{C_{xx}(1 - |\rho|^2)}\right\} \tag{26}$$

where I_0 is the modified Bessel function of the first kind of order 0. The pdf $p_r(r)$ is invariant to ψ. In the circular case $\rho = 0$, it is the *Rayleigh pdf*

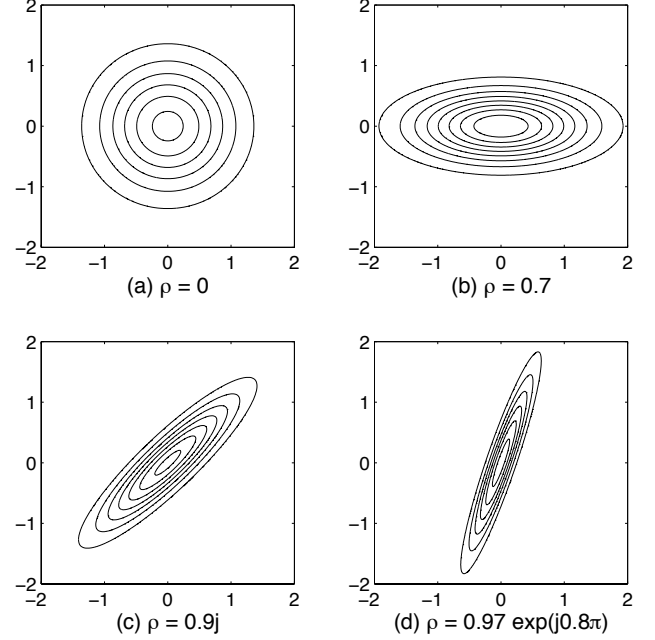$$p_r(r) = \frac{r}{C_{xx}/2} \exp\left\{-\frac{r^2}{C_{xx}}\right\}.$$

Fig. 1. Probability-density contours of complex Gaussian random variables with different complex correlation coefficient ρ.

This suggests that we call $p_r(r)$ in (26) the *improper/noncircular Rayleigh pdf*. Integrating $p_{r\theta}(r, \theta)$ over r yields the marginal pdf for θ:

$$p_\theta(\theta) = \frac{\sqrt{1 - |\rho|^2}}{2\pi[1 - |\rho|\cos(2\theta - \psi)]}, \quad -\pi < \theta \leq \pi.$$

In the circular case $\rho = 0$, this is a uniform distribution.

These marginals are illustrated in Figure 2 for $C_{xx} = 1$, $\psi = 0.8\pi$, and various values of $|\rho|$. The larger $|\rho|$ the more noncircular x becomes, and the marginal distribution of θ develops two peaks at $\theta = \psi/2 = 0.4\pi$ and $\theta = \psi/2 - \pi = -0.6\pi$. At the same time, the maximum of the pdf of r is shifted to the left.

Having derived the joint and marginal distributions of r and θ, it is also straightforward to write down the *von-Mises pdf*, which is the conditional distribution of θ given r:

$$p_{\theta|r}(\theta|r) = \frac{p_{r\theta}(r, \theta)}{p_r(r)} = \frac{1}{2\pi} \frac{\exp\{\kappa\cos(2\theta - \psi)\}}{I_0\{\kappa\}}$$

with

$$\kappa = \frac{r^2|\rho|}{C_{xx}(1 - |\rho|^2)}$$

All these results show that the parameters $\{C_{xx}, |\rho|, \psi\}$ for complex $x = x_r + jx_i$, rather than the parameters $\{C_{x_r x_r}, C_{x_i x_i}, \rho_{x_r x_i}\}$ for real (x_r, x_i), are the most natural parametrization for the joint and marginal pdfs of the polar coordinates r and θ.

It is worthwhile to comment on the fact that the Central Limit Theorem still applies to noncircular random variables. That is, adding more and more independent and identically distributed noncircular random variables leads to a sample
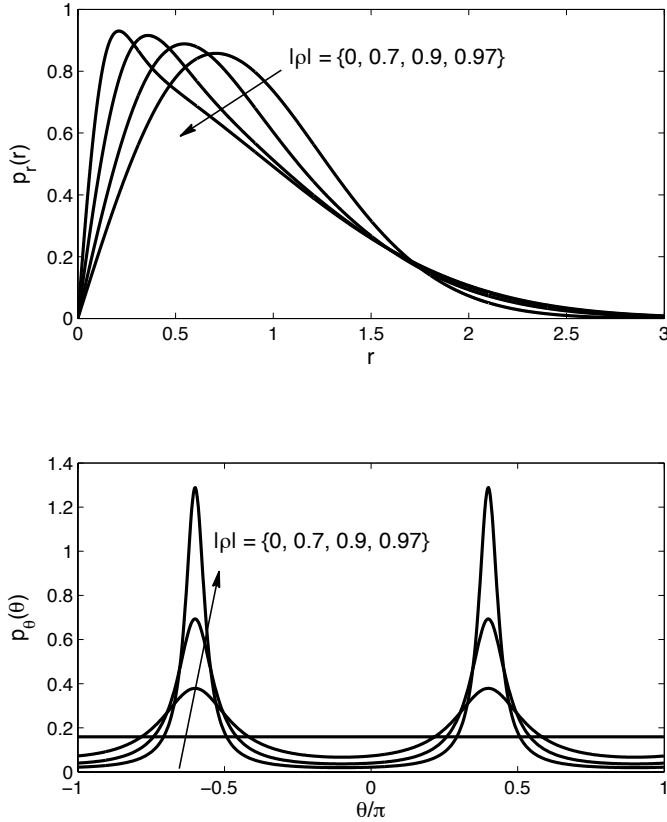
Fig. 2. Marginal pdfs for magnitude r and angle θ in the bivariate Gaussian distribution with unit variance $C_{xx} = 1$, $\psi = 0.8\pi$, and various values for $|\rho|$.

(a) Radar data

(b) 16-QAM data

(c) Wind data

Fig. 3. Scatter plots for (a) circular, (b) proper but noncircular, and (c) improper (and thus noncircular) data

average that is asymptotically Gaussian and noncircular. However, this addition has to be done *coherently*, preserving the phase of the random samples. If a large number of noncircular random variables is added *noncoherently*, *i.e.*, with randomized phase, then the phase information is washed out, and the resulting sample average becomes more and more circular.

F. Examples

Figure 3 shows scatter plots for three signals: (a) Ice Multiparameter Imaging X-Band Radar (IPIX) data from the website *http://soma.crl.mcmaster.ca/ipix/*, (b) a 16-quadrature amplitude modulated (QAM) signal, and (c) wind data obtained from *http://mesonet.agron.iastate.edu*. Figure 4 shows their corresponding covariance functions $c(m)$ and complementary covariance functions $\tilde{c}(m)$. The radar signal in (a) is narrow-band. Evidently, the gain and phase of the in-phase and quadrature channels are matched, as the data appear circular (and therefore proper). The uniform phase is due to carrier phase fluctuation from pulse-to-pulse and the amplitude fluctuations are due to variations in the scattering cross-section. The 16-QAM signal in (b) has zero complementary covariance function and is therefore proper (second-order circular). However, its distribution is not rotationally invariant and therefore
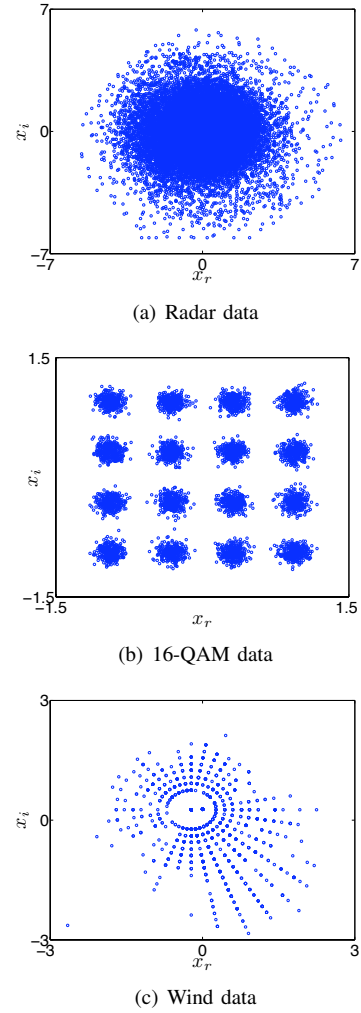
it is noncircular. The wind data in (c) is noncircular and improper.

In Figure 5 (a), we show the scatter plot of a motor component estimated using ICA of functional MRI data [106], which is naturally represented as complex valued [4]. The paradigm used in the collection of the data is a simple motor task with a box-car type time-course, *i.e.*, the stimulus has periodic on and off periods. As can be observed in the figure, the distribution of the given fMRI motor component has a highly noncircular distribution. In Figure 5 (b) and (c), we show the spatial map for the same component using a Mahalanobis Z-score threshold, which is defined as $Z_{c_{k,l}} = \sqrt{[\mathbf{u}_{k,l} - \hat{\boldsymbol{\mu}}_k]^T \, \hat{\mathbf{C}}_k^{-1} \, [\mathbf{u}_{k,l} - \hat{\boldsymbol{\mu}}_k]}$. In this expression, $\mathbf{u}_{k,l} = \left[u_{r_{k,l}}, u_{i_{k,l}} \right]^T$ is the vector of real and imaginary parts of the kth estimated source of voxel l, and $\hat{\boldsymbol{\mu}}_k$ and $\hat{\mathbf{C}}_k$ are the corresponding estimated spatial image mean vector and covariance matrix.

As demonstrated by these examples, noncircular signals do arise in practice, even though circularity has been commonly assumed in the derivation of many signal processing algorithms. As we will elaborate, taking the noncircular or

(a) Radar data

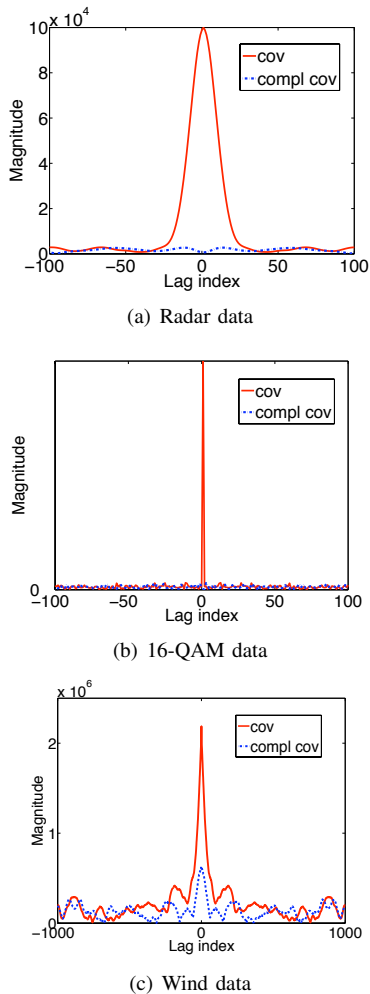(b) 16-QAM data

(c) Wind data

Fig. 4. Covariance and complementary covariance function plots for the corresponding processes in Figure 3: (a) circular, (b) proper but noncircular, (c) improper

(a) Scatter plot
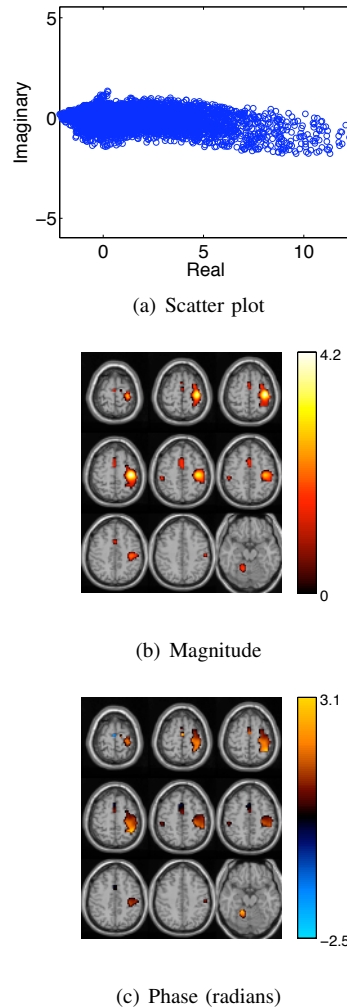

(b) Magnitude

(c) Phase (radians)

Fig. 5. (a) Scatter plot of the average voxel values of the motor component estimated using ICA from 16 subjects. (b) Magnitude and (c) phase spatial maps using Mahalanobis Z-score thresholding; only voxels with $|Z_c| > 4$ are shown.

improper nature of signals into account in their processing may provide significant payoffs. It is worth noting that, in these examples, we have classified signals as circular or noncircular simply by inspection of their scatter plots and estimated covariance functions. But such classification should be done based on sound statistical arguments. This is the topic of the next section.

III. MODEL SELECTION

When should we take noncircularity into account? On the one hand, if the signals are indeed noncircular, we would expect a noncircular model to capture their properties more accurately. On the other hand, noncircular models have more degrees of freedom than circular models, and the principle of parsimony says that one should seek simple models to avoid overfitting to noise fluctuations. So how do we choose between circular/proper and noncircular/improper models? This raises the question of how to detect whether a given signal is circular or noncircular, and more generally, how to detect the number of circular and noncircular signals in a given signal subspace. The latter problem can be combined with the detection of the dimension of the signal subspace so it becomes a simultaneous

order and model selection problem. For either problem, we show that circular models should be preferred over noncircular models when the SNR is low, the number of samples is small, or the signals' degree of noncircularity is low.

A. Circularity coefficients

Throughout this section, we drop the subscripts on covariance matrices for economy. We first derive a maximal invariant (a complete set of invariants) for the augmented covariance matrix $\underline{\mathbf{C}}$ under nonsingular strictly linear transformation. Such a set is given by the *canonical correlations* between $\mathbf{x}$ and its conjugate $\mathbf{x}^*$ [112], which [41] calls the *circularity coefficients* of $\mathbf{x}$. "Maximal invariant" means two things: (1) The circularity coefficients are invariant under nonsingular linear transformation; (2) If two jointly Gaussian random vectors $\mathbf{x}$ and $\mathbf{y}$ have the same circularity coefficients, then $\mathbf{x}$ and $\mathbf{y}$ are related by a nonsingular linear transformation, $\mathbf{x} = \mathbf{M}\mathbf{y}$.

Assuming $\mathbf{C}$ has full rank, the canonical correlations between $\mathbf{x}$ and $\mathbf{x}^*$ are determined by starting with the coherence

matrix [116]

$$\mathbf{R} = \mathbf{C}^{-1/2}\widetilde{\mathbf{C}}(\mathbf{C}^*)^{-H/2} = \mathbf{C}^{-1/2}\widetilde{\mathbf{C}}\mathbf{C}^{-T/2}. \quad (27)$$

Since $\mathbf{R}$ is complex symmetric, $\mathbf{R} = \mathbf{R}^T$, yet not Hermitian symmetric, *i.e.*, $\mathbf{R} \neq \mathbf{R}^H$, there exists a special singular value decomposition (SVD), called the *Takagi factorization* [53], which is

$$\mathbf{R} = \mathbf{F}\mathbf{K}\mathbf{F}^T. \quad (28)$$

The complex matrix $\mathbf{F}$ is unitary, and $\mathbf{K} = \mathrm{diag}(k_1, k_2, ..., k_N)$ contains the canonical correlations $1 \geq k_1 \geq k_2 \geq \cdots \geq k_N \geq 0$ on its diagonal. The squared canonical correlations k_n^2 are the eigenvalues of the squared coherence matrix $\mathbf{R}\mathbf{R}^H = \mathbf{C}^{-1/2}\widetilde{\mathbf{C}}\mathbf{C}^{-*}\widetilde{\mathbf{C}}^*\mathbf{C}^{-H/2}$, or equivalently, of the matrix $\mathbf{C}^{-1}\widetilde{\mathbf{C}}\mathbf{C}^{-*}\widetilde{\mathbf{C}}^*$ [116]. The canonical correlations between $\mathbf{x}$ and $\mathbf{x}^*$ are invariant to the choice of a square root for $\mathbf{C}$, and they are a maximal invariant for $\underline{\mathbf{C}}$ under nonsingular strictly linear transformation of $\mathbf{x}$. Therefore, any function of $\underline{\mathbf{C}}$ that is invariant under nonsingular strictly linear transformation *must* be a function of these canonical correlations *only* [112].

Following [41], we call these canonical correlations k_n the *circularity coefficients*, and the set $\{k_n\}_{n=1}^N$ the *circularity spectrum* of $\mathbf{x}$. However, the term "circularity coefficient" is not entirely accurate as the circularity coefficients only characterize *second-order* circularity, or (im-)propriety. Thus, the name *impropriety coefficients* would have been more suitable. For a scalar random variable x, there is only one circularity coefficient $k_1 = |E\{x^2\}|/E\{|x|^2\}$. For Gaussian x, k_1 characterizes the *degree of noncircularity*, and for non-Gaussian x, the *degree of impropriety*.

Differential entropy[1]: The entropy of a complex random vector $\mathbf{x}$ is defined to be the entropy of the real composite vector $\mathbf{x}_{\mathbb{R}}$. The differential entropy of a complex Gaussian random vector $\mathbf{x}$ with augmented covariance matrix $\underline{\mathbf{C}}$ is thus

$$H(\mathbf{x}) = \tfrac{1}{2}\log\left[(\pi e)^{2N}\det\underline{\mathbf{C}}\right].$$

Combining our results so far, we may factor $\underline{\mathbf{C}}$ as

$$\underline{\mathbf{C}} = \begin{bmatrix} \mathbf{C} & \widetilde{\mathbf{C}} \\ \widetilde{\mathbf{C}}^* & \mathbf{C}^* \end{bmatrix} = \begin{bmatrix} \mathbf{C}^{1/2} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}^{*/2} \end{bmatrix} \begin{bmatrix} \mathbf{F} & \mathbf{0} \\ \mathbf{0} & \mathbf{F}^* \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{K} \\ \mathbf{K} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{F}^H & \mathbf{0} \\ \mathbf{0} & \mathbf{F}^T \end{bmatrix} \begin{bmatrix} \mathbf{C}^{H/2} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}^{T/2} \end{bmatrix} \quad (29)$$

Note that each factor is an augmented matrix. This factorization establishes

$$\det\underline{\mathbf{C}} = \det{}^2\mathbf{C}\,\det(\mathbf{I} - \mathbf{K}\mathbf{K}^H) = \det{}^2\mathbf{C}\prod_{n=1}^N(1 - k_n^2). \quad (30)$$

This allows us to write the entropy of a complex noncircular Gaussian random vector $\mathbf{x}$ as [41], [112]

$$H_{\text{noncircular}} = \tfrac{1}{2}\log[(\pi e)^{2N}\det\underline{\mathbf{C}}]$$
$$= \underbrace{\log[(\pi e)^N\det\mathbf{C}]}_{H_{\text{circular}}} + \underbrace{\tfrac{1}{2}\log\prod_{n=1}^N(1 - k_n^2)}_{\leq 0}, \quad (31)$$

---

[1] Since we do not consider discrete random variables in this paper, we refer to differential entropy simply as entropy from now on.

where H_{circular} is the entropy of a circular Gaussian random vector with the same Hermitian covariance matrix $\mathbf{C}$ (but $\widetilde{\mathbf{C}} = \mathbf{0}$). The entropy $H_{\text{noncircular}}$ is maximized if and only if $\mathbf{x}$ is circular. If $\mathbf{x}$ is noncircular, the loss in entropy compared to the circular case is given by the second term in (31), which is a function of the circularity spectrum. Thus, this term can be used as a measure for the degree of impropriety of a random vector $\mathbf{x}$, and it is also a test statistic in the generalized likelihood ratio test for noncircularity. Other measures for the degree of impropriety have been proposed and studied by [112].

B. Testing for circularity

In this section, we present hypothesis tests for circularity based on a generalized likelihood ratio test (GLRT). In a GLR, the unknown parameters ($\mathbf{C}$ and $\widetilde{\mathbf{C}}$ in our case) are replaced by maximum likelihood estimates. The GLR is always invariant to transformations for which the hypothesis testing problem itself is invariant [58]. As propriety is preserved by strictly linear, but not widely linear, transformations, the hypothesis test must be invariant to strictly linear, but not widely linear, transformations. Since the GLR must be a function of a maximal invariant statistic the GLR is a function of the circularity coefficients.

Let $\mathbf{x}$ be a complex zero-mean Gaussian random vector with augmented covariance matrix $\underline{\mathbf{C}}$. Consider L independent and identically distributed (i.i.d.) random samples drawn from this distribution and arranged in the sample matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_L]$, and let $\underline{\mathbf{X}} = [\underline{\mathbf{x}}_1, \underline{\mathbf{x}}_2, ..., \underline{\mathbf{x}}_L]$ denote the corresponding augmented sample matrix. The joint probability density function of these samples is

$$p(\mathbf{X}) = \pi^{-LN}(\det\underline{\mathbf{C}})^{-L/2}\exp\left\{-\frac{L}{2}\mathrm{tr}\left(\underline{\mathbf{C}}^{-1}\widehat{\underline{\mathbf{C}}}\right)\right\}, \quad (32)$$

with augmented sample covariance matrix

$$\widehat{\underline{\mathbf{C}}} = \begin{bmatrix} \widehat{\mathbf{C}} & \widehat{\widetilde{\mathbf{C}}} \\ \widehat{\widetilde{\mathbf{C}}}^* & \widehat{\mathbf{C}}^* \end{bmatrix} = \frac{1}{L}\sum_{l=1}^L \underline{\mathbf{x}}_l\underline{\mathbf{x}}_l^H.$$

The GLR test of the hypotheses $H_0 : \mathbf{x}$ is circular ($\widetilde{\mathbf{C}} = \mathbf{0}$) vs. $H_1 : \mathbf{x}$ is noncircular ($\widetilde{\mathbf{C}} \neq \mathbf{0}$) compares the GLRT statistic

$$\lambda = \frac{\max_{\substack{\underline{\mathbf{C}} \\ \widetilde{\mathbf{C}}=\mathbf{0}}} p(\mathbf{X})}{\max_{\underline{\mathbf{C}}} p(\mathbf{X})} \quad (33)$$

to a threshold. This statistic is the ratio of likelihood with $\underline{\mathbf{C}}$ constrained to have zero off-diagonal blocks, $\widetilde{\mathbf{C}} = \mathbf{0}$, to likelihood with $\underline{\mathbf{C}}$ unconstrained. We are thus testing whether or not $\underline{\mathbf{C}}$ is block-diagonal. The unconstrained maximum likelihood (ML) estimate of $\underline{\mathbf{C}}$ is the augmented sample covariance matrix $\widehat{\underline{\mathbf{C}}}$. The ML estimate of $\underline{\mathbf{C}}$ under the constraint $\widetilde{\mathbf{C}} = \mathbf{0}$ is

$$\widehat{\underline{\mathbf{C}}}_0 = \begin{bmatrix} \widehat{\mathbf{C}} & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{C}}^* \end{bmatrix}.$$

After a little algebra, the GLR can be expressed as [94], [116]

$$\ell = \lambda^{\frac{2}{L}} = \frac{\det \widehat{\mathbf{C}}}{(\det \hat{\mathbf{C}})^2} = \prod_{n=1}^{N} (1 - \hat{k}_n^2). \qquad (34)$$

In this equation, $\hat{k}_n$ denotes the estimated circularity coefficients, which are computed from the augmented sample covariance matrix $\widehat{\mathbf{C}}$. A full-rank implementation of this test relies on the middle expression in (34). A reduced-rank implementation that considers only the r largest estimated circularity coefficients is based on the last expression in (34).

This test was first proposed by [13], in complex notation by [94], and the connection with canonical correlations was established by [116]. It is shown in [13] that

$$\ell' = \frac{1}{N} \sum_{n=1}^{N} \hat{k}_n^2$$

rather than ℓ is the *locally most powerful* (LMP) test for circularity. An LMP test has the highest possible power (*i.e.*, probability of detection) for H_1 close to H_0, where all circularity coefficients are small. Testing for circularity is reexamined by [133], which studies the null distributions of ℓ and ℓ', and derives a distributional approximation for ℓ. It also shows that no uniformly powerful (UMP) test exists for this problem because the GLR and the LMP tests are different for dimensions $N \geq 2$. (In the scalar case $N = 1$, the GLR and LMP tests are identical.)

Extensions of this test to non-Gaussian data: There have been recent extensions of this test to non-Gaussian data. In [95], the test is made asymptotically robust with respect to violations of Gaussianity, provided the distribution remains elliptically symmetric. This is achieved by dividing the GLRT statistic by an estimated normalized fourth-order moment (which is closely related to the kurtosis). Generation and estimation of samples from a complex generalized Gaussian distribution (GGD) and estimation of its parameters is addressed in [90] and a circularity test for the complex GGD is given in [89]. These results have been recently extended to elliptically symmetric distributions [93].

The complex GGD is a complex elliptically symmetric distribution with a simple probabilistic characterization. It allows for a direct connection to the Gaussian test given in (34). If x is a scalar complex GGD random variable, it has pdf

$$p(x) = \frac{\beta(c)}{\sqrt{\det(\underline{\mathbf{C}})}} \exp \left\{ - \left[\eta(c)(\underline{\mathbf{x}}^H \underline{\mathbf{C}}^{-1} \underline{\mathbf{x}}) \right]^c \right\} \qquad (35)$$

where $c > 0$ is the shape parameter, $\beta(c) = \frac{c\Gamma(2/c)}{\pi\Gamma(1/c)^2}$, $\eta(c) = \frac{\Gamma(2/c)}{2\Gamma(1/c)}$, $\Gamma(\cdot)$ is the Gamma function, $\underline{\mathbf{x}} = [x, x^*]^T$, and $\underline{\mathbf{C}}$ is the augmented covariance matrix. The complex GGD comprises a large number of symmetric distributions, from super-Gaussian (for $0 < c < 1$), Gaussian (for $c = 1$), to sub-Gaussian (for $c > 1$), including other distributions such as the bivariate Laplacian. As for all complex elliptically symmetric distributions, zero mean and propriety is equivalent to circularity.

A GLRT statistic based on the complex GGD, as derived by [89], is given by

$$\log \lambda_{\mathrm{GGD}} = (L/2) \log \det \left(\widehat{\mathbf{C}} \widehat{\mathbf{C}}_0^{-1} \right) +$$

$$\eta^{\hat{c}}(\hat{c}) \sum_{l=1}^{L} \left[\left(\underline{\mathbf{x}}_l^H \widehat{\underline{\mathbf{C}}}^{-1} \underline{\mathbf{x}}_l \right)^{\hat{c}} - \left(\underline{\mathbf{x}}_l^H \widehat{\underline{\mathbf{C}}}_0^{-1} \underline{\mathbf{x}}_l \right)^{\hat{c}} \right]. \quad (36)$$

Here, $\hat{c}$ is the ML estimate of the shape parameter c, and $\widehat{\underline{\mathbf{C}}}_0$ and $\widehat{\underline{\mathbf{C}}}$ are the ML estimates of the augmented covariance matrix, under H_0 and H_1, respectively. These ML estimates have been derived in [90]. Asymptotically for $L \to \infty$, the statistic λ_{GGD} is χ^2-distributed with two degrees of freedom. This allows choosing a threshold to yield a specified probability of false alarm. In the scalar Gaussian case, $c = 1$ and $N = 1$, the last term in (36) vanishes, and the complex GGD and Gaussian GLRT become equivalent. Using the complex GGD model and the ML estimators for its parameters, it is also possible to design a test for Gaussianity [89].

To compare the performance of the different circularity detectors, we generate $L = 500$ independent realizations from a complex GGD with unit variance $E\{|x|^2\} = 1$, as in [89][2]. Figures 6(a)–6(c) show the probability of detection versus the degree of impropriety $|\rho| = |E\{x^2\}|$ for the three detectors: the Gauss-GLRT in (34), the adjusted GLRT (Adj-GLRT) [95], and the complex GGD-GLRT in (36). Each data point in the figures is the result of the average of 1000 runs, with the detection threshold set for a probability of false alarm of 0.01. As can be observed in the figures, all three detectors yield similar performance for sub-Gaussian and identical performance for Gaussian data. So the Gauss-GLRT detector performs well even with sub-Gaussian data. When the samples are super-Gaussian, however, the complex GGD-GLRT significantly outperforms. In [89], examples are presented to show that the circularity test based on the complex GGD model provides good performance even with data that are not complex GGD, such as BPSK data.

C. Order selection with a general signal subspace model

Determining the effective order of the signal subspace is an important problem in many signal processing and communications applications. The popular solution proposed by Wax and Kailath [134] uses information theoretic criteria by considering principal component analysis (PCA) of the observed data to select the order. However, the model assumes circular multivariate Gaussian signals, and is suboptimal in the presence of noncircular signals in the subspace [73].

In [73], a noncircular PCA (ncPCA) approach is proposed along with an ML procedure for estimating the free parameters in the model. The procedure can be used for model selection in the presence of circular Gaussian noise, whereby the numbers of circular and noncircular signals are determined together with the total model order. The method reduces to Wax and Kailath's signal detection method [134] when all signals are circular, but may provide a significant performance gain in the

---

[2]Matlab code to implement the detectors and to generate complex GGD samples can be found at http://mlsp.umbc.edu/resources.

(a) Sub-Gaussian with $c = 1.5$

(b) Gaussian ($c = 1$)

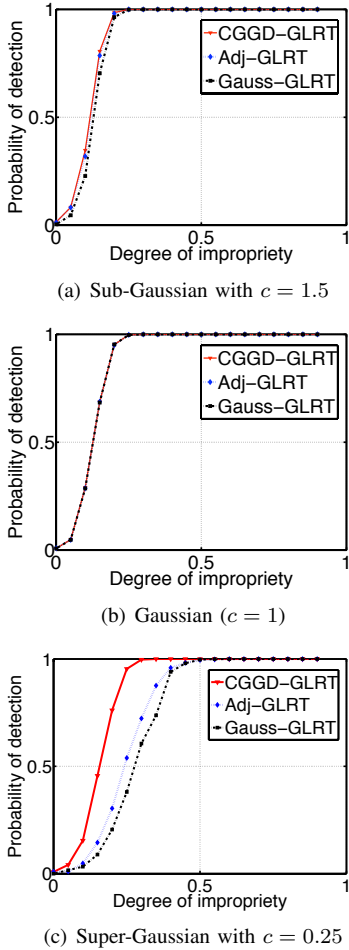(c) Super-Gaussian with $c = 0.25$

Fig. 6. Detection performance of circularity detectors versus degree of impropriety $|\rho|$, for $L = 500$ sub-Gaussian, Gaussian, and super-Gaussian complex GGD samples. Probability of false alarm is fixed at 0.01.

presence of noncircular signals. We first introduce the model for ncPCA and then demonstrate its use for model selection.

Given a set of observations $\mathbf{x}(l) = [x_1(l), \ldots, x_N(l)]^T$, with l as the observation index, we assume the linear model

$$\mathbf{x}(l) = \mathbf{A}\mathbf{s}(l) + \mathbf{n}(l), \quad l = 1, \ldots, L, \qquad (37)$$

where $\mathbf{s}(l) = [s_1(l), \ldots, s_M(l)]^T$ is the $M \times 1$ complex-valued signal vector, $\mathbf{A}$ is an $N \times M$ complex-valued matrix with full column rank, $0 \leq M < N$, and $\mathbf{n}(l)$ is the complex-valued $N \times 1$ random vector modeling the additive noise. Given the set of observations $\mathbf{x}(l)$, $l = 1, \ldots, L$, a key step in many applications is to determine the dimension M of the signal subspace. Traditional PCA provides a decomposition into signal and noise subspaces. In ncPCA, the M-dimensional signal subspace is further decomposed into noncircular and circular signals. As in [134], the noise term $\mathbf{n}(l)$ is assumed to be a circular, isotropic, stationary, complex Gaussian random process with zero mean, statistically independent of the signals, and the signals are assumed to be multivariate Gaussian. However, in contrast to [134], we let p signals be noncircular and $q = M - p$ signals be circular. The underlying assumption here is that the rank of the source covariance matrix $E\{\mathbf{s}(l)\mathbf{s}^H(l)\}$ is M and that of the complementary

covariance matrix $E\{\mathbf{s}(l)\mathbf{s}^T(l)\}$ is p. For the degenerate case, the actual number of the signals and those that are noncircular can be greater than M and/or p.

In order to determine the orders M and p, given L snapshots of $\mathbf{x}$, $\mathbf{X} \triangleq [\mathbf{x}(1), \ldots, \mathbf{x}(L)]$, we write the likelihood of $\mathbf{X}$ as

$$\ell(\mathbf{X}|\boldsymbol{\theta}) = \prod_{l=1}^{L} p[\mathbf{x}(l)|\boldsymbol{\theta}], \qquad (38)$$

where $\boldsymbol{\theta}$ denotes the set of all adjustable parameters in the likelihood function

$$p[\mathbf{x}(l)|\boldsymbol{\theta}] = \frac{1}{\pi^N \det^{1/2}(\underline{\mathbf{C}})} \exp\left\{ -\frac{\underline{\mathbf{x}}^H(l)\underline{\mathbf{C}}^{-1}\underline{\mathbf{x}}(l)}{2} \right\}.$$

For the model in (37), the covariance and complementary covariance matrices of $\mathbf{x}$ have parametric forms [72], [73]

$$\mathbf{C} = \mathbf{U}(\boldsymbol{\Lambda}^2 - \sigma^2\mathbf{I})\mathbf{U}^H + \sigma^2\mathbf{I}, \quad \widetilde{\mathbf{C}} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}\mathbf{K}'\mathbf{V}^T\boldsymbol{\Lambda}\mathbf{U}^T, \qquad (39)$$

where $\mathbf{U}$ is a complex-valued $N \times M$ matrix with orthonormal columns spanning the signal subspace, $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_M)$ is a diagonal matrix with diagonal elements $\lambda_m > \sigma^2$, σ^2 is the noise variance, $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_p]$ is an $M \times p$ complex-valued matrix with orthonormal columns, and $\mathbf{K}' = \operatorname{diag}(k_1', \ldots, k_p')$ is a $p \times p$ diagonal matrix with complex-valued entries k_i'. The absolute values of these entries equal the circularity coefficients: $|k_i'| = k_i$, $i = 1, \ldots, p$. Compared to the Takagi factorization in (28), which leads to nonnegative circularity coefficients, the decomposition in (39) incorporates an additional phase factor into the quantities k_i', making them complex valued. This is simply a notational convenience [73].

Given the ML estimates for all the free parameters $\boldsymbol{\theta} = \{\mathbf{U}, \mathbf{V}, \boldsymbol{\Lambda}, \mathbf{K}', \sigma^2\}$ in (38), we may follow [134] and select M and p using information-theoretic criteria such as Akaike's information criterion (AIC) [8], the Bayesian information criterion (BIC) [118], or the minimum description length (MDL) [105]. This leads to the estimated orders

$$(\widehat{M}, \hat{p}) = \arg\min_{(M,p)} \left[-\log\ell(\mathbf{X}|\boldsymbol{\theta}_0) + r(\boldsymbol{\theta})\eta(L) \right] \qquad (40)$$

where $\boldsymbol{\theta}_0 = \{\mathbf{U}_0, \mathbf{V}_0, \boldsymbol{\Lambda}_0, \mathbf{K}_0', \sigma_0^2\}$ is the optimal $\boldsymbol{\theta}$ that maximizes $\ell(\mathbf{X}|\boldsymbol{\theta})$, which can be obtained as in [72], and results in the likelihood

$$\frac{\log\ell(\mathbf{X}|\boldsymbol{\theta}_0)}{L} = (N - M)\log\sigma_0^2 + \sum_{m=1}^{M} \log\lambda_{0,m}^2$$
$$+ 0.5\sum_{l=1}^{p} \log\left(1 - |k_{0,l}'|^2\right) + N\log(e\pi). \qquad (41)$$

The penalty term $r(\boldsymbol{\theta})\eta(L)$ in (40) contains the degrees of freedom of $\boldsymbol{\theta}$,

$$r(\boldsymbol{\theta}) = M(2N - M) + p(2M - p + 1) + 1,$$

and $\eta(L)$, which depends on the number of samples and the chosen criterion. For example, in the BIC (or the MDL criterion) [105], [118], $\eta(L) = 0.5\log L$.

The signal detection method given in [134] can be obtained as a special case without noncircular ($p = 0$) sources. In the
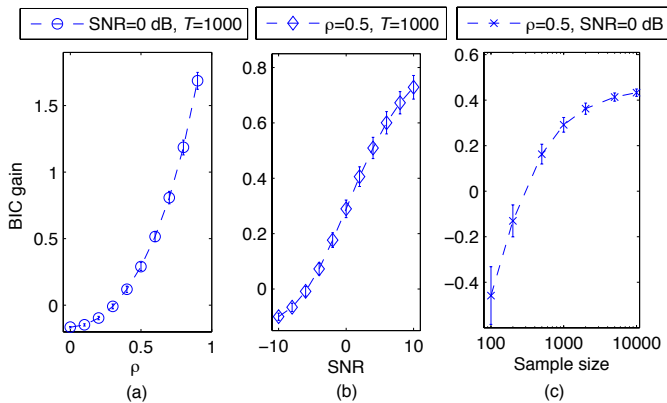
Fig. 7. BIC gain of noncircular over circular model for (a) varying degree of noncircularity, (b) varying SNR, and (c) varying sample size. Each simulation point is averaged over 1000 independent runs. A positive gain suggests that the noncircular model is preferred over the circular model.

Fig. 8. Comparison of probability of detection (Pd) and subspace distance gain for (a) ncPCA and (b) circular PCA (cPCA), as a function of SNR.

presence of noncircular signals, the presented approach will lead to a smaller $-\log\ell(\mathbf{X}|\boldsymbol{\theta}_0)$ term in (40). At the same time, however, a noncircular model has more degrees of freedom, so the penalty term $r(\boldsymbol{\theta})\eta(L)$ in (40) will increase. This requires the right tradeoff, resulting in a good model fit without overfitting. The following example demonstrates this tradeoff and shows that a circular model can be preferable when the noise level is high, the degree of noncircularity is low, and/or the number of samples is small.

In this example, $M = 7$ Gaussian sources of unit variance $E\{|s|^2\} = 1$ and identical degree of noncircularity $|\rho| = |E\{s^2\}|$ are mixed through a randomly chosen 20×7 matrix, after which circular Gaussian noise is added to the mixture. We study the BIC of a circular model with orders $(M, p) = (7, 0)$, and a noncircular model with orders $M = p = 7$. The gain of the noncircular model over the circular model is defined as (BIC of circular model)$/L$ minus (BIC of noncircular model)$/L$. It is clear that a positive gain suggests that the noncircular model is preferred over the circular model.

Figure 7 shows the overall information-theoretic (BIC) gain of using a noncircular model for varying degree of noncircularity, SNR, and sample size. The results are averaged over 1000 independent runs. We observe that the noncircular model is preferred when there is ample evidence that the signals are noncircular: the cases of large degree of noncircularity, high SNR, and large sample sizes. On the other hand, the simpler circular model is preferred when there is scarce evidence that the signals are noncircular: the cases where the degree of noncircularity is low, the SNR is low, or the number of samples is small. The ncPCA approach can model both circular and noncircular signals and hence can avoid the use of an unnecessarily complex noncircular model when a circular model should be preferred.

We also give an example in array signal processing to demonstrate the direct performance gain using ncPCA rather than a circular model [134] for signal subspace estimation and model selection. We model $M = 3$ far-field, independent narrowband sources (one BPSK signal, one QPSK signal and one 8-QAM signal, with degrees of impropriety of 1,
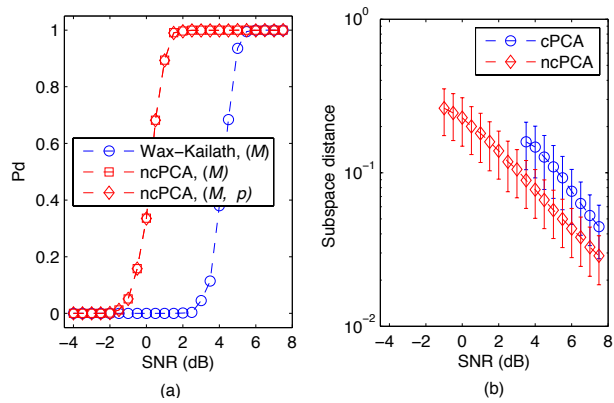
0, and 2/3, and variances 1, 2, and 6 respectively) emitting plane waves impinging upon a uniform linear array of $N = 10$ sensors with half-wavelength inter-sensor spacing. The received observations are $\mathbf{x}(l) = \sum_{m=1}^{M} s_m(l)\mathbf{a}_m + \mathbf{n}(l)$, where l is the snapshot index, $s_m(l)$ is the waveform of the mth source, $\mathbf{n}(l)$ the circular antenna noise, $\mathbf{a}_m = [1, \exp(j\pi\sin(\theta_m)), \ldots, \exp(j\pi(N-1)\sin(\theta_m))]^T$ is the steering vector associated with the mth source, and θ_m is the direction-of-arrival (DOA) of the mth source. We set $\theta_1 = -\theta_3 = 5°$ and $\theta_2 = 0$.

In Figure 8, we show the gain using ncPCA compared to a circular model [134] both in terms of probability of detection and in terms of subspace distance. Probability of detection is defined as the fraction of trials where the order was detected correctly, and substance distance is the squared Euclidean distance between the estimated and the true signal subspace. Each simulation point is averaged over 1000 independent runs. As observed in Figure 8, ncPCA outperforms circular PCA by approximately 4.0 dB in SNR. This holds for both the detection of the order M alone and the joint detection of orders (M, p), which perform almost identically for different SNR levels. In addition, ncPCA consistently leads to a smaller subspace distance. Further examples and additional discussion on the performance of ncPCA are given in [72].

IV. WIDELY LINEAR ESTIMATION

In Section II-B, we have discussed widely linear (linear-conjugate linear) transformations, which allow access to the information contained in the complementary covariance. In this section, we consider widely linear minimum mean-squared error (WLMMSE) estimation [103], widely linear minimum variance distortionless response (WLMVDR) estimation [29], [77], [26], [27], and the widely linear least-mean-square (LMS) algorithm. Using the augmented vector and matrix notation, many of the results for widely linear estimation are straightforward extensions of the corresponding results for linear estimation.

A. Widely Linear MMSE Estimation

We begin with WLMMSE estimation of the N-dimensional message (or signal) $\mathbf{x}$ from the M-dimensional measurement

$\mathbf{y}$. To extend results for LMMSE estimation to WLMMSE estimation we need only replace the signal $\mathbf{x}$ by the augmented signal $\underline{\mathbf{x}}$ and the measurement $\mathbf{y}$ by the augmented measurement $\underline{\mathbf{y}}$, and proceed as usual. Thus, most of the results for LMMSE estimation apply straightforwardly to WLMMSE estimation. It is, however, still worthwhile to summarize some of these results and to compare LMMSE with WLMMSE estimation. The widely linear estimator is

$$\hat{\underline{\mathbf{x}}} = \underline{\mathbf{W}}\,\underline{\mathbf{y}} \iff \hat{\mathbf{x}} = \mathbf{W}_1 \mathbf{y} + \mathbf{W}_2 \mathbf{y}^*, \qquad (42)$$

where

$$\underline{\mathbf{W}} = \begin{bmatrix} \mathbf{W}_1 & \mathbf{W}_2 \\ \mathbf{W}_2^* & \mathbf{W}_1^* \end{bmatrix}$$

is determined such that the mean-squared error $E\{\|\hat{\mathbf{x}} - \mathbf{x}\|^2\} = \frac{1}{2} E\{\|\hat{\underline{\mathbf{x}}} - \underline{\mathbf{x}}\|^2\}$ is minimized. Keep in mind that the augmented notation on the left hand side of (42) is simply a convenient, but redundant, representation of the right hand side of (42). Obviously, it is sufficient to estimate $\mathbf{x}$ because $\hat{\mathbf{x}}^*$ can be obtained from $\hat{\mathbf{x}}$ through conjugation.

The WLMMSE estimator is found by applying the *orthogonality principle* $(\hat{\mathbf{x}} - \mathbf{x}) \perp \mathbf{y}$ and $(\hat{\mathbf{x}} - \mathbf{x}) \perp \mathbf{y}^*$ [103], or equivalently, $(\hat{\underline{\mathbf{x}}} - \underline{\mathbf{x}}) \perp \underline{\mathbf{y}}$. This says that the error between the augmented estimator and the augmented signal must be orthogonal to the augmented measurement. This leads to $E\{\hat{\underline{\mathbf{x}}}\,\underline{\mathbf{y}}^H\} - E\{\underline{\mathbf{x}}\,\underline{\mathbf{y}}^H\} = \mathbf{0}$ and thus

$$\underline{\mathbf{W}}\,\underline{\mathbf{C}}_{yy} - \underline{\mathbf{C}}_{xy} = \mathbf{0} \Leftrightarrow \underline{\mathbf{W}} = \underline{\mathbf{C}}_{xy}\underline{\mathbf{C}}_{yy}^{-1}. \qquad (43)$$

Thus, $\hat{\underline{\mathbf{x}}} = \underline{\mathbf{C}}_{xy}\underline{\mathbf{C}}_{yy}^{-1}\underline{\mathbf{y}}$, or equivalently, [103]

$$\hat{\mathbf{x}} = (\mathbf{C}_{xy} - \widetilde{\mathbf{C}}_{xy}\mathbf{C}_{yy}^{-*}\widetilde{\mathbf{C}}_{yy}^*)\mathbf{P}_{yy}^{-1}\mathbf{y} + (\widetilde{\mathbf{C}}_{xy} - \mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\widetilde{\mathbf{C}}_{yy})\mathbf{P}_{yy}^{-*}\mathbf{y}^*.$$

In this equation, the Schur complement $\mathbf{P}_{yy} = \mathbf{C}_{yy} - \widetilde{\mathbf{C}}_{yy}\mathbf{C}_{yy}^{-*}\widetilde{\mathbf{C}}_{yy}^*$ is the error covariance matrix for linearly estimating $\mathbf{y}$ from $\mathbf{y}^*$. The augmented error covariance matrix $\underline{\mathbf{Q}}$ of the error vector $\underline{\mathbf{e}} = \hat{\underline{\mathbf{x}}} - \underline{\mathbf{x}}$ is

$$\underline{\mathbf{Q}} = E\{\underline{\mathbf{e}}\,\underline{\mathbf{e}}^H\} = \underline{\mathbf{C}}_{xx} - \underline{\mathbf{C}}_{xy}\underline{\mathbf{C}}_{yy}^{-1}\underline{\mathbf{C}}_{xy}^H.$$

A competing estimator $\hat{\underline{\mathbf{x}}}' = \underline{\mathbf{W}}'\underline{\mathbf{y}}$ will produce an augmented error $\underline{\mathbf{e}}' = \hat{\underline{\mathbf{x}}}' - \underline{\mathbf{x}}$ with covariance matrix

$$\underline{\mathbf{Q}}' = E\left\{\underline{\mathbf{e}}'\underline{\mathbf{e}}'^H\right\} = \underline{\mathbf{Q}} + (\underline{\mathbf{W}} - \underline{\mathbf{W}}')\underline{\mathbf{C}}_{yy}(\underline{\mathbf{W}} - \underline{\mathbf{W}}')^H, \quad (44)$$

which shows $\underline{\mathbf{Q}} \le \underline{\mathbf{Q}}'$. As a consequence, all real-valued increasing functions of $\underline{\mathbf{Q}}$ are minimized, in particular, $E\{\|\underline{\mathbf{e}}\|^2\} = \operatorname{tr}\underline{\mathbf{Q}} \le \operatorname{tr}\underline{\mathbf{Q}}' = E\{\|\underline{\mathbf{e}}'\|^2\}$ and $\det\underline{\mathbf{Q}} \le \det\underline{\mathbf{Q}}'$. These statements hold for the error vector $\mathbf{e}$ as well as the augmented error vector $\underline{\mathbf{e}}$ because $\underline{\mathbf{Q}} \le \underline{\mathbf{Q}}' \Rightarrow \mathbf{Q} \le \mathbf{Q}'$. The error covariance matrix $\mathbf{Q}$ of the error vector $\mathbf{e} = \hat{\mathbf{x}} - \mathbf{x}$ is the NW block of the augmented error covariance matrix $\underline{\mathbf{Q}}$, which can be evaluated as

$$\mathbf{Q} = E\{\mathbf{e}\mathbf{e}^H\} = \mathbf{C}_{xx} - (\mathbf{C}_{xy} - \widetilde{\mathbf{C}}_{xy}\mathbf{C}_{yy}^{-*}\widetilde{\mathbf{C}}_{yy}^*)\mathbf{P}_{yy}^{-1}\mathbf{C}_{xy}^H - (\widetilde{\mathbf{C}}_{xy} - \mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\widetilde{\mathbf{C}}_{yy})\mathbf{P}_{yy}^{-*}\widetilde{\mathbf{C}}_{xy}^H. \qquad (45)$$

A particular choice for a generally suboptimum filter is the LMMSE filter

$$\underline{\mathbf{W}}' = \begin{bmatrix} \mathbf{C}_{xy}\mathbf{C}_{yy}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{xy}^*\mathbf{C}_{yy}^{-*} \end{bmatrix} \iff \mathbf{W}' = \mathbf{C}_{xy}\mathbf{C}_{yy}^{-1},$$

which ignores complementary covariance matrices.

Special cases: If the signal $\mathbf{x}$ is *real*, we have $\widetilde{\mathbf{C}}_{xy} = \mathbf{C}_{xy}^*$. This leads to the simplified expression

$$\hat{\mathbf{x}} = 2\operatorname{Re}\left\{(\mathbf{C}_{xy} - \widetilde{\mathbf{C}}_{xy}\mathbf{C}_{yy}^{-*}\widetilde{\mathbf{C}}_{yy}^*)\mathbf{P}_{yy}^{-1}\mathbf{y}\right\}.$$

While the WLMMSE estimate of a real signal from a complex signal is always real [103], the LMMSE estimate is generally complex.

The WLMMSE and LMMSE estimates are identical if and only if the error of the LMMSE estimate is orthogonal to $\mathbf{y}^*$, i.e.,

$$(\mathbf{W}'\mathbf{y} - \mathbf{x}) \perp \mathbf{y}^* \iff \mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\widetilde{\mathbf{C}}_{yy} - \widetilde{\mathbf{C}}_{xy} = \mathbf{0}. \qquad (46)$$

There are two important special cases where (46) holds:

- The signal and measurement are cross-proper, *i.e.*, $\widetilde{\mathbf{C}}_{xy} = \mathbf{0}$, and the measurement is proper, $\widetilde{\mathbf{C}}_{yy} = \mathbf{0}$. Joint propriety of $\mathbf{x}$ and $\mathbf{y}$ will suffice but it is not necessary that $\mathbf{x}$ be proper.
- The measurement is *maximally improper*, *i.e.*, $\mathbf{y} = \alpha\mathbf{y}^*$ with probability 1 for constant α with $|\alpha| = 1$. In this case, $\widetilde{\mathbf{C}}_{xy} = \alpha\mathbf{C}_{xy}$ and $\widetilde{\mathbf{C}}_{yy} = \alpha\mathbf{C}_{yy}$ and $\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{yy}\alpha - \mathbf{C}_{xy}\alpha = \mathbf{0}$. WL estimation is unnecessary since $\mathbf{y}$ and $\mathbf{y}^*$ both carry exactly the same information about $\mathbf{x}$. This is irrespective of whether or not $\mathbf{x}$ is proper.

In these cases, WLMMSE estimation has no performance advantage over LMMSE estimation. The other extreme case is where a WL operation allows perfect estimation while LMMSE estimation yields a nonzero estimation error. An example of such a case is $\mathbf{y} = \mathbf{x} + \mathbf{n}$, where the signal $\mathbf{x}$ is real and the noise $\mathbf{n}$ purely imaginary. Here the WL operation $\operatorname{Re}\{\mathbf{y}\}$ yields a perfect estimate of $\mathbf{x}$, whereas the LMMSE estimate is not even real valued. If $\mathbf{y} = \mathbf{x} + \mathbf{n}$ with proper and white noise $\mathbf{n}$, then the maximum performance advantage of WLMMSE estimation over LMMSE estimation is a factor of 2 [117].

WLMMSE estimation is optimum in the Gaussian case, but may be improved upon for non-Gaussian data if we have access to higher-order statistics. The next logical extension of widely linear processing is to widely linear-quadratic processing [28], [30], [115], which requires statistical information up to fourth order. We should add the cautionary note here that there is no difference between the optimum, generally non-linear, conditional mean estimator $E(\mathbf{x}|\mathbf{y})$, and $E(\mathbf{x}|\mathbf{y}, \mathbf{y}^*)$. Conditioning on $\mathbf{y}$ *and* $\mathbf{y}^*$ changes nothing, since $E(\mathbf{x}|\mathbf{y})$ already extracts all the information there is about $\mathbf{x}$ from $\mathbf{y}$. So the "widely nonlinear" estimator $E(\mathbf{x}|\mathbf{y}, \mathbf{y}^*)$ is simply $E(\mathbf{x}|\mathbf{y})$.

B. Widely Linear MVDR Estimation

Next we extend linear minimum-variance distortionless response (LMVDR) estimators to widely linear MVDR estimators that account for complementary covariance. The measurement model is

$$\mathbf{y} = \mathbf{\Psi}\mathbf{x} + \mathbf{n}, \qquad (47)$$

where the matrix $\mathbf{\Psi} = [\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_p] \in \mathbb{C}^{M \times p}$ consists of p modes that are assumed to be known (as, for instance, in

beamforming), and the vector $\mathbf{x} = [x_1, x_2, ..., x_p]$ consists of p complex deterministic amplitudes. Without loss of generality we assume $\mathbf{\Psi}^H \mathbf{\Psi} = \mathbf{I}$. The noise $\mathbf{n}$ has covariance matrix $\mathbf{C}_{nn} = E\{\mathbf{n}\mathbf{n}^H\}$. Because $\mathbf{x}$ is modeled as unknown but deterministic, the covariance matrix of the measurement equals the covariance matrix of the noise: $\mathbf{C}_{yy} = \mathbf{C}_{nn}$. The matched filter estimator of $\mathbf{x}$ is

$$\hat{\mathbf{x}} = \mathbf{\Psi}^H \mathbf{y} \tag{48}$$

with mean $\mathbf{x}$ and error covariance matrix $E\{(\hat{\mathbf{x}} - \mathbf{x})(\hat{\mathbf{x}} - \mathbf{x})^H\} = \mathbf{\Psi}^H \mathbf{C}_{nn} \mathbf{\Psi}$. The LMVDR estimator $\hat{\mathbf{x}} = \mathbf{W}^H \mathbf{y}$, with $\mathbf{W} \in \mathbb{C}^{M \times p}$, is derived by minimizing the trace of the error covariance under an unbiasedness constraint:

$$\min \mathrm{tr}\left[\mathbf{W}^H \mathbf{C}_{nn} \mathbf{W}\right] \quad \text{under constraint } \mathbf{W}^H \mathbf{\Psi} = \mathbf{I}$$

The solution is

$$\hat{\mathbf{x}} = (\mathbf{\Psi}^H \mathbf{C}_{nn}^{-1} \mathbf{\Psi})^{-1} \mathbf{\Psi}^H \mathbf{C}_{nn}^{-1} \mathbf{y} \tag{49}$$

with mean $\mathbf{x}$ and error covariance matrix $\mathbf{Q} = E\{(\hat{\mathbf{x}} - \mathbf{x})(\hat{\mathbf{x}} - \mathbf{x})^H\} = (\mathbf{\Psi}^H \mathbf{C}_{nn}^{-1} \mathbf{\Psi})^{-1}$. If there is only a single mode $\psi \in \mathbb{C}^{M \times 1}$, then the solution is

$$\hat{x} = \frac{\psi^H \mathbf{C}_{nn}^{-1}}{\psi^H \mathbf{C}_{nn}^{-1} \psi} \mathbf{y}, \tag{50}$$

and the covariance matrix of the noise is $\mathbf{C}_{nn} = E\{\mathbf{y}\mathbf{y}^H\} - |x|^2 \psi\psi^H$.

The measurement model (47) in augmented form is

$$\underline{\mathbf{y}} = \underline{\mathbf{\Psi}}\,\underline{\mathbf{x}} + \underline{\mathbf{n}},$$

where

$$\underline{\mathbf{\Psi}} = \begin{bmatrix} \mathbf{\Psi} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Psi}^* \end{bmatrix}$$

and the noise $\mathbf{n}$ is generally improper with augmented covariance matrix $\underline{\mathbf{C}}_{nn}$. The matched filter estimator of $\mathbf{x}$, however, does not take into account noise, and thus the widely linear matched filter solution is still the linear solution (48). The WLMVDR estimator, on the other hand, is obtained as the solution to

$$\min \mathrm{tr}\left[\underline{\mathbf{W}}^H \underline{\mathbf{C}}_{nn} \underline{\mathbf{W}}\right] \quad \text{under constraint } \underline{\mathbf{W}}^H \underline{\mathbf{\Psi}} = \mathbf{I} \tag{51}$$

with

$$\underline{\mathbf{W}} = \begin{bmatrix} \mathbf{W}_1 & \mathbf{W}_2 \\ \mathbf{W}_2^* & \mathbf{W}_1^* \end{bmatrix}.$$

This solution is widely linear [29], [77], [26],

$$\underline{\hat{\mathbf{x}}} = (\underline{\mathbf{\Psi}}^H \underline{\mathbf{C}}_{nn}^{-1} \underline{\mathbf{\Psi}})^{-1} \underline{\mathbf{\Psi}}^H \underline{\mathbf{C}}_{nn}^{-1} \underline{\mathbf{y}}, \tag{52}$$

with mean $\mathbf{x}$ and augmented error covariance matrix $\underline{\mathbf{Q}} = E\{(\underline{\hat{\mathbf{x}}} - \underline{\mathbf{x}})(\underline{\hat{\mathbf{x}}} - \underline{\mathbf{x}})^H\} = (\underline{\mathbf{\Psi}}^H \underline{\mathbf{C}}_{nn}^{-1} \underline{\mathbf{\Psi}})^{-1}$. The variance of the WLMVDR estimator is less than or equal to the variance of the LMVDR estimator because of the following argument. The optimization (51) is performed under the constraint $\underline{\mathbf{W}}^H \underline{\mathbf{\Psi}} = \mathbf{I}$, or equivalently, $\mathbf{W}_1^H \mathbf{\Psi} = \mathbf{I}$ and $\mathbf{W}_2^H \mathbf{\Psi} = \mathbf{0}$. Thus, the WLMMSE optimization problem contains the LMMSE optimization problem as a special case in which $\mathbf{W}_2 = \mathbf{0}$ is enforced. As such, WLMVDR estimation cannot be worse than LMVDR estimation. However, the additional degree of

freedom of being able to choose $\mathbf{W}_2 \neq \mathbf{0}$ can reduce variance if the noise $\mathbf{n}$ is improper.

The reduction in variance of the WLMVDR compared to the LMVDR estimator is entirely due to exploiting the complementary correlation of the noise $\mathbf{n}$. Indeed it is easy to see that for proper noise $\mathbf{n}$, the WLMVDR solution (52) simplifies to the LMVDR solution (49). Since $\mathbf{x}$ is not assigned statistical properties, the solution is independent of whether or not $\mathbf{x}$ is improper. This stands in marked contrast to WLMMSE estimation discussed in the previous section. A detailed analysis of the WLMVDR estimator is provided by [26]. Further results are presented in [27].

C. Linear and Widely Linear Filtering and the LMS Algorithm

In this section, we consider the linear and widely linear filtering problem, where we estimate a scalar signal $x(k)$ from observations taken over M time instants $\mathbf{y}(k) = [y(k), y(k-1), ..., y(k-M+1)]^T$. We first derive the normal equations and the least-mean-square (LMS) algorithm [135], [136], and then extend it to the widely linear case. The linear estimate of $x(k)$ is

$$\hat{x}(k) = \mathbf{w}^H(k)\mathbf{y}(k).$$

For convenience, we will suppress the time-dependency of the filter $\mathbf{w}(k)$. We could derive the solution for $\mathbf{w}$ using orthogonality arguments as in (43). Alternatively, we can take the Wirtinger derivative of the linear mean-squared error term $J_L(\mathbf{w}) = E\{|e(k)|^2\} = E\{|x(k) - \hat{x}(k)|^2\}$, which is real valued, with respect to $\mathbf{w}^*$ (by treating $\mathbf{w}$ as a constant)

$$
\begin{aligned}
\frac{\partial E\{e(k)e^*(k)\}}{\partial \mathbf{w}^*} &= \frac{\partial E\{[x(k) - \mathbf{w}^H \mathbf{y}(k)][x^*(k) - \mathbf{w}^T \mathbf{y}^*(k)]\}}{\partial \mathbf{w}^*} \\
&= -E\left\{\mathbf{y}(k)[x^*(k) - \mathbf{w}^T \mathbf{y}^*(k)]\right\}. \tag{53}
\end{aligned}
$$

By setting (53) to zero and assuming a nonsingular covariance matrix, we obtain the normal equations

$$E\{\mathbf{y}(k)\mathbf{y}^H(k)\}\mathbf{w}_{\mathrm{opt}} = E\{x^*(k)\mathbf{y}(k)\} \iff \mathbf{w}_{\mathrm{opt}} = \mathbf{C}_{yy}^{-1}(k)\mathbf{c}_{yx}(k),$$

where $\mathbf{c}_{yx}(k) = E\{x^*(k)\mathbf{y}(k)\}$. If $x(k)$ and $y(k)$ are jointly WSS, this equation is independent of k. The weight vector $\mathbf{w}$ can also be computed adaptively using gradient descent updates

$$
\begin{aligned}
\mathbf{w}(k+1) &= \mathbf{w}(k) - \mu \frac{\partial E\{|e(k)|^2\}}{\partial \mathbf{w}^*(k)} \\
&= \mathbf{w}(k) + \mu E\{e^*(k)\mathbf{y}(k)\} \tag{54}
\end{aligned}
$$

or using stochastic gradient updates

$$\mathbf{w}(k+1) = \mathbf{w}(k) + \mu e^*(k)\mathbf{y}(k),$$

which replaces the expected value in (54) with its instantaneous estimate. This is the popular LMS algorithm. The stepsize $\mu > 0$ determines the trade-off between the rate of convergence and the minimum MSE.

Widely linear filter and LMS algorithm: As shown in (42), a widely linear filter forms the estimate of $x(k)$ through the inner product

$$\hat{x}_{WL}(k) = \mathbf{v}^H \underline{\mathbf{y}}(k) \tag{55}$$

where the weight vector $\mathbf{v} = [v_0, v_1, \cdots, v_{2M-1}]^T$ has twice the dimension of the linear filter. The minimization of the widely linear MSE cost J_{WL} is obtained analogously by setting $\partial J_{\mathrm{WL}}(\mathbf{v})/\partial \mathbf{v}^* = \mathbf{0}$. This results in the *widely linear complex normal equation*

$$E\{\underline{\mathbf{y}}(k)\underline{\mathbf{y}}^H(k)\}\mathbf{v}_{\mathrm{opt}} = E\{x^*(k)\underline{\mathbf{y}}(k)\} \iff \mathbf{v}_{\mathrm{opt}} = \underline{\mathbf{C}}_{yy}^{-1}(k)\begin{bmatrix} \mathbf{c}_{yx}(k) \\ \tilde{\mathbf{c}}_{yx}^*(k) \end{bmatrix},$$

where $\tilde{\mathbf{c}}_{yx}^*(k) = E\{x^*(k)\mathbf{y}^*(k)\}$. The difference between the widely linear MSE J_{WL} and the linear MSE J_{L} is [103]

$$\begin{aligned} J_{\mathrm{diff}} &= J_{\mathrm{L}}(\mathbf{w}_{\mathrm{opt}}) - J_{\mathrm{WL}}(\mathbf{v}_{\mathrm{opt}}) \\ &= \left(\tilde{\mathbf{c}}_{yx}^* - \widetilde{\mathbf{C}}_{yy}^*\mathbf{C}_{yy}^{-1}\mathbf{c}_{yx}\right)^H \left(\mathbf{C}_{yy}^* - \widetilde{\mathbf{C}}_{yy}^*\mathbf{C}_{yy}^{-1}\widetilde{\mathbf{C}}_{yy}\right)^{-1}\left(\tilde{\mathbf{c}}_{yx}^* - \widetilde{\mathbf{C}}_{yy}^*\mathbf{c}\right) \end{aligned}$$

The widely linear LMS algorithm is written similar to the linear case as

$$\mathbf{v}(k+1) = \mathbf{v}(k) + \mu e^*(k)\underline{\mathbf{y}}(k) \tag{57}$$

where $\mu > 0$ is the stepsize and $e(k) = x(k) - \mathbf{v}^H(k)\underline{\mathbf{y}}(k)$.

The study of the LMS filter has been an active research topic. A thorough account of its properties is given in [50], [74] based on the different types of assumptions that can be invoked to simplify the analysis. With the augmented vector notation, most of the results for the linear LMS filter can be readily extended to the widely linear case although care must be taken in the assumptions such as uncorrelatedness and orthogonality of the signals [38], [6].
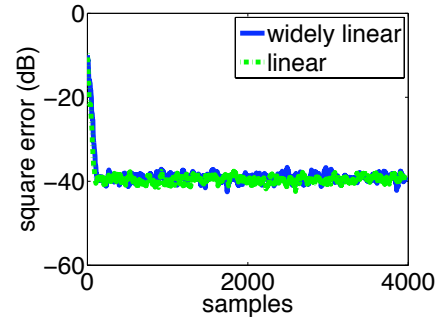
The convergence of the LMS algorithm depends on the eigenvalues of the input covariance matrix [21], [50]. For the widely linear LMS filter, this is the augmented covariance matrix. A measure typically used for the eigenvalue disparity of a given matrix is the condition number. For a Hermitian matrix $\mathbf{C}$, the condition number is the ratio of largest to smallest eigenvalue: $\kappa(\mathbf{C}) = \lambda_{\max}/\lambda_{\min}$. When the signal is proper, the augmented covariance matrix is block-diagonal and has eigenvalues that occur with even multiplicity. In this case, the condition numbers of the augmented covariance matrix $\underline{\mathbf{C}}$ and the Hermitian covariance matrix $\mathbf{C}$ are the same. If the signal is improper, then the eigenvalue spread of $\underline{\mathbf{C}}$ is always greater than the eigenvalue spread of $\mathbf{C}$. The maximally improper case leads to the most spread out eigenvalues. This can be shown via majorization theory [115], [117].

Thus, the MSE performance advantage of the widely linear LMS algorithm for improper signals comes at the price of slower convergence compared to the linear LMS algorithm. An update scheme such as the recursive least squares (RLS) algorithm [50], which is less sensitive to the eigenvalue spread, may be preferable in some cases. In the next example, we demonstrate the impact of impropriety on the convergence of LMS algorithm, and show that when the underlying system is linear, there is no performance gain with a widely linear model—a simple point, but not always acknowledged in the literature.
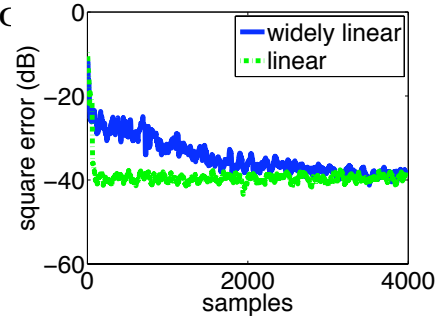
D. Examples

We define a random process

$$y(k) = \sqrt{1 - \zeta^2}y_r(k) + j\zeta y_i(k), \tag{58}$$

(a) Circular input ($\zeta = 1/\sqrt{2}$)

(b) Noncircular input ($\zeta = 0.1$)

Fig. 9.   Convergence of the linear and widely linear LMS algorithm for a linear system, for (a) circular input and (b) noncircular input

where $y_r(k)$ and $y_i(k)$ are two uncorrelated real-valued random processes, both Gaussian distributed with zero mean and unit variance. By changing the value of $\zeta \in [0, 1]$, we can change the degree of noncircularity of $y(k)$. For $\zeta = 1/\sqrt{2}$, the random process $y(k)$ becomes circular (and hence proper), whereas $\zeta = 0$ and $\zeta = 1$ lead to maximally noncircular $y(k)$. We then construct $x(k) = h(k) * y(k) + n(k)$. That is, we pass $y(k)$ through a linear system with impulse response $h(k) = \alpha(1 + \cos(2\pi(k-3)/5) - j[1 + \cos(2\pi(k-3)/10)])$, $k = 1, \ldots, 5$, where $\alpha = 0.432$ to ensure unit weight norm, and we also add uncorrelated white circular Gaussian noise $n(k)$ with signal-to-noise ratio of 20 dB.

Because $x(k)$ was obtained from $y(k)$ as the output of a linear system plus uncorrelated noise, the optimum filter for estimating $x(k)$ from $y(k)$ is obviously linear. However, we use this simple example to demonstrate what would happen if one were to use a widely linear filter instead. In order to estimate $x(k)$ using the LMS algorithm, we assemble M snapshots of $y(k)$ in the vector $\mathbf{y}(k) = [y(k), y(k-1), \cdots, y(k-M+1)]^T$. Its covariance matrix is $\mathbf{C} = \mathbf{I}$, and its complementary covariance matrix is $\widetilde{\mathbf{C}} = (1 - 2\zeta^2)\mathbf{I}$. The eigenvalues of the augmented covariance matrix $\underline{\mathbf{C}}$ can be shown to be $2\zeta^2$ and $2(1 - \zeta^2)$, each with multiplicity M. Hence, the condition number is $\kappa(\underline{\mathbf{C}}) = (1 - \zeta^2)/\zeta^2$ for $\zeta \in (0, 1/\sqrt{2}]$ and $\kappa(\underline{\mathbf{C}}) = \zeta^2/(1 - \zeta^2)$ for $\zeta \in [1/\sqrt{2}, 1)$.

In Figure 9, we show the convergence behavior of a linear and a widely linear LMS filter for estimating $x(k)$ from $\mathbf{y}(k)$. The step size is fixed at $\mu = 0.04$ for all runs, and we choose the length of the filter $M = 5$ to match the length of the system's impulse response. Figure 9 (a) shows the learning

(a) Circular input ($\zeta = 1/\sqrt{2}$)
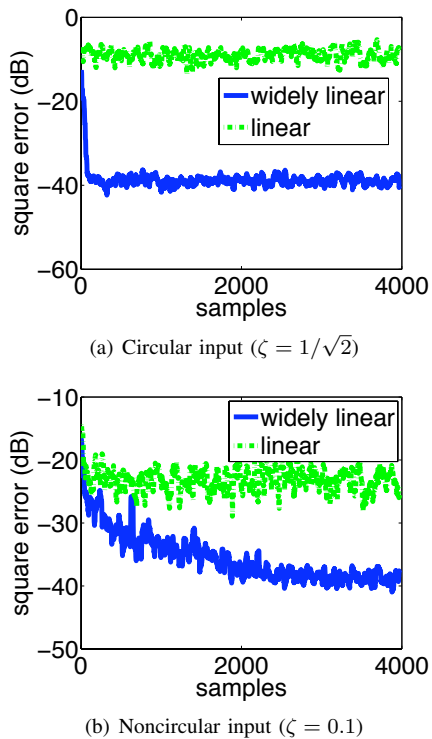

(b) Noncircular input ($\zeta = 0.1$)

Fig. 10. Convergence of the linear and widely linear LMS algorithm for a widely linear system, for (a) circular input and (b) noncircular input

curve for a circular input, *i.e.*, $\zeta = 1/\sqrt{2}$, and Figure 9 (b) for a noncircular input with $\zeta = 0.1$. The figures confirm that the linear and widely linear filters yield the same steady-state mean-square error values for circular and noncircular $y(k)$. However, in the noncircular case, the use of the widely linear LMS filter is even detrimental because of its decreased convergence rate. This is due to the fact that, for circular $y(k)$, the condition numbers for both $\mathbf{C}$ and $\underline{\mathbf{C}}$ are unity, whereas for noncircular $y(k)$, $\kappa(\mathbf{C}) = 1$ but $\kappa(\underline{\mathbf{C}}) = 99$.

Let us now consider the widely linear system $x(k) = \mathrm{Re}\{h(k) * y(k)\} + n(k)$ with the same $h(k)$ as before. Figure 10 shows the corresponding learning curves for the linear and widely linear LMS filters with the same settings for the simulation as before. As can be observed in the figures, the widely linear filter provides smaller MSE for both circular and noncircular $y(k)$. For circular input $y(k)$, the reduction in MSE by using the widely linear filter can be obtained from (56) as $J_{\mathrm{diff}} = \|\tilde{\mathbf{c}}_{yx}\|^2 = \|E\{x(k)\mathbf{y}(k)\}\|^2$. However, this performance gain again comes at the price of slower convergence due to the increased eigenvalue spread.

V. Independent Component Analysis

Data-driven signal processing methods are based on simple generative models and hence can minimize the assumptions on the nature of the data. They have emerged as promising alternatives to the traditional model-based approaches in many signal processing applications where the underlying dynamics are hard to characterize. The most common generative model used in data-driven decompositions is a linear mixing model, where the mixture (the given set of observations) is written

as a linear combination of source components, *i.e.*, it is decomposed into a mixing matrix and a source matrix. For uniqueness of the decomposition (subject to a scaling and permutation ambiguity), constraints such as sparsity, nonnegativity, or independence are applied to the two matrices. ICA is a popular data-driven blind source separation technique that imposes the constraint of statistical independence on the components, *i.e.*, the source distributions. It has been successfully applied to numerous signal processing problems in areas as diverse as biomedicine, communications, finance, and remote sensing [35], [34], [56], [5].

We consider the typical ICA problem, where $\mathbf{x}$ is a linear mixture of *independent components* (sources) $\mathbf{s}$, as described by

$$\mathbf{x} = \mathbf{Ms} \qquad (59)$$

where $\mathbf{x}, \mathbf{s} \in \mathbb{C}^N$, *i.e.*, the number of sources and observations are equal, and the *mixing matrix* $\mathbf{M} \in \mathbb{C}^{N \times N}$ is nonsingular. The objective is to blindly recover the sources $\mathbf{s}$ from the observations $\mathbf{x}$, without knowledge of $\mathbf{M}$, using the *demixing (or separating) matrix* $\mathbf{W} \in \mathbb{C}^{N \times N}$ such that the source estimates are $\mathbf{u} = \mathbf{Wx}$ where $\mathbf{u} = [u_1, \ldots, u_N]^T$. Arbitrary scaling of $\mathbf{s}$, *i.e.*, multiplication by a diagonal matrix (which may have complex entries), and reordering the components of $\mathbf{s}$, *i.e.*, multiplication by a permutation matrix, preserves the independence of its components. The product of a diagonal and a permutation matrix is a *monomial matrix*, which has exactly one nonzero entry in each column and row. Hence, we can determine $\mathbf{W}$ only up to multiplication with a monomial matrix.

A limitation of ICA for the real-valued case is that when the sources are white—*i.e.*, we cannot exploit the sample correlation structure in the data—we can only allow one Gaussian source in the mixture for successful separation. In the complex case, on the other hand, we can perform ICA of multiple Gaussian sources as long as they all have distinct circularity coefficients. In addition, the complex domain enables the separation of improper sources through a direct application of the invariance property of the circularity coefficients. When all the sources in the mixture are improper with distinct circularity coefficients, we can achieve ICA through joint diagonalization of the covariance and complementary covariance matrices using the strong uncorrelating transform (SUT) [41], [63]. For the real-valued case, separation using second-order statistics can be achieved only when the sources have sample correlation.

Using higher-order statistical information, we can perform ICA for any type of distribution, circular or noncircular, as long as there are no two complex Gaussian sources with the same circularity coefficient [41]. To achieve ICA, we can either compute the higher-order statistics explicitly, or we can generate them implicitly through the use of nonlinear functions. Among the former group is JADE (which stands for joint approximate diagonalization of eigenmatrices) [24], which computes cumulants. JADE can be used directly for ICA of complex-valued data. A recent extension of this algorithm [122] enables joint diagonalization of matrices that can be Hermitian or complex symmetric. Hence, it allows for

more efficient ICA solutions, considering also the commonly neglected complementary statistics in the original formulation of JADE.

Algorithms that rely on joint diagonalization of cumulant matrices are robust. However, their performance suffers as the number of sources increases, and the cost of computing and diagonalizing cumulant matrices may become prohibitive for separating a large number of sources. ICA techniques that use nonlinear functions to implicitly generate higher-order statistics may present attractive alternatives. Among these are ML-ICA [99], information-maximization (Infomax) [16], and maximization of non-Gaussianity (*e.g.*, the FastICA algorithm) [55], which are all intimately related to each other. These algorithms can be easily extended to the complex domain using Wirtinger calculus as shown in [7]. In addition, one can develop complex ICA algorithms that adapt to different source distributions, using general models such as complex generalized Gaussian distributions [90] as in [87], or more flexible models through efficient entropy estimation techniques [69].

A. ICA using Second-Order Statistics

We first present the second-order approach to ICA [40], [41], [63], which is based on the fact that the circularity coefficients of $\mathbf{s}$ are invariant under the linear mixing transformation $\mathbf{M}$ [113]. We discussed the circularity coefficients in Section III-A. There is a corresponding coordinate system, called the *canonical coordinate system*, where the latent description $\mathbf{s}'$ has identity correlation matrix, $E\{\mathbf{s}'\mathbf{s}'^H\} = \mathbf{I}$, and diagonal complementary correlation matrix with the circularity coefficients on the diagonal,

$$E\{\mathbf{s}'\mathbf{s}'^T\} = \mathbf{K} = \mathrm{diag}(k_1, k_2, ..., k_N).$$

In [41], vectors that are uncorrelated with unit variance, but possibly improper, are called *strongly uncorrelated*, and the transformation $\mathbf{A}_{ss}$, which transforms $\mathbf{s}$ into canonical coordinates $\mathbf{s}'$ as $\mathbf{s}' = \mathbf{A}_{ss}\mathbf{s}$, is called the *strong uncorrelating transform* (SUT). The SUT is found as

$$\mathbf{A}_{ss} = \mathbf{F}_{ss}^H \mathbf{C}_{ss}^{-1/2},$$

where $\mathbf{F}_{ss}$ is obtained from the Takagi factorization (28) for the coherence matrix of $\mathbf{s}$. If all circularity coefficients are distinct and nonzero, the SUT is unique up to the sign of its rows [41]. We will now show that the SUT $\mathbf{A}_{xx}$, computed from the *mixture* $\mathbf{x}$, is a separating matrix for the complex linear ICA problem provided that all circularity coefficients are distinct.

The assumption of *independent* components in $\mathbf{s}$ in the given ICA model (59) implies that the correlation matrix $\mathbf{C}_{ss}$ and the complementary correlation matrix $\widetilde{\mathbf{C}}_{ss}$ are both diagonal. It is therefore easy to compute canonical coordinates between $\mathbf{s}$ and $\mathbf{s}^*$, denoted by $\mathbf{s}' = \mathbf{A}_{ss}\mathbf{s}$. In the SUT $\mathbf{A}_{ss} = \mathbf{F}_{ss}^H \mathbf{C}_{ss}^{-1/2}$, $\mathbf{C}_{ss}^{-1/2}$ is a *diagonal* scaling matrix, and $\mathbf{F}_{ss}^H$ is a *permutation* matrix that rearranges the canonical coordinates $\mathbf{s}'$ such that s_1' corresponds to the largest circularity coefficient k_1, s_2' to the second largest coefficient k_2, and so on. This makes the
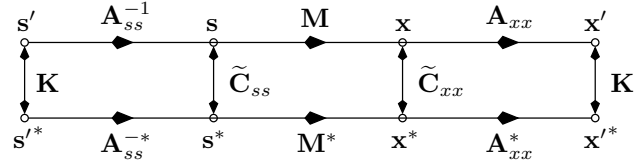

Fig. 11. Two-channel model for second-order complex ICA. The vertical arrows show the complementary correlation matrix between the upper and lower lines.

SUT $\mathbf{A}_{ss}$ monomial. As a consequence, $\mathbf{s}'$ has *independent components*.

The mixture $\mathbf{x}$ has correlation matrix $\mathbf{C}_{xx} = \mathbf{M}\mathbf{C}_{ss}\mathbf{M}^H$ and complementary correlation matrix $\widetilde{\mathbf{C}}_{xx} = \mathbf{M}\widetilde{\mathbf{C}}_{ss}\mathbf{M}^T$. The canonical coordinates between $\mathbf{x}$ and $\mathbf{x}^*$ are computed as $\mathbf{x}' = \mathbf{A}_{xx}\mathbf{x} = \mathbf{F}_{xx}^H\mathbf{C}_{xx}^{-1/2}\mathbf{x}$, and the SUT $\mathbf{A}_{xx}$ is determined analogously to $\mathbf{A}_{ss}$.

Figure 11 shows the connection between the different co-ordinate systems. The important observation is that $\mathbf{s}'$ and $\mathbf{x}'$ are both in canonical coordinates with the *same* circularity coefficients $\{k_n\}$. It remains to show that $\mathbf{s}'$ and $\mathbf{x}'$ are related by a diagonal unitary matrix $\mathbf{D} = \mathbf{A}_{xx}\mathbf{M}\mathbf{A}_{ss}^{-1}$ as $\mathbf{x}' = \mathbf{D}\mathbf{s}'$, provided that all circularity coefficients are distinct. Since $\mathbf{s}'$ and $\mathbf{x}'$ are both in canonical coordinates with the same diagonal canonical correlation matrix $\mathbf{K}$,

$$E\left\{\underline{\mathbf{s}}'\underline{\mathbf{s}}'^H\right\} = \begin{bmatrix} \mathbf{I} & \mathbf{K} \\ \mathbf{K} & \mathbf{I} \end{bmatrix}$$

$$E\left\{\underline{\mathbf{x}}'\underline{\mathbf{x}}'^H\right\} = \begin{bmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}^* \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{K} \\ \mathbf{K} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{D}^H & \mathbf{0} \\ \mathbf{0} & \mathbf{D}^T \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{D}\mathbf{D}^H & \mathbf{D}\mathbf{K}\mathbf{D}^T \\ \mathbf{D}^*\mathbf{K}\mathbf{D}^H & \mathbf{D}^*\mathbf{D}^T \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{K} \\ \mathbf{K} & \mathbf{I} \end{bmatrix}.$$

This shows that $\mathbf{D}$ is unitary and $\mathbf{D}\mathbf{K}\mathbf{D}^T = \mathbf{K}$. The latter can only be true if $D_{ij} = 0$ whenever $k_i \neq k_j$. Therefore, $\mathbf{D}$ is diagonal and unitary if all circularity coefficients are distinct. Since $\mathbf{K}$ is real, the corresponding diagonal entries of all nonzero circularity coefficients are actually ± 1. Thus, $\mathbf{x}'$ has independent components because $\mathbf{s}'$ has independent components. Hence, we have shown that the SUT $\mathbf{A}_{xx}$ is a separating matrix for the complex linear ICA problem if all circularity coefficients are distinct, and the ICA solution is $\mathbf{u} = \mathbf{W}\mathbf{x}$ with $\mathbf{u} = \mathbf{x}'$ and $\mathbf{W} = \mathbf{A}_{xx}$.

This development requires that correlation and complementary correlation matrices exist. For some distributions (e.g., Cauchy distribution) this is not the case. Ollila and Koivunen [96] have presented a generalization of the SUT that works for such distributions.

A recent source separation algorithm, entropy rate minimization (ERM) [71], uses second-order statistics along with sample correlation to improve the overall separation performance, and the second-order source separation algorithm SUT described above becomes a special case of the ERM algorithm.

B. ICA using Higher-Order Statistics

As discussed in Section V-A, ICA can be achieved using *only* second-order statistics as long as all the sources are

noncircular with distinct circularity coefficients. Using higher-order statistics, however, one can develop more powerful ICA algorithms that can achieve separation under more general conditions and separate sources as long as there are no two complex Gaussians with the same circularity coefficient in the given mixture.

A natural cost for achieving independence is mutual information, which can be expressed as the Kullback-Leibler distance between the joint and factored marginal source densities

$$\mathcal{I}(\mathbf{W}) = D\left(p_s(\mathbf{u}) \Big\| \prod_{n=1}^{N} p_{s_n}(u_n)\right) = \sum_{n=1}^{N} H(u_n) - H(\mathbf{u})$$

$$= \sum_{n=1}^{N} H(u_n) - H(\mathbf{x}) - \log|\det \mathbf{W}_{\mathbb{R}}| \qquad (60)$$

where

$$\mathbf{W}_{\mathbb{R}} = \begin{bmatrix} \mathbf{W}_r & -\mathbf{W}_i \\ \mathbf{W}_i & \mathbf{W}_r \end{bmatrix}$$

is the real representation of complex $\mathbf{W} = \mathbf{W}_r + j\mathbf{W}_i$ and $p_{s_n}(u_n) = p_{s_{n_r}, s_{n_i}}(u_{n_r}, u_{n_i})$ is the pdf of source n, $n = 1, \ldots, N$. To write (60), we used the Jacobian transformation

$$p_x(\mathbf{x}) = |\det \mathbf{W}_{\mathbb{R}}| p_u(\mathbf{W}\mathbf{x}) \qquad (61)$$

where $p_u(\mathbf{W}\mathbf{x}) = p_u(\mathbf{u})$.

When we have L independent samples $\mathbf{x}(l) \in \mathbb{C}^N$, we can use the mean ergodic theorem to write the first term in (60), the total source entropy, as

$$\sum_{n=1}^{N} H(u_n) = -\sum_{n=1}^{N} E\left\{\log p_{s_n}(\mathbf{w}_n^H \mathbf{x})\right\} \approx$$

$$-\frac{1}{L} \sum_{l=1}^{L} \sum_{n=1}^{N} \log p_{s_n}(\mathbf{w}_n^H \mathbf{x}(l)) \qquad (62)$$

where $\mathbf{w}_n$ denotes the nth row of $\mathbf{W}$. Since $H(\mathbf{x})$ is constant with respect to the weight matrix, it is easy to see that the minimization of mutual information given in (60) is equivalent to maximum likelihood estimation for the given L samples. The maximum likelihood $\mathcal{L}(\mathbf{W})$ is written as $\mathcal{L}(\mathbf{W}) = \sum_{l=1}^{L} \ell_l(\mathbf{W})$ where

$$\ell_l(\mathbf{W}) = \log p(\mathbf{x}(l)|\mathbf{W}) = \log \prod_{n=1}^{N} p_{s_n}(\mathbf{w}_n^H \mathbf{x}) + \log|\det \mathbf{W}_{\mathbb{R}}|.$$

$$(63)$$

In writing (63), we used $\mathbf{W} = \mathbf{M}^{-1}$ by ignoring the scaling and permutation ambiguity to simplify the notation. We have also omitted the sample index of $\mathbf{x}(l)$ for economy.

A simple but effective learning rule for maximizing the likelihood function can be obtained by directly using Wirtinger derivatives as explained in Section II-A. We write the real-valued cost function $\ell_l(\mathbf{W})$ as $\ell_l(\mathbf{W}, \mathbf{W}^*)$, and derive the relative (natural) gradient updates for the demixing matrix [7]

$$\Delta \mathbf{W} = (\mathbf{I} - \boldsymbol{\psi}(\mathbf{u})\mathbf{u}^H)\mathbf{W}, \qquad (64)$$

where the nth element of the score function $\boldsymbol{\psi}(\mathbf{u}) \in \mathbb{C}^N$ is the Wirtinger derivative

$$\psi_n(u_n) = -\frac{\partial \log p_{s_n}(u_n)}{\partial u_n^*}. \qquad (65)$$

This way the development avoids the need for simplifying assumptions such as circularity [14].

Another natural cost function for performing ICA is *negentropy*, which measures the entropic distance of a given distribution from the Gaussian. Independence is achieved by moving the distribution of the transformed mixture $\mathbf{w}_n^H \mathbf{x}$—the independent source estimate u_n—away from the Gaussian, *i.e.*, by maximizing non-Gaussianity. Note that in this case, the objective function is defined for each individual source rather than for the whole vector $\mathbf{u}$ as in ML ICA. With negentropy maximization as the objective, the cost function for estimating all sources is written as in (62), under the constraint that $\mathbf{W}$ be unitary. Since $\det \mathbf{W}_{\mathbb{R}} = |\det(\mathbf{W})|^2$, the second term in (63) vanishes for unitary $\mathbf{W}$. So maximizing likelihood and maximizing negentropy are equivalent goals for unitary $\mathbf{W}$.

If we approximate the negentropy—or, equivalently, the entropy (62) under a variance constraint—using a nonlinear function $G(\cdot)$

$$E\left\{\log p_{s_n}(\mathbf{w}_n^H \mathbf{x})\right\} = E\left\{|G(\mathbf{w}_n^H \mathbf{x})|^2\right\}, \qquad (66)$$

we can estimate the columns $\mathbf{w}_n$ of the unitary demixing matrix $\mathbf{W}$ in a deflationary procedure, where the sources are estimated one at a time. We can again use Wirtinger calculus to obtain either gradient updates [7] or modified Newton type updates [86], [7] such that the weight vector is updated by

$$\mathbf{w} \leftarrow E\{G'(u)[G'(u)]^*\}\mathbf{w} - E\{G^*(u)G'(u)\mathbf{x}\} + E\{\mathbf{x}\mathbf{x}^T\}E\{G^*(u)G''(u)\}\mathbf{w}^* \qquad (67)$$

where $G'(u)$ and $G''(u)$ denote the first and second derivatives of $G(\cdot)$. After each update (67), an orthogonalization and normalization step is used for $\mathbf{w}$ to ensure that the resulting $\mathbf{W}$ is unitary.

To summarize, ICA using higher-order statistics can be achieved by maximizing likelihood (which is equivalent to minimizing mutual information) or by maximizing negentropy. When the demixing matrix is constrained to be unitary—as it is for the negentropy cost—the two costs are equivalent. In both cases, the performance is optimized when the nonlinear functions, the score function in (64) and the nonlinearity $G(\cdot)$ in (67), are chosen to match the source distributions. This is discussed in more detail in the next section.

C. ICA Algorithms and Noncircularity

As shown in Section V-A, noncircularity (or rather impropriety) allows the separation of sources using only second-order statistics as long as all the sources in the mixture are noncircular with distinct circularity coefficients. Noncircularity also plays an important role in the performance of ICA algorithms. To be identifiable, the mixing matrix $\mathbf{M}$ in (59) must have full column rank and there must not be two Gaussian sources with the same circularity coefficient [41]. In terms of local stability, noncircularity plays a more critical role because algorithms become less stable as the degree of noncircularity increases. As noted in the previous section, the ML and negentropy cost functions are equivalent for unitary $\mathbf{W}$. Nevertheless, the stability conditions differ for these two cost functions
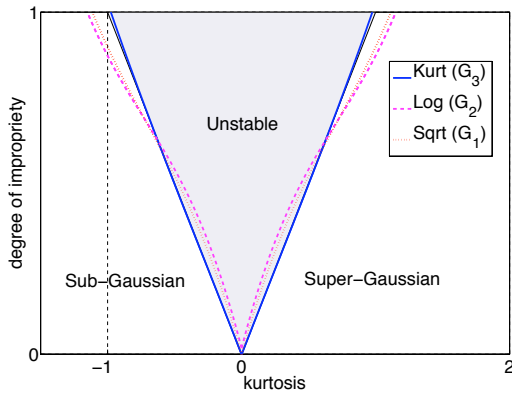
Fig. 12.   Regions of stability for negentropy cost given by $E\left\{G(|\mathbf{w}_n^H\mathbf{x}|^2)\right\}$

because, for the negentropy cost, the unitary constraint allows a desirable decoupling for each source estimate.

When the negentropy cost is approximated as $E\left\{G(|\mathbf{w}_n^H\mathbf{x}|^2)\right\}$, we can obtain local stability conditions. These are conditions for the cost function to have a local minimum or maximum [88], following the approach in [18], which assumed circularity. In Fig. 12, we plot the regions of stability for three nonlinear functions, $G_1(u) = \sqrt{a+u}$, $G_2(u) = \log(a+u)$, and $G_3(u) = u^2/2$, where a is an arbitrary constant, chosen as 0.1 as in [18]. The region of stability depends on the kurtosis (which is normalized such that the Gaussian has zero kurtosis) and the degree of impropriety $|E\{u^2\}|/E\{|u|^2\}$. We see that stability for sub-Gaussian sources, whose kurtosis is in the range $[-1,0)$, is always affected by impropriety. Examples of such sources include quadrature amplitude modulation, BPSK, and uniformly distributed signals. However, for super-Gaussian sources, whose kurtosis is greater than zero, impropriety affects stability only for sources whose kurtosis is close to the Gaussian. Hence, for super-Gaussian sources, matching the density to the improper nature of the signals is less of a concern than for sub-Gaussians.

For ML ICA, on the other hand, it is shown in [66] that, when all sources are circular and non-Gaussian, the updates converge to the inverse of the mixing matrix up to a phase shift. However, when the sources are noncircular and non-Gaussian, the stability conditions become more difficult to satisfy, especially when the demixing matrix is constrained to be unitary. When the sources are Gaussian though, ML ICA is stable as long as the circularity coefficients of all sources are distinct. This is the case that can also be solved using the second-order ICA approach described in Section V-A.

For ML ICA of *circular* sources, the nth entry of the score function vector $\boldsymbol{\psi}(\mathbf{u})$ in (65) can be shown to be [7]

$$\psi_n(u_n) = -\frac{\partial \log \phi(\sqrt{u_n u_n{}^*})}{\partial u_n{}^*} = -\frac{u_n}{2|u_n|}\left(\frac{\phi'(|u_n|)}{\phi(|u_n|)}\right)$$

because for this case we have $p_{s_n}(u_n) = \phi(|u_n|)$. Thus, the score function always has the same phase as its argument. This is the form of the score function proposed in [14] where all sources are assumed to be circular. When the source is cir-

cular Gaussian, the score function becomes linear: $\psi_n(u_n) = u_n/2\sigma_n^2$. This is expected because circular Gaussian sources cannot be separated. However, when they are noncircular, the score function is $\psi_n(u_n) = (\sigma_n^2 u_n - \tilde{\sigma}_n^2 u_n^*)/2(\sigma_n^4 - |\tilde{\sigma}_n^2|^2)$, where σ_n^2 and $\tilde{\sigma}_n^2$ are the variance and the complementary variance of the nth source. This function is widely linear in u_n, which allows separation as long as the sources have distinct circularity coefficients.

In addition to identifiability and local stability conditions, noncircularity also plays a role in terms of the overall performance of the algorithms. The performance of both ML-ICA and ICA by maximization of non-Gaussianity is optimal when the nonlinearity used in the algorithm (the score function $\boldsymbol{\psi}(\cdot)$ for ML and $G(\cdot)$ for maximization of non-Gaussianity) is matched to the density of each estimated source. This is also when the estimators assume their desirable large sample properties [54], [99].

Given the richer structure of possible distributions in the complex (two-dimensional) space compared to the real (one-dimensional) space, the pdf estimation problem becomes more challenging for complex-valued ICA. In the matching procedure, we need to consider not only the sub- or super-Gaussian but also the circular/noncircular nature of the sources. One way to do this is to employ the complex GGD (35) to derive a class of flexible ICA algorithms. Its shape parameter c and the augmented covariance matrix can be estimated during the ICA updates. This is the approach taken in the adaptive complex maximization of non-Gaussianity (ACMN) algorithm [86], [87] based on the updates given in (67). Another approach to density matching is entropy estimation through the use of a number of suitably chosen measuring functions [70]. Such an approach can provide robust performance even with a small set of chosen measuring functions [70]. Since the demixing matrix is not constrained in the maximum likelihood approach, exact density matching for each source becomes a more challenging task. This is due to the fact that in the update equation (64), the score function for each source (65) affects the whole demixing matrix estimate $\mathbf{W}$. In the complex ICA by entropy bound minimization (ICA-EBM) algorithm [69], the maximum likelihood updates are combined with a decoupling approach that enables easier density matching for each source independently. We study the performance of these algorithms in Section V-D.

Finally, it is important to note that exact density matching is not generally critical for the performance of ICA algorithms. As the discussion in Section III highlights, it is important only when the sample size is small or the sources are highly noncircular. In many cases, simpler algorithms that do not explicitly estimate the distributions, and may not even match the circular/noncircular nature of the data, can provide satisfactory performance for practical problems such as fMRI analysis [68].

D. Examples

In this section, we present two examples to show the performance of the ICA algorithms we have discussed so far, for Gaussian and non-Gaussian sources with varying degrees

of noncircularity [69]. Three indices are used to evaluate performance. Ideally, the product $\mathbf{P} = \mathbf{WM}$ of demixing and mixing matrix should be monomial. Based on that, we define the first performance index as follows: For each row of $\mathbf{P}$, we keep only the entry whose magnitude is largest. If the matrix we thus obtain is monomial, we declare success, otherwise failure. The ratio of failed trials is the first performance index. The second performance index is the average interference-to-signal-ratio (ISR), which is calculated as the average for all successful runs. The ISR for a given source is defined as the inverse ratio of the largest squared magnitude in the corresponding row of $\mathbf{P}$ to the sum of squared magnitudes for the remaining entries in that row. The third performance index is the average CPU time.

We consider three different ICA algorithms: (i) the second-order SUT [41]; (ii) ACMN, which maximizes negentropy by adaptively estimating the parameters of the GGD model in (35) [86], [87]; and (iii) complex ICA-EBM, which maximizes the likelihood and approximates the entropy (source distributions) through a set of flexible measuring functions [69].

The first example is the separation of noncircular Gaussian sources with *distinct* circularity coefficients. Five Gaussian signals with unit variance and complementary variances of $|E\{s^2\}| = 0, 0.2, 0.4, 0.6,$ and 0.8 are mixed with a randomly chosen square matrix whose real and imaginary elements are independently drawn from a zero-mean, unit-variance Gaussian distribution. Figure 13(a) shows the performance indices. We observe that the second-order SUT algorithm exhibits the best separation performance and also consumes the least CPU time. ACMN fails to separate the mixtures primarily because of the simplification it employs in the derivation. Complex ICA-EBM performs as well as SUT only for large sample sizes. The entropy estimator used in this algorithm can exactly match the entropy of a Gaussian source, which accounts for the asymptotic efficiency of complex ICA-EBM for the separation of Gaussian sources.

In the second example, we study the performance of the algorithms for complex GGD sources whose pdf is given in (35). Nine sources are generated with shape parameters $c = 0.2, 0.4, 0.6, 0.8, 1, 3, 5, 7,$ and 9, hence including four super-Gaussian, one Gaussian, and four sub-Gaussian sources. For each run, the complex correlation coefficient ρ is selected randomly inside the unit circle, resulting in different circularity coefficients. Figure 13(b) depicts the performance indices. We observe that the flexible ICA-EBM outperforms the other algorithms at a reasonable computational cost. ACMN, which is based on a complex GGD model, provides the next best performance but still with a significant performance gap primarily because of the unitary constraint it imposes on the demixing matrix $\mathbf{W}$.

In summary, the second-order SUT is efficient for estimation of Gaussian sources provided that they all have distinct circularity coefficients. By adaptively estimating the parameters of a GGD model, ACMN can provide satisfactory performance for most cases as shown in [69], [87] but it is also computationally demanding. Flexible density matching in ICA-EBM, on the other hand, provides an attractive trade-off between performance and computational cost. In addition, in

a maximum likelihood framework, the demixing matrix is not constrained. This can lead to better performance compared to algorithms that do impose constraints, such as ACMN. Except when prior information such as distinct circularity coefficients is available, a flexible algorithm such as ICA-EBM would thus provide the best choice with robust performance.

VI. CONCLUSIONS

In this overview paper, we have tried to illuminate the role that impropriety and noncircularity play in statistical signal processing of complex-valued data. We considered three questions: Why should we care about impropriety and non-circularity? When does it make sense to take it into account? How do we do it? In a nutshell, the answers to these questions are: We should care because it can lead to significant performance improvements in estimation, detection, and time-series analysis. We should take it into account whenever there is sufficient statistical evidence that signals and/or the underlying nature of the problem are indeed improper. We can do it by considering the complete statistical characterization of the signals and employing Wirtinger calculus for the optimization of cost functions with complex parameters.

In second-order methods such as mean-squared error estimation, a complete statistical characterization requires consideration of the complementary correlation; in problems such as ICA, it requires the use of flexible density models that can account for noncircularity. The caveat is that noncircular models have more degrees of freedom than circular models, and can hence lead to performance degradation in certain scenarios, even when the signals are noncircular. We noted that circular models are to be preferred when the SNR is low, the number of samples is small, or the degree of noncircularity is low. In addition, in the implementation of adaptive algorithms such as the LMS algorithm, taking the complete second-order statistical information into account may come at the expense of slower convergence.

We have reviewed some of the fundamental results, and some selected more recent developments in the field. There is a lot that we have not included, such as Cramér-Rao type performance bounds [127], [37], [97], [115] and a much more in-depth discussion of random processes. We have also paid only scant attention to the singular case of maximally improper/noncircular signals (also called "rectilinear" or "strict-sense noncircular" signals). This case deserves more attention because many algorithms developed for improper/noncircular signals either fail in the maximally improper case or may even give worse performance than algorithms developed for proper/circular signals. Examples of papers dealing explicitly with maximally improper signals are [49], [1], [2], [107]. In addition, we would like to note the growing interest in the extension of these results to hypercomplex numbers, in particular quaternions (see, *e.g.*, [124], [9], [17], [120], [130], [131], [121]).

We hope that this overview paper will encourage more researchers to take full advantage of the power of complex-valued signal processing.

REFERENCES

[1] H. Abeida and J. P. Delmas. MUSIC-like estimation of direction of arrival for non-circular sources. *IEEE Trans. Signal Processing*, 54:2678–2690, 2006.

[2] H. Abeida and J. P. Delmas. Statistical performance of MUSIC-like algorithms in resolving noncircular sources. *IEEE Trans. Signal Processing*, 56:4317–4329, 2008.

[3] M. J. Ablowitz and A. S. Fokas. *Complex Variables*. Cambridge University Press, Cambridge, UK, 2003.

[4] T. Adalı and V. D. Calhoun. Complex ICA of medical imaging data. *IEEE Signal Proc. Mag.*, 24(5):136–139, Sept. 2007.

[5] T. Adalı and H. Li. *Adaptive Signal Processing: Next Generation Solutions*, chapter Complex-valued Signal Processing. Wiley Interscience (T. Adalı and S. Haykin editors), 2010.

[6] T. Adalı, H. Li, and R. Aloysius. On properties of the widely linear MSE filter and its LMS implementation. In *Proc. Conf. Information Sciences and Systems*, Baltimore, MD, March 2009.

[7] T. Adalı, H. Li, M. Novey, and J.-F. Cardoso. Complex ICA using nonlinear functions. *IEEE Trans. Signal Processing*, 56(9):4356–4544, Sept. 2008.

[8] H. Akaike. A new look at statistical model identification. *IEEE Trans. Auto. Control*, 19:716–723, 1974 1974.

[9] P. O. Amblard and N. Le Bihan. On properness of quaternion valued random variables. In *Proc. Intl. Conf. Mathematics (IMA) in Signal Processing*, pages 23–26, 2004.

[10] P. O. Amblard and P. Duvaut. Filtrage adaptè dans le cas gaussien complexe non circulaire. In *Proc. GRETSI Conf.*, pages 141–144, 1995.

[11] P. O. Amblard, M. Gaeta, and J. L. Lacoume. Statistics for complex variables and signals – Part 1: Variables. *Signal Processing*, 53(1):1–13, 1996.

[12] P. O. Amblard, M. Gaeta, and J. L. Lacoume. Statistics for complex variables and signals – Part 2: Signals. *Signal Processing*, 53(1):15–25, 1996.

[13] S. A. Andersson and M. D. Perlman. Two testing problems relating the real and complex multivariate normal distribution. *J. Multivariate Analysis*, 15:21–51, 1984.

[14] J. Anemüller, T. J. Sejnowski, and S. Makeig. Complex independent component analysis of frequency-domain electroencephalographic data. *Neural Networks*, 16:1311–1323, 2003.

[15] L. Anttila, M. Valkama, and M. Renfors. Circularity-based I/Q imbalance compensation in wideband direct-conversion receivers. *IEEE Trans. Vehicular Techn.*, 57:2099–2113, 2008.

[16] A. Bell and T. Sejnowski. An information maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[17] N. Le Bihan and P. O. Amblard. Detection and estimation of Gaussian proper quaternion valued random processes. In *Proc. Intl. Conf. Mathematics (IMA) in Signal Processing*, pages 23–26, 2006.

[18] E. Bingham and A. Hyvärinen. A fast fixed-point algorithm for independent component analysis of complex valued signals. *Int. J. Neural Systems*, 10:1–8, 2000.

[19] D. H. Brandwood. A complex gradient operator and its application in adaptive array theory. *Proc. Inst. Elect. Eng.*, 130(1):11–16, Feb. 1983.

[20] W. M. Brown and R. B. Crane. Conjugate linear filtering. *IEEE Trans. Inform. Theory*, 15:462–465, 1969.

[21] H. J. Butterweck. A steady-state analysis of the LMS adaptive algorithm without the use of independence assumption. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, pages 1404–1407, Detroit, 1995.

[22] S. Buzzi, M. Lops, and S. Sardellitti. Widely linear reception strategies for layered space-time wireless communications. *IEEE Trans. Signal Processing*, 54:2252–2262, 2006.

[23] A. S. Cacciapuoti, G. Gelli, and F. Verde. FIR zero-forcing multiuser detection and code designs for downlink MC-CDMA. *IEEE Trans. Signal Processing*, 55:4737–4751, 2007.

[24] J.-F. Cardoso and A. Souloumiac. Blind beamforming for non-Gaussian signals. *IEE Proc. Radar Signal Processing*, 140:362–370, 1993.

[25] P. Charge, Y. Wang, and J. Saillard. A non-circular sources direction finding method using polynomial rooting. *Signal Processing*, 81:1765–1770, 2001.

[26] P. Chevalier and A. Blin. Widely linear MVDR beamformers for the reception of an unknown signal corrupted by noncircular interferences. *IEEE Trans. Signal Processing*, 55:5323–5336, 2007.

[27] P. Chevalier, J. P. Delmas, and A. Oukaci. Optimal widely linear MVDR beamforming for noncircular signals. In *Proc. ICASSP*, pages 3573–3576, 2009.

[28] P. Chevalier, P. Duvaut, and B. Picinbono. Complex transversal Volterra filters optimal for detection and estimation. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, pages 3537–3540, Toronto, ON, 1991.

[29] P. Chevalier and A. Maurice. Constrained beamforming for cyclostationary signals. In *Proc. ICASSP*, pages 3789–3792, 1997.

[30] P. Chevalier and B. Picinbono. Complex linear-quadratic systems for detection and array processing. *IEEE Trans. Signal Processing*, 44(10):2631–2634, Oct. 1996.

[31] P. Chevalier and F. Pipon. New insights into optimal widely linear array receivers for the demodulation of BPSK, MSK, and GMSK interferences — application to SAIC. *IEEE Trans. Signal Processing*, 54:870–883, 2006.

[32] P. Comon. Estimation multivariable complexe. *Traitement du signal*, 3:97–102, 1986.

[33] P. Comon. Circularité et signaux aléatoires à temps discret. *Traitement du Signal*, 11(5):417–420, 1994.

[34] P. Comon. Independent component analysis – a new concept? *Signal Processing*, 36:287–314, 1994.

[35] P. Comon and C. Jutten. *Handbook of Blind Source Separation*. Academic Press, 2010.

[36] J. P. Delmas. Asymptotically minimum variance second-order estimation for noncircular signals with application to DOA estimation. *IEEE Trans. Signal Processing*, 52:1235–1241, 2004.

[37] J. P. Delmas and H. Abeida. Stochastic Cramér-Rao bound for noncircular signals with application to DOA estimation. *IEEE Trans. Signal Processing*, 52:3192–3199, 2004.

[38] S. Douglas and D. Mandic. Performance analysis of the conventional complex LMS and augmented complex LMS algorithms. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, pages 3794–3797, Dallas, TX, March 2010.

[39] P. Duvaut. Processus et vecteurs complexes elliptiques. In *Proc. GRETSI Conf.*, pages 129–132, 1995.

[40] J. Eriksson and V. Koivunen. Complex-valued ICA using second order statistics. In *Proc. IEEE Int. Workshop on Machine Learning for Signal Processing (MLSP)*, pages 183–192, Saõ Luis, Brazil, Sept. 2004.

[41] J. Eriksson and V. Koivunen. Complex random vectors and ICA models: Identifiability, uniqueness and separability. *IEEE Trans. Info. Theory*, 52(3):1017–1029, 2006.

[42] J. Eriksson, E. Ollila, and V. Koivunen. Essential statistics and tools for complex random variables. *IEEE Trans. Signal Processing*, to appear.

[43] M. Gaeta. *Higher-Order Statistics Applied to Source Separation*. PhD thesis, INP, Grenoble, France, 1991.

[44] W. A. Gardner. Cyclic Wiener filtering: theory and method. *IEEE Trans. Commun.*, 41:151–163, 1993.

[45] G. Gelli, L. Paura, and A. R. P. Ragozini. Blind widely linear multiuser detection. *IEEE Commun. Lett.*, 4:187–189, 2000.

[46] H. Gerstacker, R. Schober, and A. Lampe. Receivers with widely linear processing for frequency-selective channels. *IEEE Trans. Commun.*, 51:1512–1523, 2003.

[47] N. R. Goodman. Statistical analysis based on a certain multivariate complex Gaussian distribution. *Annals Math. Stats.*, 34:152–176, 1963.

[48] T. L. Grettenberg. A representation theorem for complex normal processes. *IEEE Trans. Inform. Theory*, 11:305–306, 1965.

[49] M. Haardt and F. Roemer. Enhancements of unitary ESPRIT for non-circular sources. In *Proc. ICASSP*, pages 101–104, 2004.

[50] S. Haykin. *Adaptive Filter Theory*. Prentice-Hall, Inc., Upper Saddle River, NJ, fourth edition edition, 2002.

[51] P. Henrici. *Applied and Computational Complex Analysis*, volume III. Wiley, New York, NY, 1986.

[52] L. Hörmander. *An Introduction to Complex Analysis in Several Variables*. North-Holland, Oxford, U.K., 1990.

[53] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, New York, NY, 1985, 1999.

[54] A. Hyvärinen. One-unit contrast functions for independent component analysis: A statistical analysis. In *Proc. IEEE Workshop on Neural Networks for Signal Processing (NNSP)*, pages 388–397, Amelia Island, FL, Sept. 1997.

[55] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Networks*, 10(3):626–634, 1999.

[56] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, New York, NY, 2001.

[57] J. J. Jeon, J. G. Andrews, and K. M. Sung. The blind widely linear minimum output energy algorithm for DS-CDMA systems. *IEEE Trans. Signal Processing*, 54:1926–1931, 2006.

PUBLISHED: IEEE TRANS. SIGNAL PROCESSING, VOL. 59, NO. 11, PP. 4516–4528, NOV. 2011.

24

[58] S. M. Kay and J. R. Gabriel. An invariance property of the generalized likelihood ratio test. *IEEE Signal Processing Lett.*, 10:352–355, 2003.

[59] K. Kreutz-Delgado. The complex gradient operator and the CR-calculus. ECE275A: Parameter Estimation I, Lecture Supplement on Complex Vector Calculus, 2007.

[60] J. L. Lacoume. Variables et signaux alèatoires complexes. *Traitement du Signal*, 15:535–544, 1998.

[61] J. L. Lacoume and M. Gaeta. Complex random variable: A tensorial approach. In *Proc. IEEE Workshop on Higher-Order Statistics*, 1991.

[62] A. Lampe, R. Schober, and W. Gerstacker. A novel iterative multiuser detector for complex modulation schemes. *IEEE J. Sel. Areas Commun.*, 20:339–350, 2002.

[63] L. De Lathauwer and B. De Moor. On the blind separation of non-circular sources. In *Proc. European Signal Process. Conf. (EUSIPCO)*, Toulouse, France, 2002.

[64] H. Li and T. Adalı. Optimization in the complex domain for nonlinear adaptive filtering. In *Proc. 33rd Asilomar Conf. on Signals, Systems and Computers*, pages 263–267, Pacific Grove, CA, Nov. 2006.

[65] H. Li and T. Adalı. Complex-valued adaptive signal processing using nonlinear functions. *J. Advances in Signal Processing*, 2008(Article ID 765615):9 pages, 2008.

[66] H. Li and T. Adalı. Algorithms for complex ML ICA and their stability analysis using Wirtinger calculus. *IEEE Trans. Signal Processing*, 2010.

[67] H. Li and T. Adalı. Application of independent component analysis with adaptive density model to complex-valued fMRI data. *IEEE Trans. Biomedical Eng.*, in press.

[68] H. Li, T. Adalı, N. M. Correa, P. A. Rodriguez, and V. D. Calhoun. Flexible complex ICA of fMRI data. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Dallas, TX, March 2010.

[69] X.-L. Li and T. Adalı. Complex independent component analysis by entropy bound minimization. *IEEE Trans. Circuits Syst. I: Fund. Theory Apps.*, 57(7):1417–1430, July 2010.

[70] X.-L. Li and T. Adalı. Independent component analysis by entropy bound minimization. *IEEE Trans. Signal Processing*, 58(10):5151–5164, Oct. 2010.

[71] X.-L. Li and T. Adalı. Blind separation of noncircular correlated sources using Gaussian entropy rate. *IEEE Trans. Signal Processing*, 59(6):2969–2975, June 2011.

[72] X.-L. Li, T. Adalı, and M. Anderson. Noncircular principal component analysis and its application to model selection. *IEEE Trans. Signal Processing*, in press.

[73] X.-L. Li, M. Anderson, and T. Adalı. Detection of circular and noncircular signals in the presence of circular white Gaussian noise. In *Proc. Asilomar Conf. on Signals, Systems and Computers*, Pacific Grove, CA, Nov. 2010.

[74] O. Macchi. *LMS Adaptive Processing with Applications in Transmission*. Wiley, New York, NY, 1996.

[75] D. P. Mandic and V. S. L. Goh. *Complex Valued Nonlinear Adaptive Filters*. Wiley, New York, NY, 2009.

[76] P. Marchand, P. O. Amblard, and J. L. Lacoume. Statistiques d'ordre supèrieur á deux pour des signaux cyclostationnaires á valeurs complexes. In *Proc. GRETSI Conf.*, pages 69–72, 1995.

[77] T. McWhorter and P. Schreier. Widely-linear beamforming. In *Proc. 37th Asilomar Conf. Signals, Systems, Comput.*, pages 753–759, 2003.

[78] R. Meyer, W. H. Gerstacker, R. Schober, and J. B. Huber. A single antenna interference cancellation algorithm for increased GSM capacity. *IEEE Trans. Wireless Commun.*, 5:1616–1621, 2006.

[79] A. Mirbagheri, N. Plataniotis, and S. Pasupathy. An enhanced widely linear CDMA receiver with OQPSK modulation. *IEEE Trans. Commun.*, 54:261–272, 2006.

[80] C. N. K. Mooers. A technique for the cross spectrum analysis of pairs of complex-valued time series, with emphasis on properties of polarized components and rotational invariants. *Deep-Sea Research*, 20:1129–1141, 1973.

[81] D. R. Morgan. Variance and correlation of square-law detected allpass channels with bandpass harmonic signals in Gaussian noise. *IEEE Trans. Signal Processing*, 54:2964–2975, 2006.

[82] D. R. Morgan and C. K. Madsen. Wide-band system identification using multiple tones with allpass filters and square-law detectors. *IEEE Trans. Circuits Systems–I*, 53:1151–1165, 2006.

[83] A. Napolitano and M. Tanda. Doppler-channel blind identification for noncircular transmissions in multiple-access systems. *IEEE Trans. Commun.*, 52:2073–2078, 2004.

[84] F. D. Neeser and J. L. Massey. Proper complex random processes with applications to information theory. *IEEE Trans. Info. Theory*, 39:1293–1302, July 1993.

[85] R. Nilsson, F. Sjoberg, and J. P. LeBlanc. A rank-reduced LMMSE canceller for narrowband interference suppression in OFDM-based systems. *IEEE Trans. Commun.*, 51:2126–2140, 2003.

[86] M. Novey. *Complex ICA using Nonlinear Functions*. PhD thesis, University of Maryland Graduate School, Baltimore, MD, 2009.

[87] M. Novey and T. Adalı. Complex ICA by negentropy maximization. *IEEE Trans. Neural Networks*, 19(4):596–609, April 2008.

[88] M. Novey and T. Adalı. On extending the complex FastICA algorithm to noncircular sources. *IEEE Trans. Signal Processing*, 56(5):2148–2154, April 2008.

[89] M. Novey, T. Adalı, and A. Roy. Circularity and Gaussianity detection using the complex generalized Gaussian distribution. *IEEE Signal Proc. Letts.*, 16(11):993–996, Nov. 2009.

[90] M. Novey, T. Adalı, and A. Roy. A complex generalized Gaussian distribution—characterization, generation, and estimation. *IEEE Trans. Signal Processing*, 58(3):1427–1433, March 2010.

[91] S. Olhede. On probability density functions for complex variables. *IEEE Trans. Info. Theory*, 52:1212–1217, March 2006.

[92] E. Ollila. On the circularity of a complex random variable. *IEEE Signal Processing Lett.*, 15:841–844, 2008.

[93] E. Ollila, J. Eriksson, and V. Koivunen. Complex elliptically symmetric random variables – generation, characterization and circularity tests. *IEEE Trans. Signal Processing*, 2010.

[94] E. Ollila and V. Koivunen. Generalized complex elliptical distributions. In *Proc. 3rd Sensor Array Multichannel Signal Processing Workshop*, pages 460–464, Sitges, Spain, July 2004.

[95] E. Ollila and V. Koivunen. Adjusting the generalized likelihood ratio test of circularity robust to non-normality. In *Proc. IEEE International Workshop on Signal Proc.*, Perugia, Italy, June 2009.

[96] E. Ollila and V. Koivunen. Complex ICA using generalized uncorre-lating transform. *Signal Processing*, 89:365–377, April 2009.

[97] E. Ollila, V. Koivunen, and J. Eriksson. On the Cramer-Rao bound for the constrained and unconstrained complex parameters. In *Proc. IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM' 08)*, pages 414–418, Darmstadt, Germany, July 2008.

[98] K. B. Petersen and M. S. Pedersen. The Matrix Cookbook, Oct. 2008. Version 20081110.

[99] D. Pham and P. Garat. Blind separation of mixtures of independent sources through a quasi maximum likelihood approach. *IEEE Trans. Signal Processing*, 45(7):1712–1725, 1997.

[100] B. Picinbono. On circularity. *IEEE Trans. Signal Processing*, 42:3473–3482, Dec. 1994.

[101] B. Picinbono. Second-order complex random vectors and normal distributions. *IEEE Trans. Signal Processing*, 44(10):2637–2640, October 1996.

[102] B. Picinbono and P. Bondon. Second-order statistics of complex signals. *IEEE Trans. Signal Processing*, 45(2):411–419, Feb. 1997.

[103] B. Picinbono and P. Chevalier. Widely linear estimation with complex data. *IEEE Trans. Signal Processing*, 43:2030–2033, Aug. 1995.

[104] R. Remmert. *Theory of Complex Functions*. Springer-Verlag, Harrison-burg, VA, 1991.

[105] J. Rissanen. Modeling by the shortest data description. *Automatica*, 14:465–471, 1978.

[106] P. A. Rodriguez, V. D. Calhoun, and T. Adalı. De-noising, phase ambiguity correction and visualization techniques for complex-valued ICA of group fMRI data. *Pattern Recognition*, in press.

[107] F. Römer and M. Haardt. Multidimensional unitary tensor-esprit for non-circular sources. In *Proc. ICASSP*, 2009.

[108] P. Rubin-Delanchy and A. T. Walden. Simulation of improper complex-valued sequences. *IEEE Trans. Signal Processing*, 55:5517–5521, 2007.

[109] P. Ruiz and J. L. Lacoume. Extraction of independent sources from correlated inputs: A solution based on cumulants. In *Proc. Workshop on Higher-Order Spectral Analysis*, pages 146–151, June 1989.

[110] P. Rykaczewski, M. Valkama, and M. Renfors. On the connection of I/Q imbalance and channel equalization in direct-conversion transceivers. *IEEE Trans. Vehicular Techn.*, 57:1630–1636, 2008.

[111] L. L. Scharf, P. J. Schreier, and A. Hanssen. The Hilbert space geometry of the Rihaczek distribution for stochastic analytic signals. *IEEE Signal Processing Lett.*, 12:297–300, 2005.

[112] P. J. Schreier. Bounds on the degree of improperiety of complex random vectors. *IEEE Signal Process. Lett.*, 15:190–193, 2008.

[113] P. J. Schreier, T. Adalı, and L. L. Scharf. On ICA of improper and noncircular sources. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, pages 3561–3564, Taipei, Taiwan, April 2009.

25

[114] P. J. Schreier and L. L. Scharf. Second-order analysis of improper complex random vectors and processes. *IEEE Trans. Signal Processing*, 51(3):714–725, March 2003.

[115] P. J. Schreier and L. L. Scharf. *Statistical Signal Processing of Complex-Valued Data: The Theory of Improper and Noncircular Signals.* Cambridge, 2010.

[116] P. J. Schreier, L. L. Scharf, and A. Hanssen. A generalized likelihood ratio test for impropriety of complex signals. *IEEE Signal Process. Lett.*, 13(7):433–436, July 2006.

[117] P. J. Schreier, L. L. Scharf, and C. T. Mullis. Detection and estimation of improper complex random signals. *IEEE Trans. Inform. Theory*, 51(1):306–312, Jan. 2005.

[118] G. E. Schwarz. Estimating the dimensions of a model. *Ann. Stats.*, 6(2):461–464, 1978.

[119] G. Tauböck. Complex noise analysis of DMT. *IEEE Trans. Signal Processing*, 55:5739–5754, 2007.

[120] C. C. Took and D. P. Mandic. Quaternion-valued stochastic gradient-based adaptive IIR filtering. *IEEE Trans. Signal Processing*, 58:3895–3901, July 2010.

[121] C. C. Took and D. P. Mandic. Augmented second-order statistics of quaternion random signals. *Signal Processing*, 91(214–224), Feb. 2011.

[122] T. Trainini, X.-L. Li, E. Moreau, and T. Adalı. A relative gradient algorithm for joint decompositions of complex matrices. In *Proc. European Signal Process. Conf. (EUSIPCO)*, Aalborg, Denmark, August 2010.

[123] H. Trigui and D. T. M. Slock. Performance bounds for cochannel interference cancellation within the current GSM standard. *Signal Processing*, 80:1335–1346, 2000.

[124] N. N. Vakhania. Random vectors with values in quaternion Hilbert spaces. *Theory Probability Appl.*, 43(1):99–115, 1999.

[125] N. N. Vakhania and N. P. Kandelaki. Random vectors with values in complex Hilbert spaces. *Theory Probability Appl.*, 41(1):116–131, Feb. 1996.

[126] A. van den Bos. Complex gradient and Hessian. *IEE Proc.: Vision, Image, and Signal Processing*, 141(6):380–382, Dec. 1994.

[127] A. van den Bos. A Cramér-Rao lower bound for complex parameters. *IEEE Trans. Signal Processing*, 42:2859, 1994.

[128] A. van den Bos. Estimation of complex parameters. In *10th IFAC Symp.*, volume 3, pages 495–499, July 1994.

[129] A. van den Bos. The multivariate complex normal distribution—A generalization. *IEEE Trans. Inform. Theory*, 41:537–539, 1995.

[130] J. Via, D. Ramirez, I. Santamaria, and L. Vielva. Properness and widely linear processing of quaternion random vectors. *IEEE Trans. Info. Theory*, 56:3502–3515, July 2010.

[131] J. Via, D. Ramirez, I. Santamaria, and L. Vielva. Widely and semi-widely linear processing of quaternion vectors. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, pages 3946–3949, Dallas, TX, March 2010.

[132] P. Wahlberg and P. J. Schreier. Spectral relations for multidimensional complex improper stationary and (almost) cyclostationary processes. *IEEE Trans. Inform. Theory*, 54:1670–1682, 2008.

[133] A. T. Walden and P. Rubin-Delanchy. On testing for impropriety of complex-valued Gaussian vectors. *IEEE Trans. Signal Processing*, 57:825–834, March 2009.

[134] M. Wax and T. Kailath. Detection of signals by information theoretic criteria. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 33(2):387–392, April 1985.

[135] B. Widrow, J. Cool, and M. Ball. The complex LMS algorithm. *Proc. IEEE*, 63:719–720, 1975.

[136] B. Widrow and Jr. M. E. Hopf. Adaptive switching circuits. In *IRE WESCON*, volume 4, pages 96–104, 1960.

[137] W. Wirtinger. Zur formalen Theorie der Funktionen von mehr komplexen Veränderlichen. *Math. Ann.*, 97:357–375, 1927.

[138] M. Witzke. Linear and widely linear filtering applied to iterative detection of generalized MIMO signals. *Ann. Telecommun.*, 60:147–168, 2005.

[139] R. A. Wooding. The multivariate distribution of complex normal variables. *Biometrika*, 43:212–215, 1956.

[140] Y. C. Yoon and H. Leib. Maximizing SNR in improper complex noise and applications to CDMA. *IEEE Commun. Letters*, 1:5–8, 1997.

[141] V. Zarzoso and P. Comon. Robust independent component analysis by iterative maximization of the kurtosis contrast with algebraic optimal step size. *IEEE Trans. Neural Networks*, 21(2):248–261, Feb. 2010.

[142] Y. Zou, M. Valkama, and M. Renfors. Digital compensation of I/Q imbalance effects in space-time coded transmit diversity systems. *IEEE Trans. Signal Processing*, 56:2496–2508, 2008.

Tülay Adalı (S'89–M'93–SM'98–F'09) received the Ph.D. degree in electrical engineering from North Carolina State University, Raleigh, in 1992 and joined the faculty at the University of Maryland Baltimore County (UMBC), Baltimore, the same year. She is currently a Professor in the Department of Computer Science and Electrical Engineering at UMBC.

Prof. Adalı assisted in the organization of a number of international conferences and workshops including the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), the IEEE International Workshop on Neural Networks for Signal Processing (NNSP), and the IEEE International Workshop on Machine Learning for Signal Processing (MLSP). She was the General Co-Chair, NNSP (2001–2003); Technical Chair, MLSP (2004–2008); Program Co-Chair, MLSP (2008 and 2009), 2009 International Conference on Independent Component Analysis and Source Separation; Publicity Chair, ICASSP (2000 and 2005); and Publications Co-Chair, ICASSP 2008.

Prof. Adalı chaired the IEEE SPS Machine Learning for Signal Processing Technical Committee (2003–2005); Member, SPS Conference Board (1998–2006); Member, Bio Imaging and Signal Processing Technical Committee (2004–2007); and Associate Editor, IEEE Transactions on Signal Processing (2003–2006), Elsevier Signal Processing Journal (2007–2010). She is currently Chair of the MLSP Technical Committee and serving on the Signal Processing Theory and Methods Technical Committee; Associate Editor, IEEE Transactions on Biomedical Engineering and Journal of Signal Processing Systems for Signal, Image, and Video Technology; Senior Editorial Board member, IEEE Journal of Selected Areas in Signal Processing.

Prof. Adalı is a Fellow of the IEEE and the AIMBE and the recipient of a 2010 IEEE Signal Processing Society Best Paper Award and the past recipient of an NSF CAREER Award. Her research interests are in the areas of statistical signal processing, machine learning for signal processing, and biomedical data analysis.

Peter J. Schreier (S '03, M '04, SM '09) was born in Munich, Germany, in 1975. He received a Master of Science from the University of Notre Dame, Indiana, in 1999, and a Ph.D. from the University of Colorado at Boulder, in 2003, both in electrical engineering. From 2004 until January 2011, he was a faculty member in the School of Electrical Engineering and Computer Science at the University of Newcastle, Australia. Since February 2011, he has been Chaired Professor of Signal and System Theory in the Faculty of Electrical Engineering, Computer Science, and Mathematics at Universität Paderborn, Germany. This chair receives financial support from the Alfried Krupp von Bohlen und Halbach foundation.

Prof. Schreier has received fellowships from the State of Bavaria, the Studienstiftung des deutschen Volkes (German National Academic Foundation), and the Deutsche Forschungsgemeinschaft (German Research Foundation). He currently serves as Area Editor and Associate Editor of the *IEEE Transactions on Signal Processing*. He is a member of the IEEE Technical Committee on *Machine Learning for Signal Processing*.

For Louis Scharf, please use the photo and bio from the paper: Yuejie Chi, Louis Scharf, Ali Pezeshki, and Robt Calderbank, "Sensitivity of Basis Mismatch to Compressed Sensing," *IEEE Trans. Signal Proc.*, May 2011.
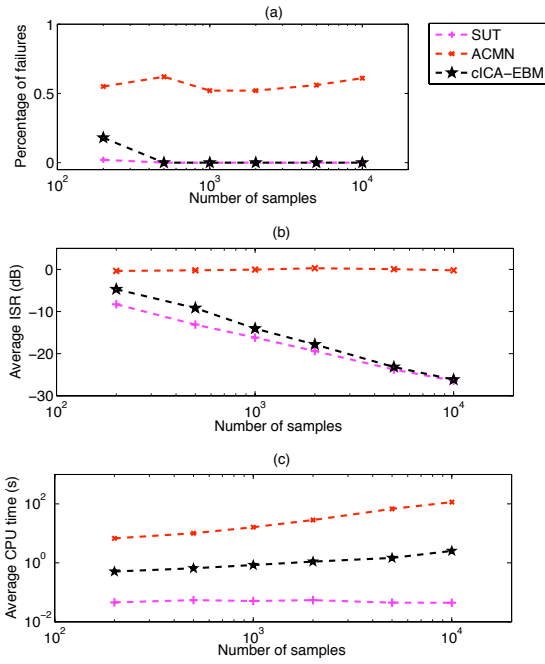
(a) Separation performance for noncircular Gaussian sources
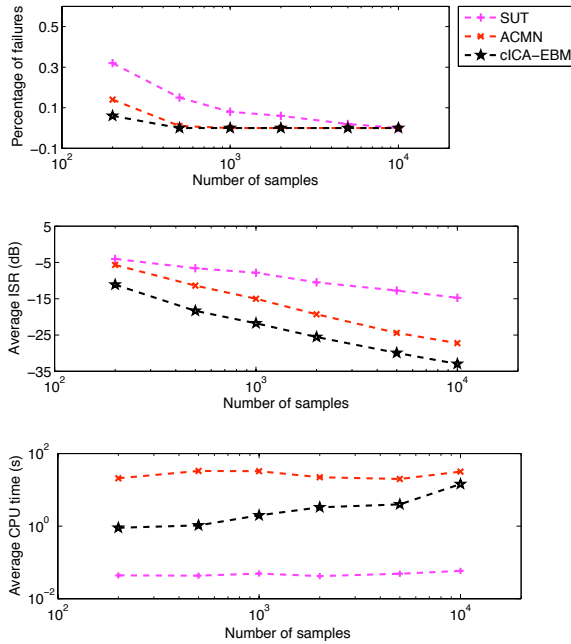
(b) Separation performance for GGD sources

Fig. 13.   Performance comparison of eight complex ICA algorithms for the separation of (a) mixtures of five Gaussian sources with degrees of noncircularity $|E\{s^2\}|/E\{|s|^2\} = 0$, 0.2, 0.4, 0.6, and 0.8; and (b) mixtures of nine noncircular complex sources drawn from GGD distributions, both as a function of sample size. Each simulation point is averaged over 100 independent runs.