Noname manuscript No.
(will be inserted by the editor)

IVA Using Complex Multivariate GGD: Application to fMRI Analysis

Rami Mowakeaa · Zois Boukouvalas · Qunfang Long · Tülay Adali

Received: date / Accepted: date*

Abstract Examples of complex-valued random phenomena in science and engineering are abound, and joint blind source separation (JBSS) provides an effective way to analyze multiset data. Thus there is a need for flexible JBSS algorithms for efficient data-driven feature extraction in the complex domain. Independent vector analysis (IVA) is a prominent recent extension of independent component analysis to multivariate sources, i.e., to perform JBSS, but its effectiveness is determined by how well the source models used match the true latent distributions and the optimization algorithm employed. The complex multivariate generalized Gaussian distribution (CMGGD) is a simple, yet effective parameterized family of distributions that account for full second- and higher-order statistics including noncircularity, a property that has been often omitted for convenience. In this paper, we marry IVA and CMGGD to derive, IVA-CMGGD, with a number of numerical optimization implementations including steepest descent, the quasi-Newton method Broyden-Fletcher-Goldfarb-Shanno (BFGS), and its limited-memory sibling limited-memory BFGS (L-BFGS) all in the complex-domain. We demonstrate the performance of our algorithm on simulated data as well as a 14-subject real-world complex-valued functional magnetic resonance imaging (fMRI) dataset against a number of competing algorithms.

Keywords IVA · noncircularity · complex-valued · BSS · fMRI

* Published in Multidimensional Signals and Systems on 10/9/2019:
Mowakeaa, R., Boukouvalas, Z., Long, Q. et al. IVA using complex multivariate GGD: application to fMRI analysis. Multidim Syst Sign Process 31, 725–744 (2020). https://doi.org/10.1007/s11045-019-00685-0

This work was supported in part by NSF-CCF 1618551 and NSF-NCS 1631838

R. Mowakeaa · Q. Long · T. Adali
University of Maryland, Baltimore County, Baltimore, MD 21250
Tel.: 410-455-1000
E-mail: {ramo1,qunfang1,adali}@umbc.edu

Z. Boukouvalas
American University, Washington, DC 20016
E-mail: boukouva@american.edu

1 Introduction

The complex-valued domain is the natural home for the processing of signals for various problems such as those in medical imaging, communications, sonar and radar array processing, and geophysics. Recently, research efforts increasingly focus on the analysis of multiple datasets jointly to exploit complimentary information among them. Independent vector analysis (IVA), a recent extension of independent component analysis (ICA) to multiple datasets, decomposes each dataset into a linear combination of statistically-independent factors taking cross-dataset dependencies into account. It has been successfully applied to a good number of applications where joint source separation is needed and one has to fully leverage the statistical information across multiple datasets, for example, in the joint analysis of electroencephalographic datasets (Bridwell et al, 2018), functional magnetic resonance imaging (fMRI) applications (Lv et al, 2018), and multichannel audio array processing (Itakura et al, 2018), to name a few. Therefore, flexible algorithms are needed in the complex domain that are data-driven since, in many cases, little prior information is known about the nature of these sources. The IVA algorithm, IVA Gaussian (IVA-G) (Anderson et al, 2012a), on the other hand, uses the noncircular Gaussian distribution as the latent source model which can exploit second-order statistics. IVA Laplacian (IVA-L) (Lee et al, 2006), uses an uncorrelated, circular Laplacian model which better matches the heavy tails of physiologically-relevant components in fMRI but fails to exploit noncircularity which has been shown in fMRI sources (Adalı et al, 2011b; Rodriguez et al, 2012). In (Kuang et al, 2017), the authors develop a noncircular complex-valued multivariate IVA algorithm based on the real-valued multivariate generalized Gaussian distribution (MGGD), which we call adaptive IVA, by explicitly incorporating the pseudo-covariance matrix in the update rule. However, a number of simplifying assumptions are made on the data that bias the results: first, the demixing matrices are constrained to be orthogonal which is generally not true (Cardoso, 1998). Second, for each source, the covariance matrix is estimated using its dominant eigenvalue only. This is informally justified based on experimental observations in multiband frequency-domain acoustic data in (Na et al, 2013) but justification for other sources is unclear.

Since the performance of IVA is intimately tied to the accuracy of the modeling assumptions made to the underlying latent sources, we target families of distributions that are simple (i.e., require few parameters to estimate), yet flexible enough to accommodate a variety of sources. In (Ollila et al, 2012), a broad survey of complex-valued, elliptically-symmetric distributions is presented, but the simple generalized Gaussian distribution derived omits noncircularity.

In (Mowakeaa et al, 2016), we develop the complex MGGD (CMGGD) family of distributions that generalizes the generalized Gaussian distribution in (Ollila et al, 2012) to noncircular random vectors. We develop an estimator for the CMGGD augmented covariance matrix which incorporates both the classical covariance matrix in addition to the pseudo-covariance matrix within the maximum likelihood framework. We illustrate the advantage of incorporating all available forms of statistical diversity—in this discussion, noncircularity, second- and higher-order statistics—into source distributions through the improved performance in estimating the augmented covariance matrix over a range of shape parameters.

In this paper, we employ the CMGGD family in the development of IVA without constraining the demixing matrices to be orthogonal—or unitary in the complex-valued case. Using the complex augmented form, we present the cost function and develop IVA-CMGGD in the complex domain. Since the cost function has no closed-form solution, we use Wirtinger calculus to derive the gradient and resort to 3 numerical optimization techniques: steepest descent (SD), quasi-Newton Broyden-Fletcher-Goldfarb-Shanno (BFGS), and limited-memory BFGS (L-BFGS), with a step size satisfying the Wolfe conditions (Sorber et al, 2012; Nocedal and Wright, 2006). The result is 3 new IVA algorithms: IVA-CMGGD-SD, IVA-CMGGD-BFGS, and IVA-CMGGD-L-BFGS respectively. To demonstrate the efficacy of our approach, we compare their performance with that of competing algorithms applied to both complex-valued simulated sources as well as a real-world, 14-subject complex-valued fMRI finger-tapping dataset used in a number of other studies (Rodriguez et al, 2012, 2011; Li et al, 2011; Xiong et al, 2012). In exploratory fMRI analysis, little is known about the distributions of the latent sources and the data, as collected, are natively complex-valued (Rodriguez et al, 2015). The phase of the complex-valued fMRI data is often dropped in favor of real-valued processing of the magnitude data alone, even though studies have shown the phase to contain novel physiologically-relevant information (Feng et al, 2009; Rowe, 2005; Arja et al, 2010). The phase images are discarded primarily because their unfamiliar and noisy nature poses a challenge when studying fMRI data (Calhoun et al, 2002; Adalı and Calhoun, 2007). Therefore, exploiting the information contained in the phase promises to increase our ability to identify areas in the brain with significant susceptibility changes that could not be previously identified by operating only on the magnitude data. Additionally, it is natural to assume that dependencies exist in neural behavior in multiple subjects under similar conditions (Dea et al, 2011). These facts motivate the need for flexible, data-driven techniques for joint analysis in the complex domain utilizing the full extent of statistical information available in simple form—an excellent match for our proposed algorithms.

The rest of the paper is organized as follows: Section 2 presents background material and the theoretical development of our proposed method. In Section 3, we introduce the experimental setup for simulated data and the real-world fMRI dataset and present the results. Section 4 concludes the paper.

2 Theory and Methods

2.1 Complex-Valued RVs

The probability density function (PDF) of a complex-valued random vector (RV) $\mathbf{x} = \mathbf{x}_r + j\mathbf{x}_i \in \mathbb{C}^K$, where $\mathbf{x}_r, \mathbf{x}_i \in \mathbb{R}^K$, is defined as the joint probability density function of its real and imaginary components:

$$p_{\mathbf{x}}(\mathbf{x}) \coloneqq p_{\mathbf{x}_r,\mathbf{x}_i}(\mathbf{x}_r, \mathbf{x}_i). \tag{1}$$

The real composite representation $\mathbf{x}_{\mathbb{R}} = \left[\mathbf{x}_r^\top, \mathbf{x}_i^\top\right]^\top$ fully captures the behavior of a random vector over its domain. The complex augmented form (Schreier and Scharf, 2010), given by $\left[\mathbf{x}^\top, \mathbf{x}^H\right]^\top$, where $(\cdot)^H$ is the Hermitian operator, is an

equivalent redundant representation related to the real composite form through an invertible linear transform (Schreier and Scharf, 2010; Adalı et al, 2011b),

$$\underline{\mathbf{x}} = \mathbf{T}_K \mathbf{x}_{\mathbb{R}} \Leftrightarrow \mathbf{x}_{\mathbb{R}} = \frac{1}{2}\mathbf{T}_K^H \underline{\mathbf{x}}, \tag{2}$$

where $\mathbf{T}_K = \begin{bmatrix} \mathbf{I} & j\mathbf{I} \\ \mathbf{I} & -j\mathbf{I} \end{bmatrix} \in \mathbb{C}^{2K \times 2K}$ is unitary up to a factor[1] of 2 and $\mathbf{I}$ is the identity matrix. In this paper, we omit the subscript in $\mathbf{T}_K$ when it is understood from the context.

Although redundant, the complex augmented form offers a number of advantages when manipulating complex RVs. For instance, since the PDF of a complex-valued random vector $\mathbf{x}$ is necessarily a real-valued function of its argument (as are all optimize-able cost functions), Wirtinger, or $\mathbb{CR}$ calculus (Schreier and Scharf, 2010; Adalı et al, 2011a; Brandwood, 1983; Wirtinger, 1927; Adalı and Schreier, 2014) can be applied to take derivatives of the PDF with respect to $\mathbf{x}$ or $\mathbf{x}^*$ while keeping the other constant. This avoids the tedious task of differentiating with respect to each of the real and imaginary components separately. And by maintaining its complex structure, this approach also retains the intuition of the RVs in their natural domain.

The complex augmented form also allows for capturing full complex second-order statistics implicitly. To illustrate this, consider the augmented covariance matrix of $\mathbf{x} \in \mathbb{C}^K$ given by:

$$\underline{\mathbf{C}} \coloneqq \mathbb{E}\left\{\underline{\mathbf{x}}\, \underline{\mathbf{x}}^H\right\} = \mathbb{E}\left\{\begin{bmatrix} \mathbf{x}\mathbf{x}^H & \mathbf{x}\mathbf{x}^\top \\ \mathbf{x}^*\mathbf{x}^H & \mathbf{x}^*\mathbf{x}^\top \end{bmatrix}\right\}, \tag{3}$$

where $(\cdot)^*$ is the complex conjugate operator. It is clear that the covariance matrix of the complex augmented form captures the covariance matrix $\mathbf{C} = \mathbb{E}\left\{\mathbf{x}\mathbf{x}^H\right\}$ as well as the pseudo-covariance matrix $\mathbf{P} = \mathbb{E}\left\{\mathbf{x}\mathbf{x}^\top\right\}$ of the complex RV. We note the distinct block pattern of $\underline{\mathbf{C}}$: the northwest block is the complex conjugate of the southeast block, and the northeast block is the complex conjugate of the southwest block. Also, an augmented covariance matrix is Hermitian, and, like its real-valued analog, positive semi-definite (Haykin, 2014).

Following definitions in (Schreier and Scharf, 2010), we call an random vector $\mathbf{x}$ proper if the pseudo-covariance matrix vanishes and improper otherwise. To quantify the impropriety of random vector $\mathbf{x}$, the noncircularity coefficients, defined as the singular values of the coherence matrix[2],

$$\mathbf{R} = \mathbf{C}^{-1/2}\mathbf{P}\mathbf{C}^{-H/2}, \tag{4}$$

are computed. Then, the noncircularity coefficients can be combined into a single metric of impropriety in a number of ways each possessing different attributes. We opt for the following (Schreier and Scharf, 2010) for degree of impropriety (DOI):

$$\rho = \frac{1}{K}\sum_{k=1}^{K} \lambda_k^2, \tag{5}$$

[1] Unitary up to a factor of 2 implies $\mathbf{T}\mathbf{T}^H = \mathbf{T}^H\mathbf{T} = 2\mathbf{I}$.

[2] The classical singular value decomposition is sufficient to obtain the noncircularity coefficients. However, if the corresponding canonical projections are needed, a special, complex-symmetric decomposition called the Takagi factorization (Horn and Johnson, 1990; Schreier and Scharf, 2010; Moreau and Adalı, 2013) is required to maintain the complex augmented form.

where λ_k is the k^{th} singular value of $\mathbf{R}$. This choice possesses the desirable property that when $\rho = 1,\ \lambda_k = 1,\ k \in \{1, \ldots, K\}$, i.e., each of the marginal variates is maximally-improper. Similarly, when $\rho = 0,\ \lambda_k = 0,\ k \in \{1, \ldots, K\}$, each is accordingly proper.

2.2 The CMGGD Family

In a number of applications such as fMRI analysis, the components of interest are highly noncircular (Li et al, 2011), and hence it is desirable to use a multivariate density model that allows for noncircularity. MGGD, as we noted, provides a desirable balance between complexity and flexibility. It has unimodal marginals that can model distributions with heavier or lighter tails than the Gaussian distribution using a single shape parameter thus able to decrease the bias due to PDF mismatch while allowing for reasonable complexity with a small number of paarmeters. The circular complex-valued generalized Gaussian distribution is given by (Ollila et al, 2012):

$$p_{\text{GG}}(\mathbf{x}) = \frac{\beta \Gamma(K) b^{-K/\beta}}{\pi^K \Gamma\left(K/\beta\right)} \left|\boldsymbol{\Sigma}\right|^{-1} \exp\left\{-\frac{1}{b}\left(\mathbf{x}^H \boldsymbol{\Sigma}^{-1} \mathbf{x}\right)^{\beta}\right\}, \tag{6}$$

where b is a scale parameter and $\boldsymbol{\Sigma} \in \mathbb{C}^{K \times K}$ is a positive semi-definite matrix proportional to the classical covariance matrix for fixed shape parameter β and b. It is clear that this definition neglects noncircularity by omitting the pseudo-covariance matrix. Instead, we introduce the complex augmented form to incorporate the pseudo-covariance (for details, see (Mowakeaa et al, 2016)):

$$p(\mathbf{x}) = \frac{K \Gamma(K) \eta^K 2^{-K/\beta}}{\pi^K \Gamma\left(1 + \frac{K}{\beta}\right)} \left|\underline{\mathbf{C}}\right|^{-1/2} \exp\left\{-\frac{1}{2}\left(\frac{\eta}{2} \underline{\mathbf{x}}^H \underline{\mathbf{C}}^{-1} \underline{\mathbf{x}}\right)^{\beta}\right\}, \tag{7}$$

where $\Gamma(\cdot)$ is the gamma function, $|\cdot|$ represents the determinant when its argument is a matrix, and η is a normalization factor unique to the specific member of the CMGGD family and is given by:

$$\eta = \frac{2^{\frac{1}{\beta}-1} \Gamma\left(\frac{K+1}{\beta}\right)}{K \Gamma\left(\frac{K}{\beta}\right)}. \tag{8}$$

It is straightforward to show that if the pseudo-covariance matrix vanishes, the PDF in (7) is simply the PDF in (6) in augmented form. Hence, CMGGD is a generalization of the generalized Gaussian distribution to incorporate noncircularity. The flexibility of the CMGGD family of distributions provides an ideal tool for IVA providing more flexibility which makes it attractive for many real-world applications.

2.3 Generative Model

IVA, a multiset extension of ICA, can perform decompositions on a number of datasets *jointly* by taking into account the dependence between corresponding sources across datasets (Kim et al, 2006). In real-world applications, the number of

mixtures often exceed the number of sources to be estimated. Principal component analysis (PCA) is a common preprocessing step used to reduce the number of mixtures to match the number of desired components to be estimated (Hyvärinen et al, 2001; Michael et al, 2014; Lee et al, 2008; Calhoun and Adalı, 2012; Wax and Kailath, 1985). It also avoids over-fitting by reducing the data to the signal subspace in noisy applications such as fMRI. The success of this step, however, is tied to the accuracy with which the enforced model order matches the true order of the signal subspace. We then write for each dataset

$$\mathbf{x}^{[k]} = \mathbf{A}^{[k]}\mathbf{s}^{[k]}, \quad k \in 1, \cdots, K, \tag{9}$$

where $\mathbf{x}^{[k]} \in \mathbb{C}^N$ is the mixture random vector from the k^{th} dataset, $\mathbf{A}^{[k]} \in \mathbb{C}^{N\times N}$ is the k^{th} mixing matrix, and $\mathbf{s}^{[k]}$ is the k^{th} latent source RV. We define a source component vector (SCV) by cascading corresponding estimated sources from each dataset row-wise: $\mathbf{s}_n = \left[s_n^{[1]}, \ldots, s_n^{[K]}\right]^\top$. The goal of IVA is to estimate K demixing matrices that produce source estimates:

$$\mathbf{y}^{[k]} = \mathbf{W}^{[k]}\mathbf{x}^{[k]}, \quad k \in 1, \cdots, K, \tag{10}$$

where $\mathbf{W}^{[k]}$ is the estimate of the k^{th} demixing matrix, that are as independent as possible.

One approach to solving (10) is to minimize the mutual information (MI) among estimated SCVs. The MI framework is a natural measure of the degree of statistical independence among SCVs and can incorporate all statistical forms of diversity. Then, the MI of N linearly-mixed independent and identically distributed (IID) complex-valued sources can be defined as (Anderson et al, 2012a):

$$\mathcal{I}\left(\mathbf{y}_1; \ldots; \mathbf{y}_N\right) = \sum_{n=1}^{N} \mathcal{H}\left(\mathbf{y}_n\right) - \sum_{k=1}^{K} \left(\log\left|\mathbf{W}^{[k]}\right|^2\right) - \mathcal{H}\left(\boldsymbol{\mathcal{X}}\right), \tag{11}$$

where $\mathcal{I}\left(\cdot\right)$ is the MI function, $\mathcal{H}\left(\cdot\right)$ denotes the (differential) entropy, $\boldsymbol{\mathcal{X}} = \left[\mathbf{x}^{[1]\top}, \ldots, \mathbf{x}^{[K]\top}\right]^\top \in \mathbb{C}^{NK\times 1}$ is the collection of all $\mathbf{x}^{[k]}$, and $\mathbf{y}_n \in \mathbb{C}^K$ is the estimate of the n^{th} SCV with elements $y_n^{[k]}$ given by:

$$y_n^{[k]} = \mathbf{w}_n^{[k]\top}\mathbf{x}^{[k]}, \tag{12}$$

where $\mathbf{w}_n^{[k]}$ is defined to be the *column* vector of elements of the n^{th} row of $\mathbf{W}^{[k]}$. The last term in (11) is constant with respect to $\mathbf{W}^{[k]}$ and can be dropped from the optimization procedure. The mutual information cost function (11) we develop here does not require the mixtures to be pre-whitened. This approach is used in a number of other works (Anderson et al, 2012a; Kim et al, 2006; Adalı et al, 2014). However, pre-whitening the data approximately resolves about half the unknowns in the demixing matrices (Cardoso, 1998) hence, in practice, it is advantageous to do so to improve convergence speed.

Minimization of the MI is equivalent to the maximization of the maximum likelihood cost function (Cardoso, 1998), making available all the theoretical advantages associated with maximum likelihood theory. As the model deviates from the true PDF, a bias is introduced in the estimate of the demixing matrix resulting into poor estimation performance. This fact emphasizes the need for flexible

modeling of source distributions in a data-driven fashion. Equally important is the choice of optimization approaches since the solution to (11) does not exist in closed form. This is the subject of Sec. 2.4.

2.4 IVA-CMGGD Cost Function and Gradient

By using the MI cost function, the IVA problem can be stated as:

$$\arg\min_{\mathbf{W}} \; J(\mathbf{W}), \tag{13}$$

where $J(\mathbf{W})$ is the mutual information-based cost function and $\mathbf{W} \in \mathbb{C}^{N\times N\times K}$ is a tensor containing all $\mathbf{W}^{[k]}$. Motivated by its inclusion of noncircularity and its simple parametric form, we deploy the CMGGD PDF in (7) to the MI cost function in (11). The resulting cost function becomes:

$$J(\mathbf{W}) = \sum_{n=1}^{N}\left[\kappa_n + \frac{1}{2}\log\left(|\underline{\mathbf{C}}_n|\right) + \frac{1}{2}\mathbb{E}\left\{\left(\frac{\eta_n}{2}\underline{\mathbf{y}}_n^H\underline{\mathbf{C}}_n^{-1}\underline{\mathbf{y}}_n\right)^{\beta_n}\right\}\right] - \sum_{k=1}^{K}\left(\log\left|\mathbf{W}^{[k]}\right|^2\right), \tag{14}$$

where $\kappa_n = -\log\left(\frac{K\Gamma(K)\eta_n^K}{\pi^K\Gamma\left(1+\frac{K}{\beta_n}\right)2^{\frac{K}{\beta_n}}}\right)$. Since distribution parameters $\underline{\mathbf{C}}_n$ and β_n are estimated from the data, we can drop terms in (14) independent of $\mathbf{W}^{[k]}$ to yield:

$$J(\mathbf{W}) = \frac{1}{2}\sum_{n=1}^{N}\mathbb{E}\left\{\left(\frac{\eta_n}{2}\underline{\mathbf{y}}_n^H\underline{\mathbf{C}}_n^{-1}\underline{\mathbf{y}}_n\right)^{\beta_n}\right\} - \sum_{k=1}^{K}\left(\log\left|\mathbf{W}^{[k]}\right|^2\right). \tag{15}$$

It is clear in (15) that both the covariance and pseudo-covariance, hence noncircularity, are exploited in the cost function[3]. In practice, the limited number of samples available may limit the performance of IVA—especially as model flexibility is increased. We expand on this further in Section 3.

Since no closed-form solution to (15) exists, we derive the gradient $\nabla J(\mathbf{W})$ for use in iterative numerical approaches, where $[\nabla J(\mathbf{W})]_{n,k} = \frac{\partial J(\mathbf{W})}{\partial \mathbf{w}_n^{[k]*}} \in \mathbb{C}^{1\times N}$. We use Wirtinger calculus to differentiate $J(\mathbf{W})$ with respect to $\mathbf{w}_n^{[k]*}$ while considering $\mathbf{w}_n^{[k]}$ as a constant. Generally, differentiating the log-determinant term in (15) includes matrix inversion at each iteration which can lead to numerical error. Therefore, we utilize the decoupling trick (Anderson et al, 2012b; Fu et al, 2015) which permits computing the derivative of the log-determinant term in (15) row-wise without inversion. Then, the gradient can be written as:

$$\frac{\partial J(\mathbf{W})}{\partial \mathbf{w}_n^{[k]*}} = \mathbb{E}\left\{\frac{\beta_n\eta_n^{\beta_n}}{2^{\beta_n+1}}\frac{\left(\mathbf{e}_k^\top\underline{\mathbf{C}}_n^{-1}\underline{\mathbf{y}}_n\right)\mathbf{x}^{[k]*}}{\left(\underline{\mathbf{y}}_n^H\underline{\mathbf{C}}_n^{-1}\underline{\mathbf{y}}_n\right)^{(1-\beta_n)}}\right\} + \frac{\mathbf{h}_n^{[k]*}}{\mathbf{h}_n^{[k]H}\mathbf{w}_n^{[k]*}}, \tag{16}$$

where $\mathbf{e}_k$ is a unit vector of appropriate length with a value of 1 in the k^{th} position, and 0 elsewhere, and $\mathbf{h}_n$ is *any* unit-norm vector orthogonal to $\mathbf{w}_{n_o}^{[k]}$, $n_o \in \{1,\ldots,n-1,n+1,\ldots,N\}$.

[3] We maintain unit-norm rows of $\mathbf{W}$ to prevent driving the cost function lower through scaling alone.

2.5 IVA-CMGGD Optimization Procedure

To minimize (15), since no closed-form solution exists, we employ iterative numerical optimization line search methods of the following form:

$$\mathbf{W} \leftarrow \mathbf{W} + \mu \mathbf{D}, \tag{17}$$

where $\mathbf{D} \in \mathbb{C}^{N \times N \times K}$ is an appropriate tensor direction that forms an obtuse angle with the gradient $\nabla J(\mathbf{W})$ and μ is a step size that satisfies the Wolfe conditions to ensure rapid convergence (Nocedal and Wright, 2006; Sorber et al, 2012). In this context, the obtuse angle is interpreted in a row-wise sense.

Steepest descent (SD), the simplest line search algorithm, uses the unit-direction $\mathbf{D} = -\frac{\nabla J(\mathbf{W})}{\|\nabla J(\mathbf{W})\|}$ to decrease the cost function locally in the Euclidean space. This approach is attractive due to its simplicity but might lead to an very slow rate of convergence even when the Hessian is reasonably well-conditioned (Nocedal and Wright, 2006).

Newton methods improve on SD for up to quadratic convergence (Nocedal and Wright, 2006) by taking into account both the gradient as well as the Hessian. However, where the Hessian is difficult to derive or expensive to compute, quasi-Newton approaches, such as BFGS, offer super-linear convergence using a positive definite Hessian estimate. At iteration i, given $\mathbf{u} = \mathrm{vec}\left(\mathbf{W}^{(\mathrm{i}+1)} - \mathbf{W}^{(\mathrm{i})}\right)$, where $\mathrm{vec}(\cdot)$ serializes its argument in a vector, and $\mathbf{v} = \mathrm{vec}\left(\nabla J^{(\mathrm{i}+1)} - \nabla J^{(\mathrm{i})}\right)$, we define the augmented forms $\underline{\mathbf{u}} = \left[\mathbf{u}^\top \mathbf{u}^H\right]^\top$ and $\underline{\mathbf{v}} = \left[\mathbf{v}^\top \mathbf{v}^H\right]^\top$. Then, the inverse Hessian can be updated in augmented form by (Sorber et al, 2012):

$$\underline{\mathbf{H}}^{-1} \leftarrow \left(\mathbf{I} - \frac{\underline{\mathbf{u}}\,\underline{\mathbf{v}}^H}{\underline{\mathbf{v}}^H \underline{\mathbf{u}}}\right) \underline{\mathbf{H}}^{-1} \left(\mathbf{I} - \frac{\underline{\mathbf{v}}\,\underline{\mathbf{u}}^H}{\underline{\mathbf{v}}^H \underline{\mathbf{u}}}\right) + \frac{\underline{\mathbf{u}}\,\underline{\mathbf{u}}^H}{\underline{\mathbf{v}}^H \underline{\mathbf{u}}}, \tag{18}$$

under the condition $\underline{\mathbf{v}}^H \underline{\mathbf{u}} > 0$. This condition is guaranteed to hold when the Wolfe conditions are used to select the step size (Nocedal and Wright, 2006). Then, the vectorized BFGS direction can be written as:

$$\underline{\mathbf{d}} = -\underline{\mathbf{H}}^{-1} \underline{\nabla J}^*, \tag{19}$$

where $\underline{\nabla J}$ is the augmented form of the vectorized gradient and $\underline{\mathbf{d}}$ is the augmented form of $\mathrm{vec}(\mathbf{D})$.

Still, the cost of storing the Hessian estimate itself may become prohibitive as the number of sources and the number of datasets increase. Limited-memory BFGS (L-BFGS) alleviates the need to store the entire Hessian by storing the previous M gradient estimates which are then used to update the augmented inverse Hessian implicitly. In this case, the L-BFGS direction is given by Algorithm 1 (Sorber et al, 2012), where $\Re\{\cdot\}$ extracts the real component of a complex argument.

After finding a suitable descent direction, a satisfactory step size μ must be selected in order to assure convergence. Several criteria exist, such as the Wolfe conditions which are given by (Sorber et al, 2012):

$$\begin{aligned} J(\mathbf{W} + \mu\mathbf{D}) &\leq J(\mathbf{W}) + c_1 \mu \underline{\mathbf{d}}^\top \underline{\nabla J}(\mathbf{W}), \text{ and} \\ \underline{\mathbf{d}}^\top \underline{\nabla J}(\mathbf{W} + \mu\mathbf{D}) &\geq c_2 \underline{\mathbf{d}}^\top \underline{\nabla J}(\mathbf{W}), \end{aligned} \tag{20}$$

Algorithm 1 L-BFGS update at iteration i (Sorber et al, 2012)

Require: $\mathbf{u}_m$, $\mathbf{v}_m$, $m \in \{i-1, i-2, \ldots, i-M\}$
$\gamma_m \leftarrow \Re\left\{\mathbf{v}_m^H \mathbf{u}_m\right\}^{-1}$
$\mathbf{g} \leftarrow 2 \operatorname{vec}\left(\frac{\partial J(\mathbf{W})}{\partial \mathbf{W}^*}\right)$
$\mathbf{d} \leftarrow -\mathbf{g}$
for $m = i-1, i-2, \ldots, i-M$ **do**
    $\alpha_m \leftarrow \gamma_m \Re\left\{\mathbf{u}_m^H \mathbf{d}\right\}$
    $\mathbf{d} \leftarrow \mathbf{d} - \alpha_m \mathbf{v}_m$
$\mathbf{d} \leftarrow \frac{1}{2}\mathbf{H}_{i-M}^{-1}\mathbf{d}$
for $m = i-M, i-M+1, \ldots, i-1$ **do**
    $\beta \leftarrow \gamma_m \Re\left\{\mathbf{v}_m^H \mathbf{d}\right\}$
    $\mathbf{d} \leftarrow \mathbf{d} + (\gamma_m - \beta)\,\mathbf{u}_m$
return d

where c_1 and c_2 are constants usually taken to be 10^{-4} and 0.9 respectively[4]. The first condition in (20) is called the sufficient decrease condition and is responsible for ensuring that the step size selected leads to a substantial decrease in the cost function while the second condition, known as the curvature condition, disqualifies step sizes that are too small. Usually, finding a step size that satisfies the Wolfe conditions is performed by interpolating the cost function between iterates until one is found. This typically involves many function and gradient evaluations at each iteration of the algorithm. To reduce the impact of these computations, we fix a set of step sizes over which the Wolfe conditions are evaluated. If none of the step sizes satisfy the conditions at a given iteration, we pick the step size that leads to the largest decrease in the value of the cost function. Convergence of IVA-CMGGD is attained when the magnitude of the gradient falls below threshold ϵ_1 while the magnitude change in the distribution parameters is below threshold ϵ_2. To avoid shrinking the cost function through the scaling of $\mathbf{W}$, we constrain each row $\mathbf{w}_n^{[k]}$ to be unit-norm by projecting each row of $\mathbf{W}$ onto the unit-sphere after each update. In all IVA-CMGGD variants, we estimate the distribution parameters during the optimization procedure every q iterations to ensure significant progress in decreasing the cost function at the current set of distribution parameters before updating them. Heuristically, we choose $q = \lfloor\sqrt{N^2K}\rfloor$, where $\lfloor\cdot\rfloor$ is the floor function. Algorithm 2 summarizes the IVA-CMGGD procedure where the distribution parameters for the n^{th} SCV are denoted compactly by $\boldsymbol{\Theta}_n = \left[\operatorname{vec}(\mathbf{C}_n)^\top \operatorname{vec}(\mathbf{P}_n)^\top \beta_n\right]^\top$.

3 Results

In this section, we apply IVA-CMGGD to both mixtures of simulated CMGGD sources as well as real-world complex-valued fMRI data. We conduct a number of experiments that highlight the performance advantage of the proposed methods. In the first set of experiments, we generate synthetic sources over the full range of the shape parameter $\beta \in [0.125, 8]$, using the CMGGD data generation method in (Mowakeaa et al, 2016) that allows for noncircular variables, to demonstrate the full capability of IVA-CMGGD. Then, to better accommodate

[4] For the conjugate gradient method, not discussed in this paper, the value of c_2 is often taken to be 0.1 (Nocedal and Wright, 2006).

Algorithm 2 IVA-CMGGD

Require: $\mathbf{X}^{[1]}, \ldots, \mathbf{X}^{[K]}$ (pre-whitening), $\mathbf{W}_0^{[1]}, \ldots, \mathbf{W}_0^{[K]}$ (unit-norm row), $\epsilon_1 > 0$, $\epsilon_2 > 0$
Whiten $\mathbf{X}^{[k]}$ for all k
$\mathbf{W} \leftarrow \mathbf{W}_0$
Estimate distribution parameters $\boldsymbol{\Theta}_n$ for all n
repeat
  Find tensor descent direction $\mathbf{D}$ using SD, BFGS, or L-BFGS
  Find μ satisfying Wolfe conditions
  $\mathbf{W} \leftarrow \mathbf{W} + \mu\mathbf{D}$
  $\mathbf{w}_n^{[k]} \leftarrow \frac{\mathbf{w}_n^{[k]}}{\left\|\mathbf{w}_n^{[k]}\right\|}$ for all n, k
  if iter $\pmod q = 1$ **then**
    Update distribution parameters $\boldsymbol{\Theta}_n$ for all n
  if $\sum_{n,k} \left\| \frac{\partial J(\mathbf{W})}{\partial \mathbf{w}_n^{[k]*}} \right\| < \epsilon_1$ & $\|\Delta\boldsymbol{\Theta}_n\| < \epsilon_2$ **then**
    Break
until Max iteration
Transform $\mathbf{W}^{[k]}$ to the domain of pre-whitened $\mathbf{X}^{[k]}$ for all k
return W

the most prominent competing algorithm, IVA-GL, which uses a Laplacian PDF model in its second stage following a Gaussian model in its first, we generate super-Gaussian sources with $\beta \in [0.125, 0.5]$. We omit IVA-G in this case due to its model mismatch for clarity. In the next set of experiments, we study performance on actual real-world complex-valued fMRI data where sources have been shown to be mostly super-Gaussian for physiologically-relevant components and both super- and sub-Gaussian in the case of artifacts (Lee et al, 1999; Calhoun and Adalı, 2006; Girolami, 1998). This allows us to test the data-driven flexibility of IVA-CMGGD variants in modeling of the unknown source distributions. We compare the proposed method IVA-CMGGD, in its three flavors SD, BFGS, and L-BFGS, against IVA-GL. Adaptive IVA was omitted after failing to produce competing results even after matching the used tolerance of 10^{-6} as in (Kuang et al, 2017) and discarding runs that met a maximum iteration count of $20,000$. (In comparison, the authors in (Kuang et al, 2017) use a maximum iteration of $1,000$.) This might be due to the sub-optimal estimation of each source covariance matrix within a 1-dimensional subspace spanned by its dominant eigenvector as well as the assumption of orthogonality on each demixing matrix (Kuang et al, 2017).

3.1 Simulated CMGGD Sources

In this section, we use simulated data to demonstrate the performance of IVA-CMGGD. For the following simulated experiments, the covariance matrix for each SCV is a random symmetric positive-definite matrix $\mathbf{C} = \mathbf{A}^H\mathbf{A}$ and the pseudo-covariance matrix is generated as $\mathbf{P} = \mathbf{C}^{1/2}\mathbf{F}\boldsymbol{\Lambda}\mathbf{F}^\top\mathbf{C}^{\top/2}$, where $\mathbf{A}$ has elements drawn from the standard complex Gaussian distribution, $\mathbf{F}$ is any unitary matrix and $\boldsymbol{\Lambda}$ is the diagonal matrix of noncircularity coefficients (Schreier and Scharf, 2010). Since the ground truth, i.e., both underlying sources and mixing matrices, are known, we use inter-symbol-interference (ISI) (Moreau and Macchi, 1994) to

find the average performance over a number of runs and is given by (Anderson et al, 2012a; Moreau and Macchi, 1994),

$$\mathrm{ISI}(\mathbf{G}) = \frac{1}{2N(N-1)}\left[\sum_{n=1}^{N}\left(\sum_{m=1}^{N}\frac{|g_{n,m}|}{\max_p |g_{n,p}|} - 1\right) + \sum_{m=1}^{N}\left(\sum_{n=1}^{N}\frac{|g_{n,m}|}{\max_p |g_{p,m}|} - 1\right)\right], \tag{21}$$

where $\mathbf{G} = \mathbf{WA}$ with elements $g_{n,m}$. ISI measures the quality of a decomposition by describing the deviation of $\mathbf{G}$ from the identity matrix. For multi-dataset decompositions, ISI can be generalized to joint ISI (jISI) for IVA as in (Anderson et al, 2012a):

$$\mathrm{ISI}_{\mathrm{joint}}(\mathcal{G}) = \mathrm{ISI}(\tilde{\mathbf{G}}) \tag{22}$$

where $\mathcal{G} = \left\{\mathbf{G}^{[k]}\right\}_{k=1}^{K}$ is the collection of matrices $\mathbf{G}^{[k]}$ and the $(i,j)^{\mathrm{th}}$ element of $\tilde{\mathbf{G}}$ is the absolute sum of corresponding elements of $\mathbf{G}^{[k]}$ over all k as in:

$$\left[\tilde{\mathbf{G}}\right]_{i,j} = \sum_{k=1}^{K}\left|\left[\mathbf{G}^{[k]}\right]_{i,j}\right|. \tag{23}$$

Joint ISI in is thus a generalization of ISI to multiple datasets and they are equivalent when $K = 1$.

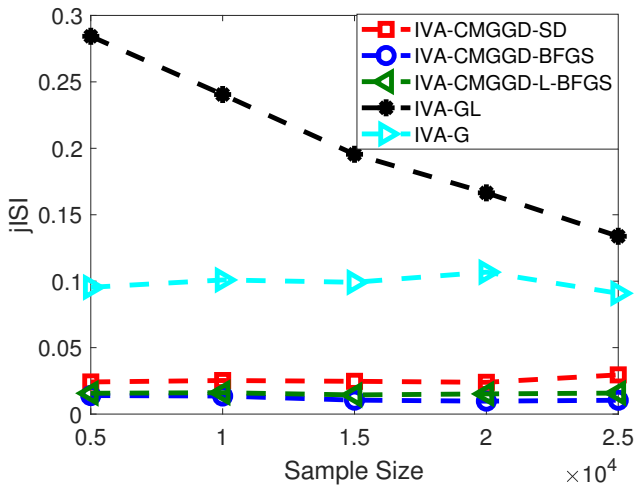

Fig. 1 Joint ISI for decomposing mixtures of $N = 3$ sources over $\beta \in [0.125, 8]$ and $K = 4$ datasets with DOI = 0.5. Each point is the average of 300 independent runs

In the first experiment, we generate $N = 3$ sources with $K = 4$ datasets over the full range of the shape parameter $\beta \in [0.125, 8]$ as we vary the number of samples over the set $V \in \{5, 10, 15, 20, 25\} \times 1000$ samples. Thus, each SCV is multivariate with dimension 4, i.e., $\mathbf{S}_n \in \mathbb{C}^{4\times V}$. The sources from each dataset are mixed using matrices $\mathbf{A}^{[k]}$ with elements drawn from the uniform distribution $\mathcal{U}(0, 1)$. Fig. 1 shows the results of this experiment averaged over 300 independent

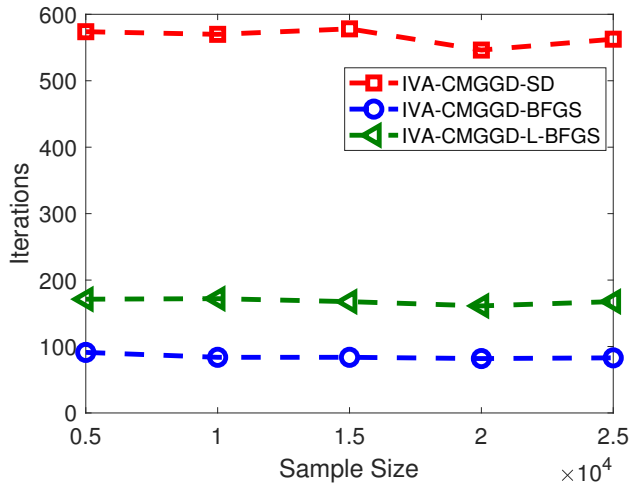

Fig. 2 Number of iterations required per algorithm for mixtures of $N = 3$ sources over $\beta \in [0.125, 8]$ and $K = 4$ datasets with DOI $= 0.5$. Each point is the average of 300 independent runs

runs. The broad parameter shape range naturally favors IVA-CMGGD variants in this experiment. We also report the number of iterations required for convergence for each of the IVA-CMGGD variants in Fig. 2. Here, as suspected, the convergence of IVA-CMGGD-BFGS requires considerably fewer iterations than IVA-CMGGD-SD, whereas IVA-CMGGD-L-BFGS, with its memory-efficient implementation, trails closely behind.

In the second experiment, we restrict the shape parameter to the super-Gaussian range $\beta \in [0.125, 0.5]$. Fig. 3 shows the results of this experiment averaged over 300 independent runs. It is clear from this figure that while all methods show a decrease in jISI as the number of samples increases, IVA-CMGGD methods are better able to model the super-Gaussian sources via their adaptive nonlinearity. The L-BFGS variant of IVA-CMGGD shows slightly reduced performance versus its more efficient siblings. This is due to the noisy nature of the descent direction approximation inherent to this flavor of IVA-CMGGD.

In the third experiment, we fix the number of samples to $V = 10,000$ samples and vary DOI over $\rho \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ while keeping other parameters as before. Fig. 4 shows the result of this experiment averaged over 300 independent runs. Here, the case for taking full statistical information into account is apparent as IVA-CMGGD variants outperform IVA-GL. We also note that the worst performance achieved by each algorithm occurs in a neighborhood of $\rho = 0.5$. As DOI increases from this point, the correlation between the real and imaginary parts increases, reducing the effective degrees of freedom of the sources. On the other hand, as DOI decreases from this point, the pseudo-covariance matrix vanishes, decreasing the effective degrees of freedom of the parameter space of each estimated source PDF along with it. Both limiting cases thus improve overall estimation performance.

The increase in run-time associated with the additional complexity of IVA-CMGGD is on the order of 5–10 times that of IVA-GL for the experiments presented here due to the increased model complexity and hence computational considerations are one of the factors that might affect the decision to select a more powerful model for a given problem.

3.2 Complex-Valued fMRI Data

We apply IVA-CMGGD to a complex-valued fMRI dataset used in a number of studies (Rodriguez et al, 2012, 2011; Li et al, 2011; Xiong et al, 2012). As detailed in (Rodriguez et al, 2012), the dataset consists of fMRI scans from 14 subjects performing a finger-tapping motor task with alternating periods of 30 s *ON* (finger tapping) and 30 s *OFF* (rest). The experiments were performed on a 3 T Siemens TRIO TIM system with a 12-channel radio-frequency (RF) coil. The fMRI experiment used a standard Siemens gradient echo planar imaging (EPI) sequence modified to store real and imaginary data separately. The data were preprocessed to correct for phase error, motion correction and spatial normalization into standard Montreal Neurological Institute space using the MATLAB toolbox for statistical parametric mapping[5] (SPM). We perform PCA to reduce the dimensionality of the data to an order of 30 components estimated by the minimum description length (MDL) criterion for complex-valued data as in (Rodriguez et al, 2012; Xiong et al, 2008).

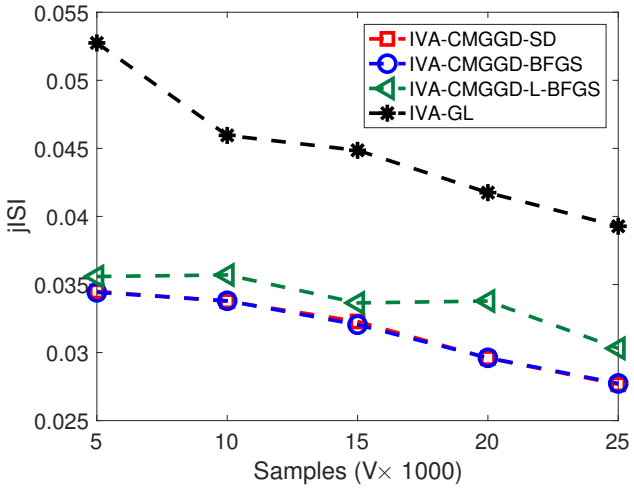

Fig. 3 Joint ISI for decomposing mixtures of $N = 3$ super-Gaussian sources and $K = 4$ datasets with DOI = 0.5. Each point is the average of 300 independent runs

Since the ground truth is not available for real-world data, ISI cannot be used to evaluate the performance of a single run. Instead, we adopt a couple of commonly used methods including analysis of average spatial maps and average TC

[5] SPM URL: http://www.fil.ion.ucl.ac.uk/spm/software/spm12/

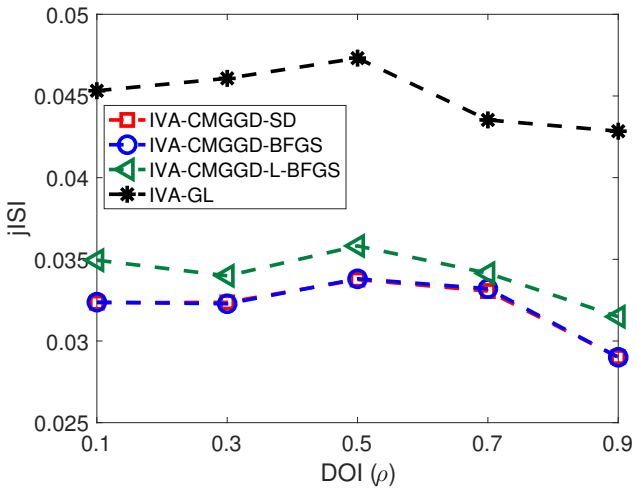

Fig. 4 Joint ISI for decomposing mixtures of $N = 3$ super-Gaussian sources and $K = 4$ datasets with $V = 10,000$ samples. Each point is the average of 300 independent runs

correlation with the task paradigm (Du et al, 2016; Rodriguez et al, 2012). In (Rodriguez et al, 2012, 2011), the authors develop a visualization technique for fMRI spatial maps that takes into account both the real and imaginary values of the estimated components of each complex-valued voxel through the Mahalanobis distance:

$$\mathbf{z}(v) = \sqrt{(\mathbf{y}_{\mathbb{R}} - \boldsymbol{\mu}_{\mathbb{R}})^{\top} \mathbf{C}_{\mathbb{R}}^{-1} (\mathbf{y}_{\mathbb{R}} - \boldsymbol{\mu}_{\mathbb{R}})}, \tag{24}$$

where $\mathbf{z}(v)$ is the real value assigned to the v^{th} voxel of estimated source component $\mathbf{y}$, $\mathbf{y}_{\mathbb{R}}$ is the real composite form of $\mathbf{y}$ and $\boldsymbol{\mu}_{\mathbb{R}}, \mathbf{C}_{\mathbb{R}}$ are the mean and covariance matrix of $\mathbf{y}_{\mathbb{R}}$ respectively. Estimated TCs are correlated with the paradigm TC after convolving it with the blood oxygenation level dependent (BOLD) haemodynamic response using SPM.

To evaluate the performance of each algorithm, we first remind that averaging resulting components or timecourses across runs cannot be used in decomposition tasks where the goal is interpretation of each component, and instead one should opt for finding and using a most representative run to evaluate the performance (Himberg and Hyvarinen, 2003; Rachakonda et al, 2007). On all but relatively small datasets, performing IVA on a large number of runs is computationally prohibitive—especially for IVA-CMGGD. Therefore, we require a computationally-efficient method to adjudicate a relatively small sample of runs in search of its most representative or *central* member—in some meaningful sense. ICASSO, a multiple-run consistency and visualization technique, tackles this problem through evaluating the quality of clustering of corresponding SCVs from different runs (Himberg and Hyvarinen, 2003). However, the resulting centroids fail to represent expected outcomes since they do not correspond to sources produced by a single run (Ma et al, 2013). In addition, ICASSO may be sensitive to outliers, especially when the number of runs is not large, and is computationally complex. Recently, a minimum-spanning tree (MST) approach has been developed for ICA

	Motor TC Correlation
IVA-CMGGD	0.9205
IVA-GL	0.9079
IVA-G	0.9025

Table 1 Central run motor component TC correlation with paradigm

analyses with the goal of finding a central run from a number of independent ones while constraining the solution space to actual runs (Du et al, 2014b, 2016, 2014a). Given a collection of sources from multiple runs, this is performed by selecting the run with sources that are a minimum distance to sources from all other runs as measured by $1-\rho_{\text{pearson}}$, where ρ_{pearson} is the Pearson correlation coefficient. Extending this approach to multivariate IVA sources can be performed by rotating the estimated sources in a consistent manner since the scale ambiguity of complex IVA leads to a misalignment in phase (Rodriguez et al, 2012). However, this approach becomes computationally expensive as the dimensionality of the dataset K, the number of sources N, and the number of samples V increase. In (Long et al, 2018), with the same objective of finding the most representative run from many, the authors take advantage of cross-ISI, defined to be the ISI where the true mixing matrix is replaced by the inverse of the estimated demixing matrix for the same dataset from another run in ICA. To illustrate, let $\mathbf{W}_{(r1)}^{[k]}$, $\mathbf{W}_{(r2)}^{[k]}$ be the estimated demixing matrices for the k^{th} dataset from runs $r1$ and $r2$ respectively. For our case, i.e., for application to IVA, we write the cross-ISI as:

$$\text{ISI}_{\text{cross}}(r_1, r_2) = \frac{1}{2}\left[\text{ISI}_{\text{joint}}(\mathcal{G}_{r1,r2}) + \text{ISI}_{\text{joint}}(\mathcal{G}_{r2,r1})\right], \tag{25}$$

where $\mathcal{G}_{r_i,r_j} = \left\{\mathbf{W}_{(r_i)}^{[k]}\mathbf{W}_{(r_j)}^{-[k]}\right\}_{k=1}^{K}$ to allow cross-ISI to be symmetric. Thus, cross-ISI, as defined in (25), is a symmetrical generalization of joint-ISI where the true mixing matrices for each dataset and each run are substituted with the corresponding inverses of demixing matrices of the same datasets, but from different runs. Finally, we compute for each run the average cross-ISI by:

$$\text{ISI}_{\text{cross}}^{r_i} = \frac{1}{R-1}\sum_{\substack{j=1\\ j\neq i}}^{R}\text{ISI}_{\text{cross}}(r_i, r_j), \tag{26}$$

where the quantity $\text{ISI}_{\text{cross}}^{r_i}$ measures the average deviation of run r_i from all other runs and represents the cost of selecting run r_i as the central run. Thus, the central run can be found through:

$$r_{\text{central}} = \arg\min_{r_i}\ \text{ISI}_{\text{cross}}^{r_i}, \tag{27}$$

where, in this paper, we perform 10 independent, randomly-initialized runs for each method. Figs. 5, 6, and 7 show three sample components from the central runs of each of the methods: the motor component, the sensorimotor component, and the default mode network (DMN) respectively. These components all show the desired compact, focal activated region estimation with little noise. Each of these images is averaged over all subjects, after accounting for phase ambiguity,

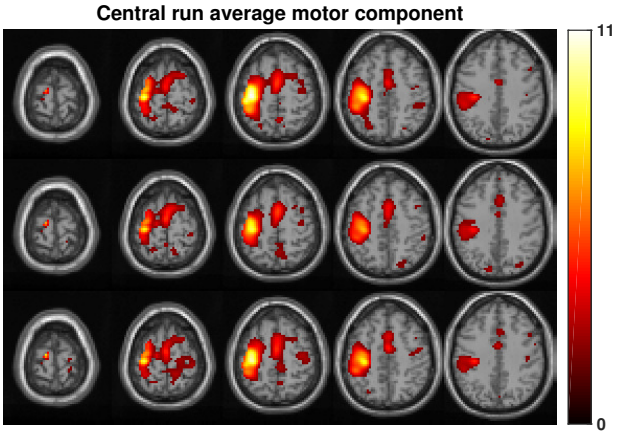

Fig. 5 Central run average motor component across all subjects: IVA-GL (top), IVA-G (middle), and IVA-CMGGD (bottom) thresholded at 1.96

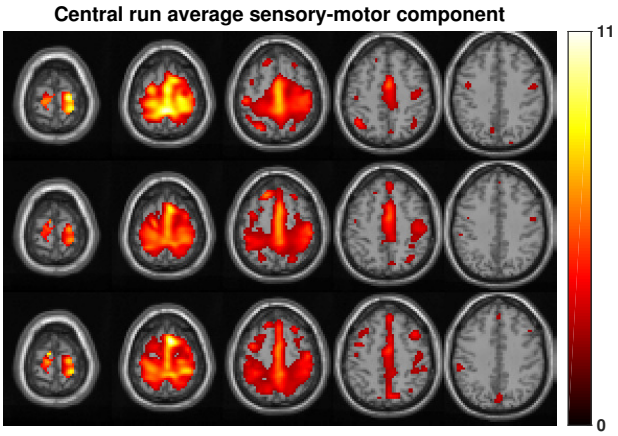

Fig. 6 Central run average sensorimotor component across all subjects: IVA-GL (top), IVA-G (middle), and IVA-CMGGD (bottom) thresholded at 1.96

and plotted using the Mahalanobis distance technique in (Rodriguez et al, 2012, 2011). To compare the central run decompositions quantitatively, we compute the correlation of the motor component TC, which is task related, to the paradigm TC after convolution with the SPM haemodynamic response. We present the central run motor component TC averaged over all subjects for each method in Fig. 8 while showing the paradigm TC for reference. Table 1 shows the correlation results. We note that IVA-CMGGD shows superior correlation due to its adaptive parametric density which better matches the true distributions. IVA-GL and IVA-G, with their simpler, more rigid models, cannot exploit all of the statistical information available. Finally, it is important to note that although the performance of the

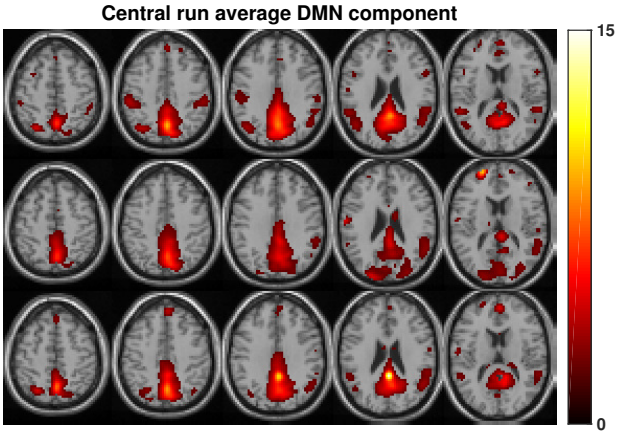

Fig. 7 Central run average DMN component across all subjects: IVA-GL (top), IVA-G (middle), and IVA-CMGGD (bottom) thresholded at 1.96

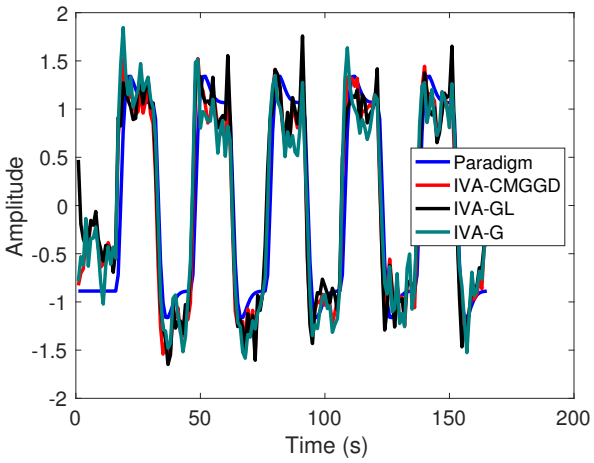

Fig. 8 Central run estimated motor component TC averaged across all subjects

central run is not necessarily the best performing run, it is the most reproducible, and thus best represents the average, or expected performance given a particular set of runs. The trade-off between run-to-run variability and expected bias is an important aspect to be considered prior to any analysis.

4 Conclusion

In this paper, we introduce IVA-CMGGD, a complex-valued, multivariate, and data-driven method for latent variable analysis based on statistical independence. By incorporating a simple, yet flexible, family of source distributions that incor-

porates the full statistical information inherent in the data, i.e., noncircularity, we show that it adapts to a wide variety of source models in a data-driven fashion. We show that this allows for better separation performance when compared to competing algorithms that impose less flexible assumptions on the analysis. We also demonstrate the applicability of IVA-CMGGD to real-world data through a 14-subject fMRI dataset with improved central run performance over competitive approaches.

Appendix A Derivation of Gradient

Since (15) is a real-valued function of complex variables, it suffices to compute the gradient with respect to $\mathbf{w}_n^{[k]*}$ using Wirtinger calculus. First, by applying the chain rule we can write:

$$\frac{\partial J(\mathbf{W})}{\partial \mathbf{w}_n^{[k]*}} = \frac{\partial J(\mathbf{W})}{\partial \mathbf{y}_n} \frac{\partial \mathbf{y}_n}{\partial w_n^{[k]*}} + \frac{\partial J(\mathbf{W})}{\partial \mathbf{y}_n^*} \frac{\partial \mathbf{y}_n^*}{\partial w_n^{[k]*}}. \tag{28}$$

Due to (12), the first term in (28) is equal to 0 leaving only the second term. Next, we subdivide the IVA-CMGGD cost function in (15) into two terms:

$$J(\mathbf{W}) = J_1(\mathbf{W}) + J_2(\mathbf{W}), \tag{29}$$

where

$$J_1(\mathbf{W}) = \frac{1}{2} \sum_{n=1}^{N} \mathbb{E} \left\{ \left(\frac{\eta_n}{2} \underline{\mathbf{y}}_n^H \underline{\mathbf{C}}_n^{-1} \underline{\mathbf{y}}_n \right)^{\beta_n} \right\}, \tag{30}$$

and

$$J_2(\mathbf{W}) = - \sum_{k=1}^{K} \left(\log \left| \mathbf{W}^{[k]} \right|^2 \right). \tag{31}$$

Differentiating $J_1(\mathbf{W})$ yields:

$$\begin{aligned} \frac{\partial J_1(\mathbf{W})}{\partial \mathbf{w}_n^{[k]*}} &= \frac{1}{2} \mathbb{E} \left\{ \beta_n \left(\frac{\eta_n}{2} \underline{\mathbf{y}}_n^H \underline{\mathbf{C}}_n^{-1} \underline{\mathbf{y}}_n \right)^{\beta_n - 1} \left(\frac{\eta_n}{2} \mathbf{e}_k^\top \underline{\mathbf{C}}_n^{-1} \underline{\mathbf{y}}_n \right) \mathbf{x}^{[k]*} \right\} \\ &= \mathbb{E} \left\{ \frac{\beta_n \eta_n^{\beta_n}}{2^{\beta_n + 1}} \frac{\left(\mathbf{e}_k^\top \underline{\mathbf{C}}_n^{-1} \underline{\mathbf{y}}_n \right) \mathbf{x}^{[k]*}}{\left(\underline{\mathbf{y}}_n^H \underline{\mathbf{C}}_n^{-1} \underline{\mathbf{y}}_n \right)^{(1-\beta_n)}} \right\}, \end{aligned} \tag{32}$$

where $\mathbf{e}_k$ is defined as in (16) and $\frac{\partial \mathbf{y}_n^*}{\partial \mathbf{w}_n^{[k]*}} = \mathbf{x}^{[k]*}$.

In order to differentiate $J_2(\mathbf{W})$, we utilize a decoupling procedure (Anderson et al, 2012a), originally established in (Li and Zhang, 2007). The purpose is to factorize each summand in (31) into the product of two terms: one dependent on $\mathbf{w}_n^{[k]}$ and the other independent of it. By defining $\tilde{\mathbf{W}}_n^{[k]}$ to be the $(N-1) \times N$ matrix containing rows of $\mathbf{W}^{[k]}$ other than the n^{th}, and by defining

$$\bar{\omega}_n^{[k]} = \sqrt{\left| \det \left(\tilde{\mathbf{W}}_n^{[k]} \tilde{\mathbf{W}}_n^{[k]H} \right) \right|}, \tag{33}$$

the decoupling procedure admits the following representation for $J_2(\mathbf{W})$:

$$J_2(\mathbf{W}) = -\sum_{k=1}^{K} \log\left(\left|\mathbf{w}_n^{[k]H}\mathbf{h}_n^{[k]*}\right|^2 \bar{\omega}_n^{[k]2}\right), \tag{34}$$

where $\mathbf{h}_n^{[k]}$ is a unit-length vector orthogonal to each of $\left\{\mathbf{w}_n^{[m]}\right\}_{m\neq k}$. Then, the gradient of $J_2(\mathbf{W})$ can be computed as:

$$\begin{aligned}\frac{\partial J_2(\mathbf{W})}{\partial \mathbf{w}_n^{[k]*}} &= \frac{\mathbf{h}_n^{[k]*}\mathbf{h}_n^{[k]\top}\mathbf{w}_n^{[k]}}{\mathbf{w}_n^{[k]H}\mathbf{h}_n^{[k]*}\mathbf{h}_n^{[k]\top}\mathbf{w}_n^{[k]}} \\ &= \frac{\mathbf{h}_n^{[k]*}}{\mathbf{h}_n^{[k]H}\mathbf{w}_n^{[k]*}}.\end{aligned} \tag{35}$$

Summing (32) and (35) yields the result in (16).

Acknowledgements The hardware used in the computational studies is part of the UMBC High Performance Computing Facility (HPCF). The facility is supported by the U.S. National Science Foundation through the MRI program (grant nos. CNS-0821258, CNS-1228778, and OAC-1726023) and the SCREMS program (grant no. DMS-0821311), with additional substantial support from the University of Maryland, Baltimore County (UMBC). See hpcf.umbc.edu for more information on HPCF and the projects using its resources.

Conflict of Interest: The authors declare that they have no conflict of interest.

References

Adalı T, Calhoun VD (2007) Complex ICA of brain imaging data. IEEE Signal Processing Magazine 24(5):136

Adalı T, Schreier P (2014) Optimization and estimation of complex-valued signals: Theory and applications in filtering and blind source separation. IEEE Signal Processing Magazine 31(5):112–128, DOI 10.1109/MSP.2013.2287951

Adalı T, Schreier P, Scharf L (2011a) Complex-valued signal processing: The proper way to deal with impropriety. IEEE Transactions on Signal Processing 59(11):5101–5125, DOI 10.1109/TSP.2011.2162954

Adalı T, Schreier PJ, Scharf LL (2011b) Complex-valued signal processing: The proper way to deal with impropriety. IEEE Transactions on Signal Processing 59(11):5101–5125

Adalı T, Anderson M, Fu GS (2014) Diversity in independent component and vector analyses: Identifiability, algorithms, and applications in medical imaging. IEEE Signal Processing Magazine 31(3):18–33

Anderson M, Li XL, Adalı T (2012a) Complex-valued independent vector analysis: Application to multivariate Gaussian model. Signal Processing 92(8):1821–1831

Anderson M, Li XL, Rodriguez P, Adalı T (2012b) An effective decoupling method for matrix optimization and its application to the ICA problem. In: Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, IEEE, pp 1885–1888

Arja SK, Feng Z, Chen Z, Caprihan A, Kiehl KA, Adalı T, Calhoun VD (2010) Changes in fMRI magnitude data and phase data observed in block-design and event-related tasks. NeuroImage 49(4):3149–3160

Brandwood DH (1983) A complex gradient operator and its application in adaptive array theory. Communications, Radar and Signal Processing, IEE Proceedings F 130(1):11–16, DOI 10.1049/ip-f-1.1983.0003

Bridwell DA, Rachakonda S, Silva RF, Pearlson GD, Calhoun VD (2018) Spatiospectral decomposition of multi-subject EEG: Evaluating blind source separation algorithms on real and realistic simulated data. Brain Topography 31(1):47–61

Calhoun V, Adalı T, Pearlson G, van Zijl P, Pekar J (2002) Independent component analysis of fMRI data in the complex domain. Magnetic Resonance in Medicine 48(1):180–192, DOI 10.1002/mrm.10202, URL http://dx.doi.org/10.1002/mrm.10202

Calhoun VD, Adalı T (2006) Unmixing fMRI with independent component analysis. IEEE Engineering in Medicine and Biology Magazine 25(2):79–90

Calhoun VD, Adalı T (2012) Multisubject independent component analysis of fMRI: a decade of intrinsic networks, default mode, and neurodiagnostic discovery. IEEE reviews in biomedical engineering 5:60–73

Cardoso JF (1998) Blind signal separation: Statistical principles. Proceedings of the IEEE 86(10):2009–2025

Dea JT, Anderson M, Allen E, Calhoun VD, Adalı T (2011) IVA for multi-subject fMRI analysis: A comparative study using a new simulation toolbox. In: Machine Learning for Signal Processing (MLSP), 2011 IEEE International Workshop on, IEEE, pp 1–6

Du W, Fu GS, Calhoun VD, Adalı T (2014a) Performance of complex-valued ICA algorithms for fMRI analysis: Importance of taking full diversity into account. In: Image Processing (ICIP), 2014 IEEE International Conference on, IEEE, pp 3612–3616

Du W, Ma S, Fu GS, Calhoun VD, Adalı T (2014b) A novel approach for assessing reliability of ICA for fMRI analysis. In: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, IEEE, pp 2084–2088

Du W, Levin-Schwartz Y, Fu GS, Ma S, Calhoun VD, Adalı T (2016) The role of diversity in complex ICA algorithms for fMRI analysis. Journal of Neuroscience Methods 264:129–135

Feng Z, Caprihan A, Blagoev KB, Calhoun VD (2009) Biophysical modeling of phase changes in BOLD fMRI. NeuroImage 47(2):540–548

Fu GS, Phlypo R, Anderson M, Adalı T (2015) Complex independent component analysis using three types of diversity: Non-Gaussianity, nonwhiteness, and noncircularity. IEEE Transactions on Signal Processing 63(3):794–805

Girolami M (1998) An alternative perspective on adaptive independent component analysis algorithms. Neural Computation 10(8):2103–2114

Haykin SO (2014) Adaptive Filter theory. Pearson

Himberg J, Hyvarinen A (2003) ICASSO: Software for investigating the reliability of ICA estimates by clustering and visualization. In: Neural Networks for Signal Processing, 2003. NNSP'03. 2003 IEEE 13th Workshop on, IEEE, pp 259–268

Horn RA, Johnson CR (1990) Matrix analysis. Cambridge university press

Hyvärinen A, Karhunen J, Oja E (2001) Independent component analysis, vol 46. John Wiley & Sons

Itakura K, Bando Y, Nakamura E, Itoyama K, Yoshii K, Kawahara T (2018) Bayesian multichannel audio source separation nased on integrated source and spatial models. IEEE/ACM Transactions on Audio, Speech, and Language Processing

Kim T, Eltoft T, Lee TW (2006) Independent vector analysis: An extension of ICA to multivariate components. In: International Conference on Independent Component Analysis and Signal Separation, Springer, pp 165–172

Kuang LD, Lin QH, Gong XF, Cong F, Calhoun VD (2017) Adaptive independent vector analysis for multi-subject complex-valued fMRI data. Journal of Neuroscience Methods 281(Supplement C):49 – 63, DOI https://doi.org/10.1016/j.jneumeth.2017.01.017, URL http://www.sciencedirect.com/science/article/pii/S0165027017300316

Lee I, Kim T, Lee TW (2006) Complex fastIVA: A robust maximum likelihood approach of MICA for convolutive BSS. In: ICA, Springer, vol 6, pp 625–632

Lee JH, Lee TW, Jolesz FA, Yoo SS (2008) Independent vector analysis (IVA): Multivariate approach for fMRI group study. Neuroimage 40(1):86–109

Lee TW, Girolami M, Sejnowski TJ (1999) Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. Neural Computation 11(2):417–441

Li H, Correa NM, Rodriguez PA, Calhoun VD, Adalı T (2011) Application of independent component analysis with adaptive density model to complex-valued fMRI data. IEEE Transactions on Biomedical Engineering 58(10):2794–2803

Li XL, Zhang XD (2007) Nonorthogonal joint diagonalization free of degenerate solution. IEEE Transactions on Signal Processing 55(5):1803–1814

Long Q, Jia C, Boukouvalas Z, Gabrielson B, Emge D, Adalı T (2018) Consistent run selection for independent component analysis: Application to fMRI analysis. In: ICASSP, accepted, IEEE

Lv H, Wang Z, Tong E, Williams L, Zaharchuk G, Zeineh M, Goldstein-Piekarski A, Ball T, Liao C, Wintermark M (2018) Resting-state functional MRI: Everything that nonexperts have always wanted to know. American Journal of Neuroradiology

Ma S, Phlypo R, Calhoun VD, Adalı T (2013) Capturing group variability using IVA: A simulation study and graph-theoretical analysis. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, IEEE, pp 3128–3132

Michael AM, Anderson M, Miller RL, Adalı T, Calhoun VD (2014) Preserving subject variability in group fMRI analysis: performance evaluation of GICA vs. IVA. Frontiers in Systems Neuroscience 8

Moreau E, Adalı T (2013) Blind Identification and Separation of Complex-valued Signals. Wiley Online Library

Moreau E, Macchi O (1994) A one stage self-adaptive algorithm for source separation. In: Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on, IEEE, vol 3, pp III–49

Mowakeaa R, Boukouvalas Z, Adalı T, Cavalcante C (2016) On the characterization, generation, and efficient estimation of the complex multivariate GGD. In: Sensor Array and Multichannel Signal Processing Workshop (SAM), 2016 IEEE, IEEE, pp 1–5

Na Y, Yu J, Chai B (2013) Independent vector analysis using subband and subspace nonlinearity. EURASIP Journal on Advances in Signal Processing

2013(1):74
Nocedal J, Wright SJ (2006) Numerical optimization. Springer
Ollila E, Tyler DE, Koivunen V, Poor HV (2012) Complex elliptically symmetric distributions: Survey, new results and applications. Signal Processing, IEEE Transactions on 60(11):5597–5625
Rachakonda S, Egolf E, Correa N, Calhoun V (2007) Group ICA of fMRI toolbox (GIFT) manual. Dostupné z http://www nitrc org/docman/view php/55/295/v1 3d_ GIFTManual pdf [cit 2011-11-5]
Rodriguez P, Calhoun V, Adalı T (2012) De-noising, phase ambiguity correction and visualization techniques for complex-valued ICA of group fMRI data. Pattern Recognition 45(6):2050–2063
Rodriguez PA, Correa NM, Eichele T, Calhoun VD, Adalı T (2011) Quality map thresholding for de-noising of complex-valued fMRI data and its application to ICA of fMRI. Journal of Signal Processing Systems 65(3):497–508
Rodriguez PA, Anderson M, Calhoun VD, et al (2015) General nonunitary constrained ICA and its application to complex-valued fmri data. IEEE Transactions on Biomedical Engineering 62(3):922–929
Rowe DB (2005) Modeling both the magnitude and phase of complex-valued fMRI data. Neuroimage 25(4):1310–1324
Schreier PJ, Scharf LL (2010) Statistical signal processing of complex-valued data: The theory of improper and noncircular signals. Cambridge University Press
Sorber L, Barel MV, Lathauwer LD (2012) Unconstrained optimization of real functions in complex variables. SIAM Journal on Optimization 22(3):879–898
Wax M, Kailath T (1985) Detection of signals by information theoretic criteria. IEEE Transactions on Acoustics, Speech, and Signal Processing 33(2):387–392
Wirtinger W (1927) Zur formalen theorie der funktionen von mehr komplexen veränderlichen. Mathematische Annalen 97(1):357–375
Xiong W, Li YO, Li H, Adalı T, Calhoun VD (2008) On ICA of complex-valued fMRI: Advantages and order selection. In: Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, IEEE, pp 529–532
Xiong W, Li YO, Correa N, Li XL, Calhoun VD, Adalı T (2012) Order selection of the linear mixing model for complex-valued fMRI data. Journal of Signal Processing Systems 67(2):117–128