Large-Scale Independent Vector Analysis (IVA-G) via Coresets

Ben Gabrielson, Hanlu Yang, Trung Vu, Vince Calhoun, and Tülay Adali

***Abstract*—Joint blind source separation (JBSS) involves the factorization of multiple matrices, i.e. "datasets", into "sources" that are statistically dependent across datasets and independent within datasets. Despite this usefulness for analyzing multiple datasets, JBSS methods suffer from considerable computational costs and are typically intractable for hundreds or thousands of datasets. To address this issue, we present a methodology for how a subset of the datasets can be used to perform efficient JBSS over the full set. We motivate two such methods: a numerical extension of independent vector analysis (IVA) with the multivariate Gaussian model (IVA-G), and a recently proposed analytic method resembling generalized joint diagonalization (GJD). We derive nonidentifiability conditions for both methods, and then demonstrate how one can significantly improve these methods' generalizability by an efficient representative subset selection method. This involves selecting a *coreset* (a weighted subset) that minimizes a measure of discrepancy between the statistics of the coreset and the full set. Using simulated and real functional magnetic resonance imaging (fMRI) data, we demonstrate significant scalability and source separation advantages of our "coreIVA-G" method vs. other JBSS methods.**

***Index Terms*—Joint Blind Source Separation, Independent Vector Analysis, Multiset Canonical Correlation Analysis.**

I. Introduction

The goal of joint blind source separation (JBSS) is to factorize several datasets arranged as matrices into components that maximize a measure of statistical dependence across the datasets while maximizing independence within each dataset. In BSS terminology, each individual component is called a "source". By this understanding, JBSS naturally generalizes blind source separation (BSS) of a single dataset by exploiting an additional statistical power: source dependence across the datasets. This not only estimates sources with greater interpretability, but also aligns sources across datasets and provides additional means to compare datasets via their uncovered source dependencies. JBSS has been frequently used for analyzing medical imaging datasets [1]–[4], but has seen applications in various other domains, such as remote sensing [5], frequency domain analysis [6], molecular property prediction [7], and various other applications.

The primary characteristic of JBSS is exploiting cross-dataset dependencies via constructing sets of dependent sources, typically called "source component vectors" (SCVs). Each SCV includes one source from each dataset, and JBSS methods typically operate by maximizing dependence within each SCV and independence across different SCVs.

Differences between JBSS methods largely hinge on the measure of statistical dependence being utilized. Mutual information is the primary measure used in independent vector analysis (IVA) [6], [8], a multi-dataset generalization of independent component analysis (ICA) for BSS. IVA algorithms parameterize each SCV by a multidimensional probability density function (PDF) to model statistical dependencies within and across SCVs. IVA methods offer some of the most powerful estimation capabilities of JBSS, yet IVA algorithms are burdened by higher computational expenses.

On the other hand, simpler methods such as multiset canonical correlation analysis (MCCA) [2], [5], [9], [10] and variants of generalized joint diagonalization (GJD) [11]–[13] exploit only source correlations as the measure of dependence, which leads to significantly more efficient algorithms with possibly less powerful estimation capabilities. An IVA algorithm most comparable with these methods is one assuming a multivariate Gaussian distribution (IVA-G) [14]. As dependence between Gaussian random vectors is described only by correlation, IVA-G similarly enjoys lower computational complexity and thus IVA-G has become a practical algorithm for performing IVA. Theoretically, algorithms exploiting only source correlation have been shown capable of estimating underlying SCVs, so long as the SCVs do not possess covariance matrices that are related to each other in certain aspects [14], [15].

Despite the powerful statistical capabilities of JBSS, many JBSS methods are computationally infeasible for very high-dimensional data, particularly with too many datasets (e.g., hundreds or more). This is especially a concern given the availability of larger numbers of datasets, and the benefits of including as many datasets as possible in the decomposition for capturing the underlying distribution and relationships in the data. Complexity of JBSS with respect to the number of datasets K can be shown as at least $\mathrm{O}(K^2)$ for even the simplest JBSS methods, however, a recent JBSS method proposed in [16] called "regIVA-G" allows for $\mathrm{O}(K)$ complexity. This method operates by performing JBSS first on a small subset of the K datasets to learn "regressor" sources, and sources from the remaining datasets are then estimated by maximizing/minimizing correlation with these regressor sources. The regIVA-G method is so named because it uses a multivariate Gaussian assumption for the SCV, and thus can be interpreted as a regression-based extension of IVA-G. It was further demonstrated that regIVA-G allows a specified dimension-parameterization for the SCVs: whereas IVA-G assumes a K-dimensional distribution for the K-dimensional SCVs, regIVA-G effectively parameterizes the SCV dimensions by the number of datasets in the subset. This allows regIVA-G the ability to provide lower-dimensional parameterizations to over-parameterized SCVs. Thus, regIVA-G was demonstrated as both feasible to large numbers of

This work was supported in part by grants NSF 2316420, NIH R01MH118695, NIH R01MH123610, and NIH R01AG073949. The computational hardware used is part of the UMBC High Performance Computing Facility (HPCF), supported by the U.S. NSF through the MRI and SCREMS programs (grants CNS-0821258, CNS-1228778, OAC-1726023, DMS-0821311), with additional substantial support from the University of Maryland, Baltimore County (UMBC).

Ben Gabrielson, Hanlu Yang, Trung Vu, and Tülay Adali are with the Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore County, Baltimore MD. (bengabr1@umbc.edu).

Vince Calhoun is with the Tri-Institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), Georgia State University, Georgia Institute of Technology, Emory University, Atlanta, Georgia.

datasets and flexible to the effective dimensionality of SCVs.

In this paper, we provide a comprehensive methodology for “regIVA-G”: scaling IVA-G on a subset to a much larger set of datasets. Our paper provides the following contributions:

- As an alternative to the analytic method proposed in [16], we propose a numerical method to regIVA-G based on maximum-likelihood IVA-G [14].
- For the analytic and numerical methods, we give theoretical understanding of their capabilities via deriving *nonidentifiability conditions*: statistical conditions for when the methods cannot identify sources in a new dataset.
- Whereas [16] used random subsets, we propose selecting a subset that minimizes a novel discrepancy-based cost function [17], [18] between the statistics of the subset and the statistics of the full set. We derive this discrepancy measure directly from the analytic method’s objective function, noting the discrepancy is applicable to most other JBSS methods. We then introduce an efficient subset selection method to minimize the discrepancy based on coresets (weighted subsets) [18]–[22], motivating the name “coreIVA-G” for performing IVA-G with coresets.

We compare performance of the methods with coreset vs. random subsets, alongside other efficient JBSS algorithms, on simulated and real functional magnetic resonance imaging (fMRI) datasets. Our results demonstrate that coreIVA-G can significantly outperform other comparable methods in both computational and source separation performance.

The paper is organized as follows. Section II formulates the JBSS problem. Section III introduces IVA and IVA-G. Section IV introduces the regIVA-G methodology for scaling IVA-G on a subset to a larger set of datasets, and introduces two methods for scaling to a new dataset. Section V derives nonidentifiability conditions for both methods. Section VI introduces a subset selection method by minimizing a discrepancy measure between the statistics of the subset and the full set. Section VII demonstrates performance with respect to simulated data and real fMRI data. Section VIII concludes with takeaways and discusses future areas of improvement.

II. JBSS Problem Formulation

We start with the JBSS problem formulation. We have K datasets, each modeled as linear mixtures of N sources. At some sample t, the generative model is:

$$\mathbf{x}^{[k]}(t) = \mathbf{A}^{[k]}\ \mathbf{s}^{[k]}(t)\ , \quad t = 1, \ldots, T, \quad k = 1, \ldots, K, \tag{1}$$

with $\mathbf{x}^{[k]} = [x_1^{[k]}, \ldots, x_N^{[k]}]^\top \in \mathbb{R}^N$ denoting the N observed signals within the kth dataset, $\mathbf{A}^{[k]} \in \mathbb{R}^{N\times N}$ denoting an unknown invertible “mixing” matrix, $\mathbf{s}^{[k]} = [s_1^{[k]}, \ldots, s_N^{[k]}]^\top \in \mathbb{R}^N$ denoting the kth dataset’s N latent source signals, and $(.)^\top$ denotes the transpose. JBSS methods generally do not make any model assumptions on the $\mathbf{A}^{[k]}$ (other than being full rank), only modeling the $\mathbf{s}^{[k]}$. Note we assume that for each dataset the number of mixtures is equal to the number of sources N. In practice, an overdetermined system of more mixtures than sources is reduced to N mixtures, typically using principal component analysis (PCA) on each dataset.

The goal of JBSS is to estimate the K datasets’ sources, via estimating K demixing matrices $\mathbf{W}^{[k]} \in \mathbb{R}^{N\times N}$ that demix the datasets into the estimated sources $\mathbf{y}^{[k]} = \mathbf{W}^{[k]}\ \mathbf{x}^{[k]}$, with $\mathbf{y}^{[k]} = [y_1^{[k]}, \ldots, y_N^{[k]}]^\top \in \mathbb{R}^N$. The nth row of demixing matrix $\mathbf{W}^{[k]}$ is given by $(\mathbf{w}_n^{[k]})^\top$, and is used to estimate the nth source within the kth dataset, via $y_n^{[k]} = (\mathbf{w}_n^{[k]})^\top\ \mathbf{x}^{[k]}$.

With T samples of data, the observed datasets are represented by matrices $\mathbf{X}^{[k]} = [\mathbf{x}_1^{[k]}, \ldots, \mathbf{x}_N^{[k]}]^\top \in \mathbb{R}^{N\times T}$, and the model (1) is given as $\mathbf{X}^{[k]} = \mathbf{A}^{[k]}\ \mathbf{S}^{[k]}$, with sources given by $\mathbf{S}^{[k]} = [\mathbf{s}_1^{[k]}, \ldots, \mathbf{s}_N^{[k]}]^\top \in \mathbb{R}^{N\times T}$, and estimated sources given by $\mathbf{Y}^{[k]} = \mathbf{W}^{[k]}\ \mathbf{X}^{[k]} = [\mathbf{y}_1^{[k]}, \ldots, \mathbf{y}_N^{[k]}]^\top \in \mathbb{R}^{N\times T}$.

To model dependencies across datasets, JBSS formulations assume that sources of the same index n are dependent across

K	number of total datasets	(dataset index $k = 1, \ldots, K$)
N	number of SCVs	(SCV index $n = 1, \ldots, N$)
T	number of samples	(sample index $t = 1, \ldots, T$)
$\mathbf{x}^{[k]}$ / $\mathbf{X}^{[k]}$	kth dataset	$(\in \mathbb{R}^N$ / $\mathbb{R}^{N\times T})$
$\mathbf{s}^{[k]}$ / $\mathbf{S}^{[k]}$	kth dataset’s true sources	$(\in \mathbb{R}^N$ / $\mathbb{R}^{N\times T})$
$\mathbf{y}^{[k]}$ / $\mathbf{Y}^{[k]}$	kth dataset’s estimated sources	$(\in \mathbb{R}^N$ / $\mathbb{R}^{N\times T})$
$s_n^{[k]}$ / $\mathbf{s}_n^{[k]}$	nth true source in kth dataset	$(\in \mathbb{R}$ / $\mathbb{R}^T)$
$y_n^{[k]}$ / $\mathbf{y}_n^{[k]}$	nth estimated source in kth dataset	$(\in \mathbb{R}$ / $\mathbb{R}^T)$
$\mathbf{A}^{[k]}$	kth dataset’s mixing matrix	$(\in \mathbb{R}^{N\times N})$
$\mathbf{W}^{[k]}$	kth dataset’s estimated demixing matrix	$(\in \mathbb{R}^{N\times N})$
$\mathbf{a}_n^{[k]}$	nth column of $\mathbf{A}^{[k]}$	$(\in \mathbb{R}^N)$
$(\mathbf{w}_n^{[k]})^\top$	nth demixing vector (row vector) in $\mathbf{W}^{[k]}$	$(\in \mathbb{R}^N)$
$\mathbf{s}_n$ / $\mathbf{S}_n$	nth true SCV over all K datasets	$(\in \mathbb{R}^K$ / $\mathbb{R}^{K\times T})$
$\mathbf{y}_n$ / $\mathbf{Y}_n$	nth estimated SCV over all K datasets	$(\in \mathbb{R}^K$ / $\mathbb{R}^{K\times T})$
$\mathbf{C}_{\mathbf{x}}^{[i,j]}$ / $\hat{\mathbf{C}}_{\mathbf{x}}^{[i,j]}$	(i, j) datasets’ cross-covariance matrix / sample cross-covariance matrix	$(\in \mathbb{R}^{N\times N})$
$\mathbf{C}_{\mathbf{y}_n}$ / $\hat{\mathbf{C}}_{\mathbf{y}_n}$	$\mathbf{y}_n$’s covariance matrix / sample covariance matrix	$(\in \mathbb{R}^{K\times K})$
K_b	number of datasets in the regIVA-G subset	
S_{K_b}	index set of the K_b subset	
$\tilde{\mathbf{s}}_n$ / $\tilde{\mathbf{S}}_n$	nth true SCV of the K_b subset	$(\in \mathbb{R}^{K_b}$ / $\mathbb{R}^{K_b\times T})$
$\tilde{\mathbf{y}}_n$ / $\tilde{\mathbf{Y}}_n$	nth estimated SCV of the K_b subset	$(\in \mathbb{R}^{K_b}$ / $\mathbb{R}^{K_b\times T})$
$\tilde{\mathbf{s}}_n^{[i]}$ / $\tilde{\mathbf{S}}_n^{[i]}$	$\tilde{\mathbf{s}}_n$ appended with nth true source in ith remaining dataset	$(\in \mathbb{R}^{K_b+1}$ / $\mathbb{R}^{(K_b+1)\times T})$
$\tilde{\mathbf{y}}_n^{[i]}$ / $\tilde{\mathbf{Y}}_n^{[i]}$	$\tilde{\mathbf{y}}_n$ appended with nth estimated source in ith remaining dataset	$(\in \mathbb{R}^{K_b+1}$ / $\mathbb{R}^{(K_b+1)\times T})$
$\mathbf{C}_{\tilde{\mathbf{s}}_n^{[i]}}$	covariance matrix of $\tilde{\mathbf{s}}_n^{[i]}$	$(\in \mathbb{R}^{(K_b+1)\times(K_b+1)})$
$\hat{\mathbf{C}}_{\tilde{\mathbf{y}}_n^{[i]}}$	sample covariance matrix of $\tilde{\mathbf{y}}_n^{[i]}$	$(\in \mathbb{R}^{(K_b+1)\times(K_b+1)})$
$(\mathbf{c}_n^{[i]})_m$	vector of cross-correlations of $s_n^{[i]}$ with $\tilde{\mathbf{s}}_m$	$(\in \mathbb{R}^{K_b})$
$(\hat{\mathbf{c}}_n^{[i]})_m$	vector of cross-correlations of $\mathbf{y}_n^{[i]}$ with $\tilde{\mathbf{Y}}_m$	$(\in \mathbb{R}^{K_b})$
$\hat{\mathbf{R}}_m^{[i]}$	$(\frac{1}{T-1})^2\ \mathbf{X}^{[i]}\ \tilde{\mathbf{Y}}_m^\top\ \tilde{\mathbf{Y}}_m\ \mathbf{X}^{[i]\top}$	$(\in \mathbb{R}^{N\times N})$
$\tilde{\mathbf{\Omega}}_n$	$\frac{1}{K_b}\ [\tilde{\mathbf{S}}_n^\top\ \tilde{\mathbf{S}}_n - \sum_{\substack{m=1 \\ m\neq n}}^N\ \tilde{\mathbf{S}}_m^\top\ \tilde{\mathbf{S}}_m]$	$(\in \mathbb{R}^{T\times T})$
$\mathbf{\Omega}_n$	$\frac{1}{K}\ [\mathbf{S}_n^\top\ \mathbf{S}_n - \sum_{\substack{m=1 \\ m\neq n}}^N\ \mathbf{S}_m^\top\ \mathbf{S}_m]$	$(\in \mathbb{R}^{T\times T})$
$\mathbf{X}^{[k]\top}\ \mathbf{X}^{[k]}$	kth dataset’s projection embedding	$(\in \mathbb{R}^{T\times T})$
$\mathbf{\Psi}$	mean embedding of all K datasets	$(\in \mathbb{R}^{T\times T})$

Table 1. Notations used in this paper. Vectors are given as column vectors, e.g. $\mathbf{w}_n^{[k]}$ is a column vector, and $(\mathbf{w}_n^{[k]})^\top$ a row vector, with $^\top$ denoting the transpose. Datasets and sources (e.g. $\mathbf{x}^{[k]}$, $\mathbf{s}^{[k]}$, and $\mathbf{y}^{[k]}$) are represented as either a random vector, or by T observed samples of a random vector (e.g. $\mathbf{x}^{[k]} \in \mathbb{R}^N$ / $\mathbf{X}^{[k]} \in \mathbb{R}^{N\times T}$).

the K datasets, thus forming N sets of K sources. In IVA terminology, each of these sets is referred to as a “source component vector” (SCV). The nth SCV is denoted by $\mathbf{s}_n = [s_n^{[1]}, \ldots, s_n^{[K]}]^\top \in \mathbb{R}^K$, and is estimated by $\mathbf{y}_n = [y_n^{[1]}, \ldots, y_n^{[K]}]^\top \in \mathbb{R}^K$. Over T samples, the nth SCV is represented by the matrix $\mathbf{S}_n = [\mathbf{s}_n^{[1]}, \ldots, \mathbf{s}_n^{[K]}]^\top \in \mathbb{R}^{K\times T}$, estimated by $\mathbf{Y}_n = [\mathbf{y}_n^{[1]}, \ldots, \mathbf{y}_n^{[K]}]^\top \in \mathbb{R}^{K\times T}$. Typically each SCV is modeled as independent from all other SCVs, thus any two sources across the datasets are modeled as dependent only if they correspond to the same index n (nth SCV).

JBSS algorithms can only identify demixing matrix vectors $(\mathbf{w}_n^{[k]})^\top$ (and thus the estimated sources $y_n^{[k]}$) up to scaling and permutation ambiguity within each dataset. JBSS additionally orders sources to align with the order of SCVs, such that the nth source within a dataset corresponds to the nth SCV.

Additionally, JBSS implementations typically involve standardizing and prewhitening each dataset prior to estimation, as this considerably simplifies the calculations involved in solving these problems [23]. It is notable that when datasets are standardized and prewhitened, then provided that the SCVs are uncorrelated (and thus sources are uncorrelated within datasets), it follows that the residual mixing matrices $\mathbf{A}^{[k]}$ become asymptotically orthogonal for the observed datasets $\mathbf{X}^{[k]}$ as $T \to \infty$. We will assume for the remainder of the paper that datasets are standardized and prewhitened prior to JBSS, thus each mixture and source is zero mean unit variance. However, as in practice we deal with finite T, we do not generally assume that the $\mathbf{A}^{[k]}$ are orthogonal.

III. IVA and IVA-G: Background

This section explains the JBSS methodology of IVA, and explains that a multivariate Gaussian parameterization of the SCVs leads to the IVA-G method. We explain that despite IVA-G being perhaps the most efficient IVA method, IVA-G is computationally limited, thus motivating the regIVA-G methodology outlined in the following section.

A. Independent Vector Analysis (IVA)

The fundamental assumption of IVA is that the N SCVs are independent, and thus JBSS can be performed by minimizing a measure of dependence among the SCVs. A useful and general measure of dependence is the mutual information among the N SCVs. Given estimated SCVs $\mathbf{y}_n$ (determined by demixing matrices $\mathbf{W}^{[k]}$), this leads to the general IVA cost function:

$$\mathcal{J}_{\text{IVA}}\left(\mathcal{W}\right) \triangleq \sum_{n=1}^{N}\mathcal{H}\left\{\mathbf{y}_n\right\} - \sum_{k=1}^{K}\log\left|\det\left(\mathbf{W}^{[k]}\right)\right| \tag{2}$$

where $\mathcal{W}$ is the collection of $\mathbf{W}^{[k]}$ for $k = 1, \ldots, K$, and $\mathcal{H}\{\mathbf{y}_n\}$ is defined as the entropy of SCV estimate $\mathbf{y}_n$, which is specifically defined by its PDF [8]. The term $\log|\det\left(\mathbf{W}^{[k]}\right)|$ acts as a penalty effectively ensuring that sources are close to being uncorrelated within each dataset.

B. IVA with multivariate Gaussian Distribution (IVA-G)

IVA-G [14] is a variant of the general IVA cost (2) where each SCV’s PDF is modeled as multivariate Gaussian with independent and identically distributed (i.i.d.) samples t. The IVA-G cost is thus given by :

$$\mathcal{J}_{\text{IVA-G}}(\mathcal{W}) = \frac{1}{2}\sum_{n=1}^{N}\log\left|\det\left(\hat{\mathbf{C}}_{\mathbf{y}_n}\right)\right| - \sum_{k=1}^{K}\log\left|\det\left(\mathbf{W}^{[k]}\right)\right| + c \tag{3}$$

where we define $\hat{\mathbf{C}}_{\mathbf{y}_n} = \frac{1}{T-1}\,\mathbf{Y}_n\,\mathbf{Y}_n^\top \in \mathbb{R}^{K\times K}$ as the sample covariance matrix of SCV $\mathbf{y}_n$, and $c = \frac{1}{2}NK\,\log(2\pi e)$ is a constant. Minimizing (3) can also be explained as minimizing correlation amongst the N SCVs while also maximizing the correlation within each SCV [14], [16].

Despite its efficiency among IVA methods, IVA-G nonetheless suffers from computational complexity. We consider the minimum computations required for IVA-G numerical methods provided in [14]: computing the gradient. Here, we ignore the one-time initial costs of estimating the $\hat{\mathbf{C}}_{\mathbf{x}}^{[i,j]} = \frac{1}{T-1}\,\mathbf{X}^{[i]}\,\mathbf{X}^{[j]\top} \in \mathbb{R}^{N\times N}$, the dataset cross-covariances, for $1 \leq i, j \leq K$. Asides from estimating the $\hat{\mathbf{C}}_{\mathbf{x}}^{[i,j]}$, each iteration requires updating all NK demixing vectors $\mathbf{w}_n^{[k]}$, where each $\mathbf{w}_n^{[k]}$ update involves an update of $\mathbf{W}^{[k]}$ of $\mathrm{O}(N^3)$ complexity, an update of $\hat{\mathbf{C}}_{\mathbf{y}_n}$ of $\mathrm{O}(N^2K)$ complexity, and an update of $\hat{\mathbf{C}}_{\mathbf{y}_n}^{-1}$ of $\mathrm{O}(K^3)$ complexity. If IVA-G requires q iterations to converge, this leads a total complexity of $\mathrm{O}(q(N^4K + N^3K^2 + NK^4))$. This leads IVA-G to becoming computationally infeasible for large K, motivating the need for the low complexity alternative methods described in the following sections.

IV. regIVA-G: Subset-based methods for large-scale IVA-G

We now provide an overview of the main methodology of the paper: using a subset of the datasets to efficiently perform IVA-G on all datasets. The methodology was first introduced in the preliminary work of [16] and was called “regIVA-G” due to the solution being a multilinear-regression of the subset’s estimated SCVs. For simplicity, we refer to this methodology as “regIVA-G” when using a general choice of subset (e.g. a random subset), and later refer to the methodology as “coreIVA-G” when using a *coreset* (weighted subset) selected to best represent the statistics of the K datasets.

The regIVA-G methodology is illustrated in Fig. 1. The three steps of regIVA-G are summarized as follows:

1) partitioning step: divide the K datasets into two groups ($K = K_b + K_a$), the K_b *regressors* and the K_a *regressed*:
   - K_b datasets that form the subset estimated by IVA-G (whose sources will form *regressors*)
   - K_a remaining datasets (“new” datasets) that will be *regressed* onto the regressor sources of the subset
2) subset estimation step: perform IVA-G on the K_b subset, estimating the subset’s $\mathbf{W}^{[k]}$ and corresponding N SCVs $\tilde{\mathbf{Y}}_n$, which we call *regressor SCVs*.
3) regression step: use the regressor SCVs to separately estimate N sources in each of the K_a remaining datasets. Each source in remaining dataset $\mathbf{X}^{[i]}$ is estimated such that it is maximally correlated to one regressor SCV and maximally uncorrelated to the $N - 1$ other SCVs.

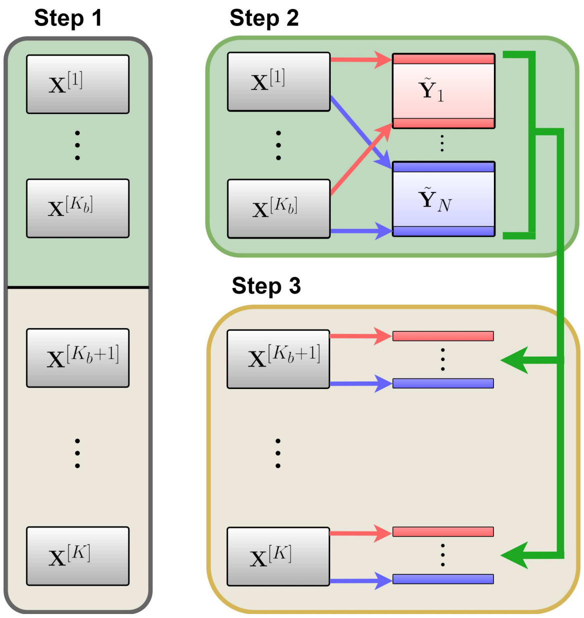

Fig. 1. illustration of the regIVA-G methodology.

Unlike IVA-G or other JBSS methods, this methodology estimates sources in each remaining dataset independently of other remaining datasets. As a result, this methodology exhibits asymptotically linear complexity with respect to K, provided that K_b is fixed and $K_a \to \infty$. With large K, regIVA-G's complexity is dominated by the O(K) regression step.

In the following sections, we overview two methods for performing the regression step: one being the explicit numerical minimization of the IVA-G cost, and the other being the GJD-type analytic solution proposed in [16].

A. *regIVA-G-N (Numerical method)*

The regIVA-G method for numerically minimizing the IVA-G cost, which we refer to as regIVA-G-N, is simply described: for each remaining $\mathbf{X}^{[i]}$, estimate $\mathbf{W}^{[i]}$ by performing IVA-G on the K_b subset's datasets appended with $\mathbf{X}^{[i]}$ (thus performing on K_b+1 total datasets), while updating only that ith dataset's $\mathbf{W}^{[i]}$ (fixing constant the subset's $\mathbf{W}^{[k]}$ to those $\mathbf{W}^{[k]}$ estimated in the subset estimation step).

We first provide notations for the regIVA-G-N subproblem over K_b+1 datasets as to differentiate from IVA-G over all K datasets, namely we introduce new notations concerning the ith subproblem's SCVs. We first denote $\tilde{\mathbf{s}}_n \in \mathbb{R}^{K_b}$ as the nth SCV of the subset, and $\tilde{\mathbf{s}}_n^{[i]} = [\tilde{\mathbf{s}}_n^\top, s_n^{[i]}]^\top \in \mathbb{R}^{K_b+1}$ as the nth SCV of the subset appended with $s_n^{[i]}$, the nth source in the ith dataset. These quantities respectively correspond to the estimated SCVs: $\tilde{\mathbf{y}}_n \in \mathbb{R}^{K_b}$ and $\tilde{\mathbf{y}}_n^{[i]} = [\tilde{\mathbf{y}}_n^\top, y_n^{[i]}]^\top \in \mathbb{R}^{K_b+1}$. Over T samples, the SCVs are represented by the matrices $\tilde{\mathbf{S}}_n \in \mathbb{R}^{K_b \times T}$, $\tilde{\mathbf{S}}_n^{[i]} = [\tilde{\mathbf{S}}_n^\top, \mathbf{s}_n^{[i]}]^\top \in \mathbb{R}^{(K_b+1)\times T}$, $\tilde{\mathbf{Y}}_n \in \mathbb{R}^{K_b \times T}$ and $\tilde{\mathbf{Y}}_n^{[i]} = [\tilde{\mathbf{Y}}_n^\top, \mathbf{y}_n^{[i]}]^\top \in \mathbb{R}^{(K_b+1)\times T}$.

Using this notation, we can write the cost function per each ith remaining dataset as a modified version of (3). If we denote the nth estimated SCV's sample covariance by $\hat{\mathbf{C}}_{\tilde{\mathbf{y}}_n^{[i]}} = \frac{1}{T-1} (\tilde{\mathbf{Y}}_n^{[i]}) (\tilde{\mathbf{Y}}_n^{[i]})^\top \in \mathbb{R}^{(K_b+1)\times(K_b+1)}$, and ignore the constant term, the regIVA-G-N cost function is given by:

$$\mathcal{J}_{\text{regIVA-G-N}}(\mathbf{W}^{[i]}) = \frac{1}{2}\sum_{n=1}^{N} \log\left|\det\left(\hat{\mathbf{C}}_{\tilde{\mathbf{y}}_n^{[i]}}\right)\right| - \log\left|\det\left(\mathbf{W}^{[i]}\right)\right| \tag{4}$$

All methods for minimizing (3) are also applicable to (4); the only difference with (4) is that all $\mathbf{W}^{[k]}$ are fixed for $k \neq i$.

For each of the K_a remaining datasets, regIVA-G-N's regression step involves IVA-G to minimize (4) over K_b+1 datasets. Given IVA-G's complexity described in Section III.B, it follows that the complexity of regIVA-G-N's regression step is O($K_a\ q(N^4 + N^3(K_b+1) + N(K_b+1)^3)$).

B. *regIVA-G-A (Analytic method)*

The analytic method proposed in [16], which we refer to as regIVA-G-A, is a highly efficient alternative to the previously described numerical method with cost described in (4). In contrast to (3) and (4), regIVA-G-A measures the degree of source dependence by the squared correlation between sources, analogous to generalized joint diagonalization (GJD) costs [11]–[13]. Furthermore, unlike regIVA-G-N in (4) where each dataset's sources are estimated jointly (via $\mathbf{W}^{[i]}$), regIVA-G-A involves separate estimation of each source (via each $\mathbf{w}_n^{[i]}$).

We now describe regIVA-G-A's objective function. Correlation of the nth source in $\mathbf{X}^{[i]}$ with each of the mth SCV's sources is given by $(\hat{\mathbf{c}}_n^{[i]})_m$ = $\frac{1}{T-1}\ \tilde{\mathbf{Y}}_m\ \mathbf{X}^{[i]\top}\ \mathbf{w}_n^{[i]} \in \mathbb{R}^{K_b}$. The degree of correlation with that SCV is measured by the sum of squared correlations with sources in the mth SCV: $||(\hat{\mathbf{c}}_n^{[i]})_m||_2^2$ = $(\hat{\mathbf{c}}_n^{[i]})_m^\top\ (\hat{\mathbf{c}}_n^{[i]})_m = \mathbf{w}_n^{[i]\top}\ \hat{\mathbf{R}}_m^{[i]}\ \mathbf{w}_n^{[i]}$, where we define $\hat{\mathbf{R}}_m^{[i]} \triangleq (\frac{1}{T-1})^2\ \mathbf{X}^{[i]}\ \tilde{\mathbf{Y}}_m^\top\ \tilde{\mathbf{Y}}_m\ \mathbf{X}^{[i]\top} \in \mathbb{R}^{N\times N}$. Using these $\hat{\mathbf{R}}_m^{[i]}$, regIVA-G-A's objective function measures the corresponding nth source's degree of correlation with its nth SCV, weighted against the correlation with the $N-1$ other SCVs:

$$\mathcal{J}_{\text{regIVA-G-A}}\ (\mathbf{w}_n^{[i]}) =\ \mathbf{w}_n^{[i]\top}\Big[\hat{\mathbf{R}}_n^{[i]} - \sum_{\substack{m=1\\ m\neq n}}^{N} \hat{\mathbf{R}}_m^{[i]}\Big]\mathbf{w}_n^{[i]} \tag{5}$$

Subject to $||\mathbf{w}_n^{[k]}||_2 = 1$, the $\mathbf{w}_n^{[i]}$ that maximizes (5) is estimated by the principal eigenvector of $\big[\hat{\mathbf{R}}_n^{[i]} - \sum_{\substack{m=1\\ m\neq n}}^{N} \hat{\mathbf{R}}_m^{[i]}\big]$.

For each remaining dataset, regIVA-G-A involves calculating the N SCVs' $\hat{\mathbf{R}}_n^{[i]}$ (of complexity O(N^3K_b)), the N $\big[\hat{\mathbf{R}}_n^{[i]} - \sum_{\substack{m=1\\ m\neq n}}^{N} \hat{\mathbf{R}}_m^{[i]}\big]$ (of O(N^3)), and the principal eigenvector of each $\big[\hat{\mathbf{R}}_n^{[i]} - \sum_{\substack{m=1\\ m\neq n}}^{N} \hat{\mathbf{R}}_m^{[i]}\big]$ per $\mathbf{w}_n^{[i]}$ (of O(N^3)). Thus, complexity of regIVA-G-A's regression step is O($K_a(K_b+2)N^3$).

It is notable that because the $\mathbf{w}_n^{[i]}$ are separately estimated in each ith dataset, this analytic method does not explicitly maximize source uncorrelatedness within each dataset (unlike the numerical method). Instead, uncorrelatedness within datasets is indirectly achieved by maximizing uncorrelatedness with the subset's SCVs. This difference between the regIVA-G methods leads to differences in estimation capabilities, highlighted in Section VII when simulating correlated SCVs.

In the next section, we derive conditions on the data's generative model for which regIVA-G is unable to uniquely identify the true sources (via demixing vectors $\mathbf{w}_n^{[k]}$) subject to scale and permutation ambiguity, which we refer to as the nonidentifiability conditions of regIVA-G.

V. REGIVA-G NONIDENTIFIABILITY CONDITIONS

This section is dedicated to deriving the *nonidentifiability conditions* of regIVA-G: statistical conditions on the data's generative model for which regIVA-G is unable to identify sources $\mathbf{S}^{[i]}$ in the regression step of a new dataset $\mathbf{X}^{[i]}$. When these conditions are *not* satisfied, precise inference of $\mathbf{S}^{[i]}$ is possible such that one can achieve $\mathbf{Y}^{[i]} = \mathbf{S}^{[i]}$, via achieving $\mathbf{W}^{[i]} = (\mathbf{A}^{[i]})^{-1}$, subject to scale and permutation ambiguities.

Denoting $\mathbf{a}_n^{[i]}$ as the nth column of $\mathbf{A}^{[i]}$ and $(\mathbf{w}_n^{[i]})^\top$ as the nth row of $\mathbf{W}^{[i]}$, identifiability occurs for the nth source if $\mathbf{w}_n^{[i]} = \mathbf{a}_n^{[i]}$ subject to scale and permutation ambiguity.

These proofs use notations defined earlier in the paper, and proceed under the following assumptions:

- $T \to \infty$, thus the data's true statistics are known (e.g., $\mathbf{C}_{\mathbf{x}}^{[i,j]} = \mathrm{E}\left\{\mathbf{x}^{[i]}\mathbf{x}^{[j]\top}\right\} \in \mathbb{R}^{N\times N}$ for $1 \leq i,j \leq K$)
- uncorrelated SCVs ($\mathrm{E}\left\{\mathbf{s}_m\ \mathbf{s}_n^\top\right\} = \mathbf{0}$ for $m \neq n$)
- prewhitened datasets (thus $\mathbf{A}^{[k]}$ are orthogonal matrices)
- the subset's sources have been estimated exactly ($\tilde{\mathbf{Y}}_n = \tilde{\mathbf{S}}_n \in \mathbb{R}^{K_b\times T}$), thus our only concern is identifying the remaining datasets' sources (via their $\mathbf{W}^{[i]}$)

Nonidentifiability of regIVA-G depends on the SCV covariances, which we require notation for. From Section IV.A, we remind the nth regIVA-G SCV is defined $\tilde{\mathbf{s}}_n^{[i]} = [\tilde{\mathbf{s}}_n^\top, s_n^{[i]}]^\top \in \mathbb{R}^{K_b+1}$, which is the subset's nth SCV $\tilde{\mathbf{s}}_n \in \mathbb{R}^{K_b}$ appended with the new dataset's nth source $s_n^{[i]}$. We also define:

- $\mathbf{C}_{\tilde{\mathbf{s}}_n^{[i]}} = \mathrm{E}\{\tilde{\mathbf{s}}_n^{[i]}\tilde{\mathbf{s}}_n^{[i]\top}\} \in \mathbb{R}^{(K_b+1)\times(K_b+1)}$ as the covariance matrix (also correlation matrix) of $\tilde{\mathbf{s}}_n^{[i]}$
- $(\mathbf{c}_n^{[i]})_m = \mathrm{E}\{s_n^{[i]}\tilde{\mathbf{s}}_m^\top\}^\top \in \mathbb{R}^{K_b}$ as the vector containing correlations of the new dataset's nth source $s_n^{[i]}$ with all sources in the subset's mth SCV $\tilde{\mathbf{s}}_m$.

For convenience, we denote $i = K_b{+}1$ within the regression step, such that $[(\mathbf{c}_n^{[i]})_n^\top, 1]$ is the last row/column of $\mathbf{C}_{\tilde{\mathbf{s}}_n^{[i]}}$.

The following theorem states the nonidentifiability conditions shared by regIVA-G-N and regIVA-G-A.

Theorem 1 (regIVA-G nonidentifiability conditions): We follow all assumptions listed at the beginning of Section V. Considering the nth source in the ith remaining dataset $s_n^{[i]}$, corresponding to demixing vector $\mathbf{w}_n^{[i]}$, it follows that $s_n^{[i]}$ is *nonidentifiable* ($\mathbf{w}_n^{[i]} \neq \mathbf{a}_n^{[i]}$ subject to scale and permutation ambiguity) if and only if for any $1 \leq m \neq n \leq N$, both $(\mathbf{c}_n^{[i]})_n = \mathbf{0} \in \mathbb{R}^{K_b}$ and $(\mathbf{c}_m^{[i]})_m = \mathbf{0} \in \mathbb{R}^{K_b}$.

Simply stated: *For $s_n^{[i]}$ to be nonidentifiable, $s_n^{[i]}$ and another source $s_m^{[i]}$ must be uncorrelated to their corresponding SCVs.*

We first derive these conditions for regIVA-G-N's (4), and later derive these conditions for regIVA-G-A's (5).

A. regIVA-G-N nonidentifiability

Proof: We outline the two main steps in this proof:

1) show regIVA-G-N's regression step is actually a particular variation of ICA, and thus regIVA-G-N can be more easily studied via the ICA nonidentifiability conditions.
2) show that regIVA-G-N's Fisher Information Matrix (FIM) is singular if and only if $(\mathbf{c}_m^{[i]})_m = \mathbf{0}$ and $(\mathbf{c}_n^{[i]})_n = \mathbf{0}$.

Before connecting regIVA-G-N to ICA, we first introduce some preliminaries regarding the IVA cost function.

IVA requires specification of $p_n(\mathbf{y}_n)$, the (chosen) differentiable PDF of the nth SCV (for $n = 1,\dots,N$). Associated with this PDF is the *score function* for the nth SCV [8], [14]:

$$[\boldsymbol{\Phi}(\mathbf{S}_n)]_{kt} = \frac{\partial\ \log\ p_n(\mathbf{S}_n)}{\partial\ s_n^{[k]}(t)}; \qquad \begin{array}{l} k = 1,2,\dots,\ K \\ t = 1,2,\dots,\ T \end{array}$$

Regarding the regIVA-G-N subproblem over $K_b{+}1$ datasets, we define the score function for the ith subproblem's nth true SCV by $\boldsymbol{\phi}(\tilde{\mathbf{s}}_n^{[i]}) = [\boldsymbol{\phi}(\tilde{\mathbf{s}}_n)^\top, \phi(\tilde{s}_n^{[i]})]^\top \in \mathbb{R}^{K_b+1}$ and nth estimated SCV by $\boldsymbol{\phi}(\tilde{\mathbf{y}}_n^{[i]}) = [\boldsymbol{\phi}(\tilde{\mathbf{y}}_n)^\top, \phi(\tilde{y}_n^{[i]})]^\top \in \mathbb{R}^{K_b+1}$.

Now we connect regIVA-G-N to ICA. If we assume the statistics are known ($T \to \infty$), and we multiply the general IVA mutual information cost function in (2) by T, the negative of (2) becomes equivalent to the log-likelihood [8], [14]. Within the general IVA log-likelihood gradient, by fixing all $\mathbf{W}^{[k]}$ to constant quantities for $k \neq i$ (as done by regIVA-G-N), we obtain regIVA-G-N's log-likelihood gradient:

$$\frac{\partial\mathcal{J}_{\text{IVA}}(\mathbf{W}^{[i]})}{\partial\mathbf{W}^{[i]}} = -\boldsymbol{\Phi}(\tilde{\mathbf{Y}}^{[i]})\mathbf{X}^{[i]\top} + T\left(\mathbf{W}^{[i]-1}\right)^\top \tag{6}$$

where $\boldsymbol{\phi}(\tilde{\mathbf{y}}^{[i]}) = [\phi(\tilde{y}_1^{[i]}),\dots,\phi(\tilde{y}_N^{[i]})] \in \mathbb{R}^N$ are the ith dataset's score function components, observed over T samples by $\boldsymbol{\Phi}(\tilde{\mathbf{Y}}^{[i]}) = [\boldsymbol{\phi}(\tilde{\boldsymbol{y}}_1^{[i]}),\dots,\boldsymbol{\phi}(\tilde{\boldsymbol{y}}_N^{[i]})] \in \mathbb{R}^{N\times T}$.

We now note that this gradient (6) takes the exact same form as the gradient of the log-likelihood for ICA [24]. Because of this equivalency, regIVA-G nonidentifiability can be more easily studied in terms of ICA nonidentifiability.

We refer to [1], [24] for the ICA nonidentifiability conditions. We focus on the log-likelihood's Fisher Information Matrix (FIM), $\mathbf{F}(\mathbf{W}^{[i]}) \in \mathbb{R}^{N^2\times N^2}$, as nonidentifiability conditions are those conditions that make the FIM singular. Evaluated at the optimum $\mathbf{W}^{[i]}\mathbf{A}^{[i]} = \mathbf{I}$ (subject to scale / permutation ambiguity), the FIM is block diagonal, with:

- N positive scalars (for each source);
- $\frac{N(N-1)}{2}$ matrices $\mathbf{F}_{m,n} \in \mathbb{R}^{2\times 2}$ (for each pair of sources).

As the scalars are positive, invertibility of the FIM depends only on invertibility of the $\mathbf{F}_{m,n}$ matrices [1], [24], given by:

$$\mathbf{F}_{m,n} = \begin{bmatrix} \mathcal{K}_{m,n} & 1 \\ 1 & \mathcal{K}_{n,m} \end{bmatrix},\ 1 \leq m,n \leq N,$$

and provided IVA-G assumes unit variance sources with i.i.d. samples, then $\mathcal{K}_{n,m} = \mathrm{E}\left\{\phi(s_n^{[i]})^2\right\}\mathrm{E}\left\{s_m^{[i]2}\right\} = \mathrm{E}\left\{\phi(s_n^{[i]})^2\right\} \in \mathbb{R}$ (not a function of m).

Then as IVA-G assumes a multivariate Gaussian PDF [14], the nth SCV's score function is $\boldsymbol{\phi}(\tilde{\mathbf{s}}_n^{[i]}) = \mathbf{C}_{\tilde{\mathbf{s}}_n^{[i]}}^{-1}\ \tilde{\mathbf{s}}_n^{[i]} \in \mathbb{R}^{K_b+1}$, and $\mathrm{E}\left\{\boldsymbol{\phi}(\tilde{\mathbf{s}}_n^{[i]})\boldsymbol{\phi}(\tilde{\mathbf{s}}_n^{[i]})^\top\right\} = \mathbf{C}_{\tilde{\mathbf{s}}_n^{[i]}}^{-1}\ \mathbf{C}_{\tilde{\mathbf{s}}_n^{[i]}}\ \mathbf{C}_{\tilde{\mathbf{s}}_n^{[i]}}^{-1} = \mathbf{C}_{\tilde{\mathbf{s}}_n^{[i]}}^{-1}$. Thus, the $\mathbf{F}_{m,n}$ are defined by $\mathcal{K}_{n,m} = (\mathbf{C}_{\tilde{\mathbf{s}}_n^{[i]}}^{-1})_{(i,i)}$, which is the (i,i)th diagonal entry of the inverse correlation matrix $\mathbf{C}_{\tilde{\mathbf{s}}_n^{[i]}}^{-1}$.

For the $\mathbf{F}_{m,n}$ to be invertible, it follows $\mathcal{K}_{n,m} \neq \mathcal{K}_{m,n}^{-1}$, or i.e., $(\mathbf{C}_{\tilde{\mathbf{s}}_n^{[i]}}^{-1})_{(i,i)} \neq 1/(\mathbf{C}_{\tilde{\mathbf{s}}_m^{[i]}}^{-1})_{(i,i)}$. However, since $\mathbf{C}_{\tilde{\mathbf{s}}_n^{[i]}}^{-1}$ is the inverse of a correlation matrix, its diagonal entries must obey $(\mathbf{C}_{\tilde{\mathbf{s}}_n^{[i]}}^{-1})_{(i,i)} \geq 1$ [25], and $(\mathbf{C}_{\tilde{\mathbf{s}}_n^{[i]}}^{-1})_{(i,i)} = 1$ is achieved only when $(\mathbf{c}_n^{[i]})_n = \mathbf{0}$ (i.e., $s_n^{[i]}$ is uncorrelated to its own SCV) [25].

Thus, having $(\mathbf{C}_{\tilde{\mathbf{s}}_n^{[i]}}^{-1})_{(i,i)} = 1/(\mathbf{C}_{\tilde{\mathbf{s}}_m^{[i]}}^{-1})_{(i,i)}$ requires that $(\mathbf{C}_{\tilde{\mathbf{s}}_n^{[i]}}^{-1})_{(i,i)} = 1/(\mathbf{C}_{\tilde{\mathbf{s}}_m^{[i]}}^{-1})_{(i,i)} = 1$, which requires $(\mathbf{c}_n^{[i]})_n = \mathbf{0}$ and $(\mathbf{c}_m^{[i]})_m = \mathbf{0}$. Thus completes the proof for regIVA-G-N.

B. regIVA-G-A nonidentifiability

Proof: We first refer back to the regIVA-G-A objective function in (5). Evaluated at $\tilde{\mathbf{Y}}_m = \tilde{\mathbf{S}}_m$, (5) is equivalent to

$$\mathcal{J}_{\text{regIVA-G-A}}\left(\mathbf{w}_n^{[i]}\right) \;=\; \mathbf{w}_n^{[i]\top}\mathbf{X}^{[i]}\tilde{\mathbf{\Omega}}_n\mathbf{X}^{[i]\top}\mathbf{w}_n^{[i]} \qquad (7)$$

where we define $\tilde{\mathbf{\Omega}}_n = c^2 \left[\tilde{\mathbf{S}}_n^\top\tilde{\mathbf{S}}_n - \sum_{\substack{m=1\\m\neq n}}^N \tilde{\mathbf{S}}_m^\top\tilde{\mathbf{S}}_m\right] \in \mathbb{R}^{T\times T}$, and we define $c = \frac{1}{T-1}$ for sake of brevity.

This $\tilde{\mathbf{\Omega}}_n$ is equivalently $\tilde{\mathbf{\Omega}}_n = \mathbf{Q}_n\mathbf{Q}_n^\top$, where we define $\mathbf{Q}_n = c\left[z\tilde{\mathbf{S}}_1^\top, \ldots, \tilde{\mathbf{S}}_n^\top, \ldots, z\tilde{\mathbf{S}}_N^\top\right] \in \mathbb{C}^{T\times NK_b}$ as the horizontal concatenation of the N SCVs, where all SCVs except the nth are multiplied by imaginary number z. It follows that:

$$\begin{aligned}\mathbf{X}^{[i]}\tilde{\mathbf{\Omega}}_n\mathbf{X}^{[i]\top} &= \mathbf{A}^{[i]}\mathbf{S}^{[i]}\mathbf{Q}_n\mathbf{Q}_n^\top\mathbf{S}^{[i]\top}\mathbf{A}^{[i]\top}\\ &= \mathbf{A}^{[i]}\mathbf{Q}_n^{[i]}\mathbf{Q}_n^{[i]\top}\mathbf{A}^{[i]\top},\end{aligned}$$

where we define $\mathbf{Q}_n^{[i]} = \mathbf{S}^{[i]}\ \mathbf{Q}_n \in \mathbb{C}^{N\times NK_b}$ as the correlations of all SCVs with each source of the ith dataset. Given uncorrelated SCVs, $\mathbf{Q}_n^{[i]}$ is represented by the block diagonal matrix of N vector blocks: $\mathbf{Q}_n^{[i]} = \bigoplus_{m=1}^N \gamma(m,n)\ (\mathbf{c}_m^{[i]})_m^\top$, with $\gamma(m,n)$ equals 1 when $m = n$ and equals z otherwise.

With $\mathbf{Q}_n^{[i]}$ having this block diagonal structure, it follows that $\mathbf{Q}_n^{[i]}\ \mathbf{Q}_n^{[i]\top} \in \mathbb{R}^{N\times N}$ is a diagonal matrix, with the mth diagonal element given by $\gamma(m,n)^2\ (\mathbf{c}_m^{[i]})_m^\top\ (\mathbf{c}_m^{[i]})_m$, here $\gamma(m,n)^2$ equals 1 when $m = n$ and equals -1 otherwise.

We now consider the eigendecomposition of $\mathbf{Q}_n^{[i]}\ \mathbf{Q}_n^{[i]\top}$. With $\mathbf{Q}_n^{[i]}\ \mathbf{Q}_n^{[i]\top}$ being a diagonal matrix, its eigenvectors are given as an identity matrix, and its eigenvalues are its diagonal elements. The principal eigenvalue is $(\mathbf{c}_n^{[i]})_n^\top\ (\mathbf{c}_n^{[i]})_n \geq 0$, which is the only eigenvalue capable of being positive.

With $\mathbf{A}^{[i]}$ orthogonal, it follows that $\mathbf{A}^{[i]}\ \mathbf{Q}_n^{[i]}\mathbf{Q}_n^{[i]\top}\ \mathbf{A}^{[i]\top}$ has the same eigenvalues of $\mathbf{Q}_n^{[i]}\mathbf{Q}_n^{[i]\top}$, but the corresponding eigenvectors are the columns of $\mathbf{A}^{[i]}$. Thus when the principal eigenvalue is positive ($(\mathbf{c}_n^{[i]})_n^\top\ (\mathbf{c}_n^{[i]})_n > 0$), it follows that the corresponding principal eigenvector of $\mathbf{A}^{[i]}\ \mathbf{Q}_n^{[i]}\mathbf{Q}_n^{[i]\top}\ \mathbf{A}^{[i]\top}$ is uniquely $\pm\mathbf{a}_n^{[i]}$, in which case identifiability is achieved. Therefore, nonidentifiability occurs only when the principal eigenvalue $(\mathbf{c}_n^{[i]})_n^\top\ (\mathbf{c}_n^{[i]})_n$ is non-unique.

As $(\mathbf{c}_n^{[i]})_n^\top\ (\mathbf{c}_n^{[i]})_n$ is the only nonnegative eigenvalue, and all other eigenvalues are nonpositive, then $(\mathbf{c}_n^{[i]})_n^\top\ (\mathbf{c}_n^{[i]})_n$ is non-unique only when $(\mathbf{c}_n^{[i]})_n^\top\ (\mathbf{c}_n^{[i]})_n$ is equal to 0 *and* one of the $N-1$ other eigenvalues is also equal to 0. This only occurs for the nth and mth SCVs when $(\mathbf{c}_n^{[i]})_n^\top\ (\mathbf{c}_n^{[i]})_n = 0$ and $(\mathbf{c}_m^{[i]})_m^\top\ (\mathbf{c}_m^{[i]})_m = 0$, requiring that $(\mathbf{c}_n^{[i]})_n = \mathbf{0}$ and $(\mathbf{c}_m^{[i]})_m = \mathbf{0}$. Thus completes the proof for regIVA-G-A.

Therefore, for either regIVA-G method, identifiability of $\mathbf{S}^{[i]}$ is possible so long as there is not more than one source in $\mathbf{S}^{[i]}$ that is uncorrelated with its SCV. An example of this occurance is when several sources are random "noise" sources uncorrelated to all other sources in the system. Yet if there is only one source $\mathbf{s}_n^{[i]}$ where $(\mathbf{c}_n^{[i]})_n = \mathbf{0}$, then $\mathbf{s}_n^{[i]}$ is still identifiable because the remaining sources have $(\mathbf{c}_m^{[i]})_m \neq \mathbf{0}$.

It is also notable that this condition specifically depends on the subset's sources; with a different choice of subset, the correlations $(\mathbf{c}_n^{[i]})_n$ will be different. Thus, there can be cases where identifiability is not possible with one subset and possible with another.

In the next section, we discuss how performance of the regIVA-G methodology can be improved by a specific choice of the subset, referred to as "coreIVA-G".

VI. COREIVA-G: REGIVA-G WITH CORESET SELECTION

Performance of regIVA-G is predicated on the subset choice. Intuitively, the best subset is one that is most "representative" of all K datasets. A perfectly representative K_b subset ideally should produce the same results as using all K datasets in place of the subset, resulting in an estimation that is comparable to IVA-G on all K datasets simultaneously.

This section develops a measure of a subset's representativeness in the context of IVA-G. To simplify derivations, we assume SCVs are uncorrelated and datasets $\mathbf{X}^{[k]}$ are prewhitened, thus the $\mathbf{A}^{[k]}$ are orthogonal. However as we show later, this measure is also applicable in general when SCVs may not be uncorrelated or the datasets not prewhitened, and also applicable to other JBSS methods that only model $\mathbf{S}^{[k]}$, opening the possibility of subset-based methods for efficiently optimizing other JBSS objective functions.

A. coreIVA-G subset selection: Cost function

We start with the regIVA-G-A's objective function (5), and assume the subset's sources are exactly identified such that $\tilde{\mathbf{Y}}_n = \tilde{\mathbf{S}}_n$. Our goal will be to compare (5) evaluated over a particular K_b subset to (5) evaluated over all K datasets.

From Section V.B, we note that (5) can be rewritten as in (7): $\mathcal{J}_{\text{regIVA-G-A}}\left(\mathbf{w}_n^{[i]}\right) \;=\; \mathbf{w}_n^{[i]\top}\mathbf{X}^{[i]}\tilde{\mathbf{\Omega}}_n\mathbf{X}^{[i]\top}\mathbf{w}_n^{[i]}$. Here, we scale (7) via $\tilde{\mathbf{\Omega}}_n$ by $\frac{(T-1)^2}{K_b}$, such that we now redefine $\tilde{\mathbf{\Omega}}_n = \frac{1}{K_b}\ [\tilde{\mathbf{S}}_n^\top\ \tilde{\mathbf{S}}_n - \sum_{\substack{m=1\\m\neq n}}^N \tilde{\mathbf{S}}_m^\top\ \tilde{\mathbf{S}}_m] \in \mathbb{R}^{T\times T}$.

To consider the "representativeness" of the K_b subset's $\tilde{\mathbf{\Omega}}_n$ matrix over all K datasets, we also consider this matrix over all K datasets. Thus, we similarly define $\mathbf{\Omega}_n = \frac{1}{K}\ [\mathbf{S}_n^\top\ \mathbf{S}_n - \sum_{\substack{m=1\\m\neq n}}^N \mathbf{S}_m^\top\ \mathbf{S}_m] \in \mathbb{R}^{T\times T}$ evaluated over all K datasets.

Comparing the regIVA-G-A (7) over K_b datasets to (7) over K datasets, these objective functions only differ between $\tilde{\mathbf{\Omega}}_n$ and $\mathbf{\Omega}_n$. Thus, we can measure the "representativeness" of a K_b subset by the distance of its $\tilde{\mathbf{\Omega}}_n$ from $\mathbf{\Omega}_n$. If we denote S_{K_b} as the index set that specifies a K_b subset's datasets from the K total, representativeness of that subset can be measured by the squared Frobenius distance measure $\mathcal{R}\left(S_{K_b}\right)$:

$$\begin{aligned}\mathcal{R}\left(S_{K_b}\right) &= \left\|\tilde{\mathbf{\Omega}}_n - \mathbf{\Omega}_n\right\|_{\mathrm{F}}^2 &(8)\\ &= \left\|\mathbf{\Sigma}_n - \sum_{\substack{m=1\\m\neq n}}^{N}\mathbf{\Sigma}_m\right\|_{\mathrm{F}}^2 &(9)\end{aligned}$$

where we define $\mathbf{\Sigma}_n = \frac{1}{K_b}\ \tilde{\mathbf{S}}_n^\top\ \tilde{\mathbf{S}}_n - \frac{1}{K}\ \mathbf{S}_n^\top\ \mathbf{S}_n \in \mathbb{R}^{T\times T}$.

Assuming the N SCVs are uncorrelated to each other, it is straightforward to show that $\langle\ \mathrm{vec}(\mathbf{\Sigma}_m)\ ,\ \mathrm{vec}(\mathbf{\Sigma}_n)\ \rangle = 0$ for $m \neq n$, where vec(.) denotes the vectorization. This is useful considering vectorized quantities: $\|\mathbf{a}+\mathbf{b}\|_{\mathrm{F}}^2 = \|\mathbf{a}\|_{\mathrm{F}}^2 + \|\mathbf{b}\|_{\mathrm{F}}^2 + 2\langle\mathbf{a},\mathbf{b}\rangle$, as (9) can be equivalently represented using vectorized forms of matrices, and $\|\mathbf{a}+\mathbf{b}\|_{\mathrm{F}}^2 = \|\mathbf{a}\|_{\mathrm{F}}^2 + \|\mathbf{b}\|_{\mathrm{F}}^2 = \|\mathbf{a}-\mathbf{b}\|_{\mathrm{F}}^2$ when $\langle\mathbf{a},\mathbf{b}\rangle = 0$. Thus, (9) can be rewritten as:

$$\mathcal{R}\left(S_{K_b}\right) = \left\|\boldsymbol{\Sigma}_n - \sum_{\substack{m=1\\m\neq n}}^{N} \boldsymbol{\Sigma}_m\right\|_{\mathrm{F}}^2 = \sum_{n=1}^{N} \left\|\boldsymbol{\Sigma}_n\right\|_{\mathrm{F}}^2 = \left\|\sum_{n=1}^{N} \boldsymbol{\Sigma}_n\right\|_{\mathrm{F}}^2$$

This allows us to write $\mathcal{R}\left(S_{K_b}\right)$ in terms of "embeddings" of each dataset's sources ${\mathbf{S}^{[k]}}^\top \mathbf{S}^{[k]}$:

$$\begin{aligned}\left\|\sum_{n=1}^{N} \boldsymbol{\Sigma}_n\right\|_{\mathrm{F}}^2 &= \left\|\sum_{n=1}^{N}\left[\frac{1}{K_b}\sum_{k\in S_{K_b}} \mathbf{s}_n^{[k]}{\mathbf{s}_n^{[k]}}^\top - \frac{1}{K}\sum_{k\in S_K} \mathbf{s}_n^{[k]}{\mathbf{s}_n^{[k]}}^\top\right]\right\|_{\mathrm{F}}^2 \\ &= \left\|\frac{1}{K_b}\sum_{k\in S_{K_b}} {\mathbf{S}^{[k]}}^\top \mathbf{S}^{[k]} - \frac{1}{K}\sum_{k\in S_K} {\mathbf{S}^{[k]}}^\top \mathbf{S}^{[k]}\right\|_{\mathrm{F}}^2 \end{aligned} \tag{10}$$

where S_K represents the index set of all K datasets.

This (10) is particularly useful because if we assume the SCVs are uncorrelated and the data is whitened (thus the $\mathbf{A}^{[k]}$ are orthogonal matrices), it follows that ${\mathbf{S}^{[k]}}^\top \mathbf{S}^{[k]} = {\mathbf{S}^{[k]}}^\top {\mathbf{A}^{[k]}}^\top \mathbf{A}^{[k]} \mathbf{S}^{[k]} = {\mathbf{X}^{[k]}}^\top \mathbf{X}^{[k]}$. This means that without even knowing the underlying sources, $\mathcal{R}\left(S_{K_b}\right)$ can be written just in terms of the original whitened datasets:

$$\mathcal{R}\left(S_{K_b}\right) = \left\|\frac{1}{K_b}\sum_{k\in S_{K_b}} {\mathbf{X}^{[k]}}^\top \mathbf{X}^{[k]} - \boldsymbol{\Psi}\right\|_{\mathrm{F}}^2 \tag{11}$$

where $\boldsymbol{\Psi} = \frac{1}{K}\sum_{k\in S_K} {\mathbf{X}^{[k]}}^\top \mathbf{X}^{[k]} \in \mathbb{R}^{T\times T}$ is what we call the "mean projection embedding" (MPE) of all K datasets, a fixed quantity that we aim to approximate with our subset.

This allows directly measuring a subset's representativeness *before* JBSS, motivating combinatorial optimization methods to select a subset that "best" minimizes (11).

We now discuss how (11) can be used to define "representativeness" in the general JBSS context, not just for IVA-G.

Most JBSS methods represent datasets as *linear subspaces* (specifically all JBSS methods that model only the $\mathbf{S}^{[k]}$ and not the $\mathbf{A}^{[k]}$), as these methods are invariant to the $\mathbf{A}^{[k]}$. Both $\mathbf{X}^{[k]}$ and $\mathbf{S}^{[k]}$ effectively provide orthonormal bases for the same N dimensional linear subspace of $\mathbb{R}^T$. While a choice of orthonormal basis for this kth subspace is not unique, the quantity ${\mathbf{X}^{[k]}}^\top\ \mathbf{X}^{[k]} \in \mathbb{R}^{T\times T}$, commonly known as a *projection matrix*, provides a unique, canonical representation of that kth subspace. This is because ${\mathbf{X}^{[k]}}^\top\ \mathbf{X}^{[k]}$ is invariant to any realization of orthonormal basis $\mathbf{X}^{[k]}$ (as ${\mathbf{X}^{[k]}}^\top\ \mathbf{X}^{[k]} = {\mathbf{X}^{[k]}}^\top\ \mathbf{A}\ \mathbf{A}^\top\ \mathbf{X}^{[k]}$ for any orthogonal matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$).

While ${\mathbf{X}^{[k]}}^\top\ \mathbf{X}^{[k]}$ is more commonly referred to as a projection matrix (e.g. in ordinary least squares regression), in the study of linear subspaces via the Grassmannian manifold [26], ${\mathbf{X}^{[k]}}^\top\ \mathbf{X}^{[k]}$ is called the "projection embedding" of its linear subspace (as it "embeds" the subspace into a unique coordinate in $\mathbb{R}^{T^2}$). When these ${\mathbf{X}^{[k]}}^\top\ \mathbf{X}^{[k]}$ are averaged across K datasets, the resultant "mean projection embedding" (MPE) $\boldsymbol{\Psi}$ provides a statistic capturing information shared across the subspaces (namely, subspaces that are "most shared" across the $\mathbf{X}^{[k]}$). In that sense, (11) can be interpreted as a discrepancy-based cost function [17], [18], [21] measuring distance between a subset's and the full set's MPEs.

This discrepancy can be further decreased if we consider *weighted* subsets. If we assign weight $\lambda_k \in \mathbb{R}$ to each kth dataset, where weights can be organized into a vector $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_{K_b}] \in \mathbb{R}^{K_b}$, then we can consider a weighted variation of (11):

$$\mathcal{R}\left(S_{K_b}, \boldsymbol{\lambda}\right) = \left\|\frac{1}{K_b}\sum_{k\in S_{K_b}} \lambda_k\ {\mathbf{X}^{[k]}}^\top \mathbf{X}^{[k]}\ -\ \boldsymbol{\Psi}\right\|_{\mathrm{F}}^2, \tag{12}$$

Methods that use weighted subsets to minimize discrepancy-based costs are referred to as *coreset* methods [18]–[22], and thus we refer to "coreIVA-G" as the "regIVA-G" methodology where a weighted subset is constructed to minimize (12). In coreIVA-G, we also apply these weights λ_k to the subset estimation and regression steps, such that regIVA-G's objectives in (4) and (5) coincide with the weighted discrepancy in (12).

B. overview of coreIVA-G

We now overview the coreIVA-G methodology:

1) partitioning step: select K_b datasets that minimize (11), simultaneously learning coreset weights λ_k. Divide the K total into this K_b coreset and the K_a remaining datasets.
2) subset estimation step: perform a weighted IVA-G on the coreset, estimating weighted regressor SCVs.
3) regression step: use the weighted regressor SCVs to separately estimate sources in each of the K_a remaining datasets, using either regIVA-G-N or regIVA-G-A.

Weighting can simply be done by multiplying each coreset $\mathbf{X}^{[k]}$ by its respective λ_k, and steps 2-3 are performed using weighted datasets in place of their nonweighted versions.

For the purpose of efficiently minimizing (12), we consider greedy methods that progressively add one dataset to the subset until K_b datasets are selected. Furthermore, it is significantly more efficient to use kernel methods to minimize (12), as a greedy method would otherwise require constructing the subset's MPE at each ith step for $i = 1, \ldots, K_b$. Instead, kernel methods define the cost only in terms of the kernels between datasets, which only need to be calculated once at the beginning of the subset selection. The canonical choice of kernel for (12) is the inner product between the datasets' embeddings: $\langle\ \mathrm{vec}({\mathbf{X}^{[k]}}^\top \mathbf{X}^{[k]}), \mathrm{vec}({\mathbf{X}^{[k]}}^\top \mathbf{X}^{[k]})\ \rangle$. However, it is more convenient to use the "projection kernel" [26]:

$$\mathrm{ker}(i,j) = \frac{1}{N}\left\|\frac{1}{T-1}\mathbf{X}^{[i]}{\mathbf{X}^{[j]}}^\top\right\|_{\mathrm{F}}^2 = \frac{1}{N}\left\|\hat{\mathbf{C}}_{\mathbf{x}}^{[i,j]}\right\|_{\mathrm{F}}^2$$

The kernel is normalized in [0 1], where 0 indicates the subspaces are orthogonal, and 1 indicates the subspaces are equivalent. This is especially useful as the $\hat{\mathbf{C}}_{\mathbf{x}}^{[i,j]}$ can be calculated as done with IVA-G, thus $\hat{\mathbf{C}}_{\mathbf{x}}^{[i,j]}$ can be used for both the kernel and the JBSS procedure. Our implementation of minimizing (12) is the "weighted kernel herding" (WKH) method [22], a greedy method with theoretical guarantees such as the property of "weak submodularity". It is notable that when greedily performing WKH such that each new kth dataset is learned aside its weight λ_k, this ensures that (12) can only decrease or stay constant as the subset size increases.

A final consideration is how the subset size K_b should be determined for coreIVA-G. Provided that the kernels between the K datasets are organized into a matrix $\mathbf{\Theta} \in \mathbb{R}^{K \times K}$ such that $(\mathbf{\Theta})_{i,j} = \text{ker}(i,j)$, we may assume that the optimal K_b is the number of datasets necessary to model $\mathbf{\Theta}$ with a low-rank approximation. Thus we may assume K_b is the "rank" of $\mathbf{\Theta}$, motivating techniques using the eigenspectra of $\mathbf{\Theta}$ to select K_b. However with WKH, as (12) can only decrease or stay constant as the subset size increases. a practical choice of K_b can be made at the point which (12) stops decreasing, which agrees with the aforementioned rank-based methods provided that $\mathbf{\Theta}$ is low-rank. In practice when K_b is not specified in the greedy procedure, we select K_b when the K_bth weight w_{K_b} is sufficiently small: $w_{K_b} \leq \tau$. We find $\tau = 0.001$ is a good choice for the general case.

In the next section, we demonstrate performance of several JBSS algorithms, including regIVA-G (with a random subset of datasets) and coreIVA-G (with a WKH subset), applied to separating simulated data. We demonstrate how each algorithm's separation performance depends on the statistics of the underlying sources. After that, we demonstrate performance on real fMRI sources over a large number of datasets.

VII. Results

We use joint inter-symbol-interference (joint-ISI or jISI) to study separation performance of JBSS when $\mathbf{A}^{[k]}$ are known, such as in the case of simulations. jISI is given by:

$$\text{ISI}_{\text{JNT}}\left(\boldsymbol{\mathcal{W}}, \boldsymbol{\mathcal{A}}\right) \triangleq \frac{1}{2N(N-1)} \left[\sum_{n=1}^{N} \left(\sum_{m=1}^{N} \frac{\bar{g}_{[n,m]}}{\max_p \bar{g}_{[n,p]}} - 1 \right) + \sum_{m=1}^{N} \left(\sum_{n=1}^{N} \frac{\bar{g}_{[n,m]}}{\max_p \bar{g}_{[p,m]}} - 1 \right) \right]$$

With $\boldsymbol{\mathcal{W}}$ as the set of all $\mathbf{W}^{[k]}$, $\boldsymbol{\mathcal{A}}$ as the set of all $\mathbf{A}^{[k]}$, $\mathbf{G}^{[k]} = \mathbf{W}^{[k]}\ \mathbf{A}^{[k]}$ is the "mixing-demixing matrix" of the kth dataset, $g_{[m,n]}^{[k]}$ the $[m,n]$ entry in $\mathbf{G}^{[k]}$, and $\bar{g}_{[m,n]} = \frac{1}{K} \sum_{k=1}^{K} |g_{[m,n]}^{[k]}|$. jISI is given in [14] as an extension of the inter-symbol-interference measure (ISI) for BSS introduced in [27]. jISI is normalized in [0 1], and collectively measures how close each $\mathbf{G}^{[k]}$ matrix is to a permuted diagonal matrix, with 0 jISI indicative of perfect separation.

We also use cross joint inter-symbol-interference (cross-jISI) as an alternative performance measure when $\mathbf{A}^{[k]}$ are not known, such as with real-world data. Cross-jISI is also normalized in [0 1] and measures "consistency" of a JBSS algorithm's estimated sources across different initializations of the data: if cross-jISI is nearly 0, then essentially the same sources are estimated regardless of an algorithm's initialization [28]. The cross-jISI between two "runs" (initializations) uses nearly the same formula for jISI except $\boldsymbol{\mathcal{W}}$ is the set of all $\mathbf{W}^{[k]}$ estimated for one run and $\boldsymbol{\mathcal{A}}$ is the set of inverses of all $\mathbf{W}^{[k]}$ estimated for another run. For our experiments, our reported cross-jISI values are averaged across all pairs of runs, recording the average "distance" between any two runs.

As our paper focuses on efficient JBSS, we limit our results to the source correlation-based JBSS methods. These include the MCCA-SUMCORR solution (often simply called MCCA, which we also call in this paper) [9], [10], a nonorthogonal GJD algorithm called GNJD which is state of the art among GJD algorithms [12], IVA-G [14], regIVA-G (random subset) [16], and coreIVA-G. For numerical algorithms, we implemented the default stopping criteria of each algorithm and limited to a maximum of 1000 $\mathbf{W}^{[k]}$ updates. We utilize the efficient "Newton" method [14] when implementing IVA-G.

For all performance evaluations done in Sections VII and VIII, we use the computational resources provided by the UMBC High Performance Computing Facility (HPCF), thus CPU time is reflective of HPCF's capabilities.

A. *Performance with simulated data*

Our SCV generative model for simulated data is as follows. We model each SCV $\mathbf{s}_n$ as a K-dimensional multivariate Gaussian distributed random vector, with mean $\mathbf{0} \in \mathbb{R}^K$ and some specified covariance $\mathbf{C}_{\mathbf{s}_n} \in \mathbb{R}^{K \times K}$. JBSS algorithms that exploit source correlation have statistical capabilities and nonidentifiability conditions dependent on these $\mathbf{C}_{\mathbf{s}_n}$, therefore we provide a comprehensive model for the $\mathbf{C}_{\mathbf{s}_n}$ as follows:

$$\mathbf{C}_{\mathbf{s}_n} = \alpha \mathbf{B}_n + \beta \mathbf{1}\mathbf{1}^\top + \sigma \mathbf{Q} + \zeta \mathbf{I}_K$$

where the following quantities are defined:

- α, β, σ, and ζ are weights in [0 1] that all sum to 1, such that $\mathbf{C}_{\mathbf{s}_n}$ is a correlation matrix.
- $\mathbf{B}_n \in \mathbb{R}^{K \times K}$ is block matrix of R_n^2 total blocks, thus R_n blocks on the main diagonal. Each diagonal block is of a random size (uniformly distributed in $[1\ (K - R_n)]$), constrained such that the diagonal block sizes sum to K. All elements in the (i,j)th block in $\mathbf{B}_n$ equal the (i,j)th element in a matrix $\mathbf{Q}_{\mathbf{B}_n} \in \mathbb{R}^{R_n \times R_n}$, which is randomly generated from the Wishart distribution, normalized such that $\mathbf{Q}_{\mathbf{B}_n}$ is a normalized similarity matrix, and then elementwise-squared such that $\mathbf{Q}_{\mathbf{B}_n}$ is strictly nonnegative. We note that R_n can be seen as the "effective rank" of $\mathbf{B}_n$ (number of unique eigenvalues), and an increase in R_n corresponds to a decrease in the "off-block" values. In a sense $\mathbf{B}_n$ can be understood as the "group structure" of $\mathbf{s}_n$, modeling groups of correlated sources sometimes seen with medical imaging datasets [1], [3], and R_n can be understood as the number of groups in that SCV.
- $\beta\ \mathbf{1}\mathbf{1}^\top \in \mathbb{R}^{K \times K}$ is a matrix where all elements equal β. This matrix can be understood as the minimum threshold of correlation within that SCV (any two sources within $\mathbf{s}_n$ must have correlation of at least β).
- $\mathbf{Q} \in \mathbb{R}^{K \times K}$ is a rank K matrix randomly generated from the Wishart distribution, then normalized and elementwise squared such that $\mathbf{Q}$ is a positive definite, strictly positive normalized similarity matrix (like $\mathbf{Q}_{\mathbf{B}_n}$). This matrix can be understood as adding random variations in correlation to the otherwise simple structured $\mathbf{C}_{\mathbf{s}_n}$, effectively ensuring all eigenvalues and eigenvectors of $\mathbf{C}_{\mathbf{s}_n}$ are unique. This effectively makes the $\mathbf{C}_{\mathbf{s}_n}$ farther from the JBSS nonidentifiability conditions and results in an improved JBSS separation performance.
- $\mathbf{I}_K \in \mathbb{R}^{K \times K}$ is an identity matrix that models the covariance of additive noise of the nth SCV. We model all additive noise signals as being uncorrelated to all other noise signals in the system (as otherwise their correlatedness defines dependence that helps JBSS).

Furthermore, all SCVs are generated jointly together in a concatenated form $\mathbf{s} = [\mathbf{s}_1^\top, \ldots, \mathbf{s}_N^\top]^\top \in \mathbb{R}^{NK}$, which allows us to not only specify the SCV covariance matrices $\mathbf{C}_{\mathbf{s}_n}$ but also the cross-covariance between separate SCVs. To this end, we additionally introduce $\gamma \in [0\ 1]$ as the cross-covariance value shared between any two SCVs, thus any two sources of two different SCVs have correlation γ. Many JBSS methods assume the SCVs are completely independent and thus $\gamma = 0$, however it is notable that JBSS is still possible with dependent SCVs so long as they are *maximally independent*, which are still identifiabile as JBSS methods merely maximize independence among SCVs. This is an important aspect to include in simulations as real-world SCVs are often dependent, such as with medical imaging data. Increasing γ demonstrates a more difficult separation problem for the JBSS methods.

Our simulated experiments test for varying the values of each variable individually, in addition to varying the number of SCVs N and the number of datasets K. Notably, time complexity of JBSS algorithms primarily depends on the data dimensions N and K and less on the statistics of the data.

Each experiment varies one variable while fixing all others to a fixed value specified here. Unless otherwise varied, we resort to these default values for variables: $\beta = 0$, $\sigma = 0$, $\zeta = 0.1$, $\gamma = 0$, $N = 8$, $K = 30$, and $T = 50000$. With $N = 8$, we default to 4 of the SCVs having $R_n = 2$ and the other 4 SCVs having $R_n = 3$. Due to the challenging nature of the default variables chosen, the default $\mathbf{C}_{\mathbf{s}_n}$ have a simple block structure that has a highly non-unique eigendecomposition, which allows us a better lens to magnify the different estimation capabilities of the algorithms.

All SCVs are jointly generated from T samples of the multivariate Gaussian random vector $\mathbf{s} = [\mathbf{s}_1, \ldots, \mathbf{s}_N] \in \mathbb{R}^{NK}$ according to specified $\mathbf{C}_{\mathbf{s}_n}$ and the specified γ. Sources are then distributed to their datasets $\mathbf{S}^{[k]}$, then mixed with values in $\mathbf{A}^{[k]}$ drawn from the standard Gaussian distribution.

Each variable's experiment measures jISI and cross-jISI in separate sub-experiments. For the jISI sub-experiment, we perform 1000 data simulations and report average jISI with initializations $\mathbf{W}^{[k]} = \mathbf{I}$. For the cross-jISI sub-experiment, we perform 50 data simulations and provide each simulation with 20 random initializations of $\mathbf{W}^{[k]}$ (all algorithms share the same initializations), and report average cross-jISI over these 50 simulations. The cross-jISI experiments omit MCCA-SUMCORR as it is an analytic solution invariant to initializations, thus its cross-jISI can be treated as 0.

We also note that regIVA-G-N and regIVA-G-A perform nearly the same for all experiments in terms of jISI and cross-jISI except for when the experiment is varying the SCV cross-correlation γ. This was also observed for coreIVA-G-N and coreIVA-G-A. Thus to simplify those experiment's plots, we refer to regIVA-G as the performance shared by both regIVA-G-A and regIVA-G-N, and coreIVA-G as the performance shared by both coreIVA-G-A and coreIVA-G-N.

Fig. 2 plots the algorithms' CPU time performances with varying the number of datasets K and the number of SCVs N. We first note that MCCA-SUMCORR (MCCA), regIVA-G-A, and coreIVA-G-A are the most efficient of all tested algorithms and have nearly overlapping CPU times when varying either K or N. The MCCA-SUMCORR solution performed here is an analytic solution where the $\mathbf{W}^{[k]}$ are obtained from the N principal eigenvectors of $\hat{\mathbf{C}}_{\mathbf{x}} = \frac{1}{T-1}\ \mathbf{X}\ \mathbf{X}^\top \in \mathbb{R}^{NK \times NK}$, where we define $\mathbf{X} = [\mathbf{X}^{[1]^\top}, \ldots, \mathbf{X}^{[K]^\top}]^\top \in \mathbb{R}^{NK \times T}$ as the vertical concatenation of the K datasets [10]. This leads MCCA-SUMCORR to have a computational complexity of O($(NK)^3$) which is among the lowest complexities of all JBSS algorithms. While MCCA-SUMCORR is efficient for large K and N, we expect regIVA-G-A to outperform in CPU time when $K \to \infty$ due to its asymptotically linear complexity

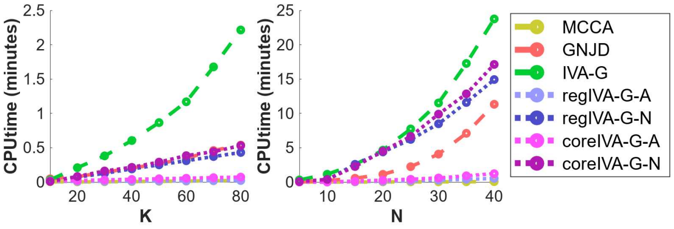

Fig. 2. CPU time (minutes) w.r.t. varying number of datasets K (fixing N=8) and number of sources N (fixing K=30). MCCA, regIVA-G-A and coreIVA-G-A overlap in varying K and N. GNJD and coreIVA-G-N overlap in varying K.

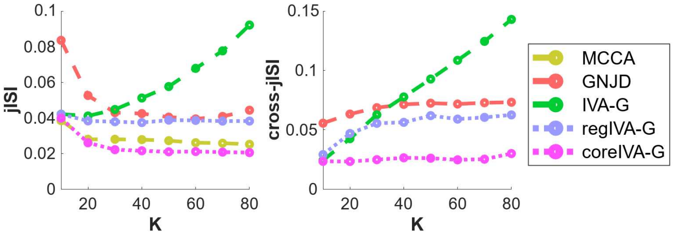

Fig. 3. jISI and cross-jISI w.r.t. varying number of datasets K. regIVA-G methods overlap ("regIVA-G"), coreIVA-G methods overlap ("coreIVA-G"). All cross-jISI figures (including this one) omit MCCA, as MCCA's cross-jISI is always 0.

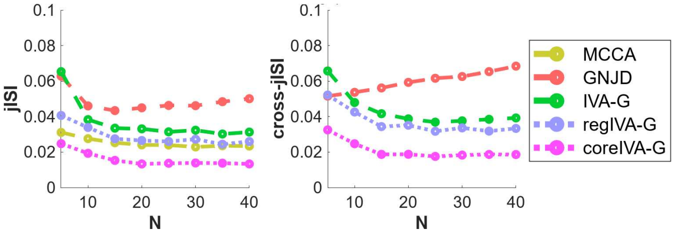

Fig. 4. jISI and cross-jISI w.r.t. varying number of SCVs N.

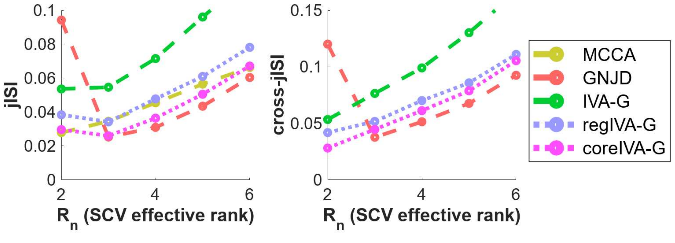

Fig. 5. jISI and cross-jISI w.r.t. varying the SCVs' "effective rank" R_n (number of blocks in each $\mathbf{C}_{\mathbf{s}_n}$).

with K. On the other hand, regIVA-G-N and coreIVA-G-N are significantly slower algorithms, primarily due to their numerical optimization of each ith remaining dataset, resulting in CPU times comparable with GNJD. Finally, IVA-G is the most expensive of all tested algorithms. These plots were observed using the default values of all variables, however we note that across all experiments, each algorithm's time was observed to essentially only depend on the dimensions K and N and not depend on the statistics of the data.

Fig. 3 plots the algorithms' average jISI and cross-jISI performances with varying the number of datasets K. Due to the challenging nature of the default variables chosen, IVA-G and GNJD have significantly worse estimation capabilities than the other algorithms. We note that when the $\mathbf{C}_{\mathbf{s}_n}$ have a simple low effective rank structure, performance of IVA-G suffers when K is very large since IVA-G overparameterizes SCVs. Conversely, GNJD performs worse when K is small. On the other hand, MCCA-SUMCORR assumes a generative model where each SCV is a "common source" shared across the K datasets, thus modeling each SCV $\mathbf{S}_n$ as an effectively rank 1 matrix [9]. Apparently this simpler parameterization allows MCCA-SUMCORR to outperform with simpler $\mathbf{C}_{\mathbf{s}_n}$. Finally, we observed coreIVA-G to be the best jISI performing algorithm with increasing K. As regIVA-G and coreIVA-G use a smaller number of datasets K_b to model the remaining datasets, K_b becomes the effective dimensionality of the SCVs (thus avoiding SCV overparameterization), allowing these methods to maintain good performance with large K.

Fig. 4 plots the algorithms' average jISI and cross-jISI performances with varying the number of SCVs N. Like in the previous experiment with varying K, IVA-G and GNJD perform the poorest among all tested algorithms, whereas the other algorithms perform significantly better in jISI in order of regIVA-G, MCCA-SUMCORR, and coreIVA-G (best).

Fig. 5 plots the algorithms' average jISI and cross-jISI performances with varying the number of blocks in each SCV R_n. We observe all algorithms perform worse with increasing R_n, which we believe is the result of less correlation in the $\mathbf{C}_{\mathbf{s}_n}$ due to the way the $\mathbf{C}_{\mathbf{s}_n}$ are generated, and the $\mathbf{C}_{\mathbf{s}_n}$ possibly being closer to nonidentifiability conditions. We notably observed that GNJD performed poorly with $R_n = 2$, but was among the best performing algorithms when R_n is large, with performance similar to coreIVA-G. Like in the previous experiments, coreIVA-G is among the best performing of all tested algorithms.

Fig. 6 plots the algorithms' average jISI and cross-jISI performances with varying β, the minimum correlation between any two sources in the same SCV. All algorithms except GNJD perform better with increasing β, whereas GNJD performs significantly worse with larger β. This may possibly be due to GNJD's separation performance being more sensitive to when all SCVs have $\mathbf{C}_{\mathbf{s}_n}$ with more similar elements (elements become closer together). Like the previous simulations, coreIVA-G is among the best performing of all tested algorithms.

Fig. 7 plots the algorithms' average jISI and cross-jISI performances with varying σ, the amount of Wishart random "variability" added to the $\mathbf{C}_{\mathbf{s}_n}$, which deviates the $\mathbf{C}_{\mathbf{s}_n}$ from having a non-unique eigendecomposition and deviates from JBSS nonidentifiability. All algorithms except MCCA-SUMCORR perform better with larger σ, whereas MCCA-SUMCORR performs slightly worse with larger σ. We anticipate that this is due to the fact that as MCCA-SUMCORR effectively models each SCV (and thus each $\mathbf{C}_{\mathbf{s}_n}$) as a rank 1 matrix, adding the Wishart variability tends the $\mathbf{C}_{\mathbf{s}_n}$ closer to a rank K model and thus tends farther from the MCCA-SUMCORR assumed model. IVA-G in particular performs the best when σ is high, which we attribute to the K-dimensional maximum likelihood SCV model providing the best SCV model in this scenario. Like the previous simulations, coreIVA-G is among the best performing of all tested algorithms, only

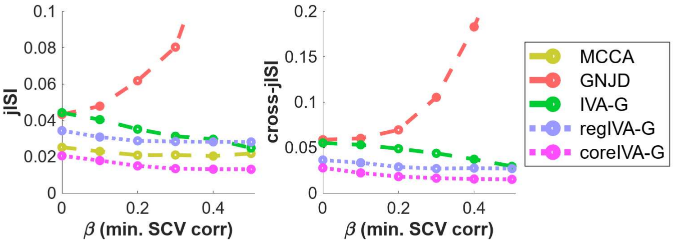

Fig. 6. jISI and cross-jISI w.r.t. varying β, the minimum correlation between any two sources in the same SCV.

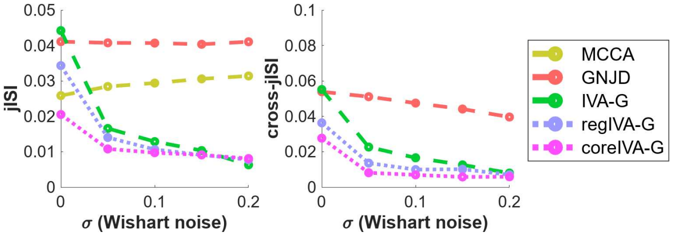

Fig. 7. jISI and cross-jISI w.r.t. varying σ, the amount of Wishart random "variability" added to the $\mathbf{C}_{\mathbf{s}_n}$.

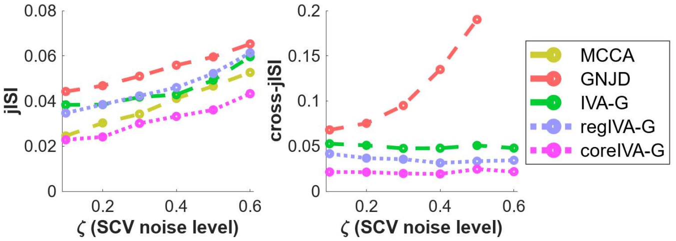

Fig. 8. jISI and cross-jISI w.r.t. varying ζ, the level of additive noise in the SCVs.

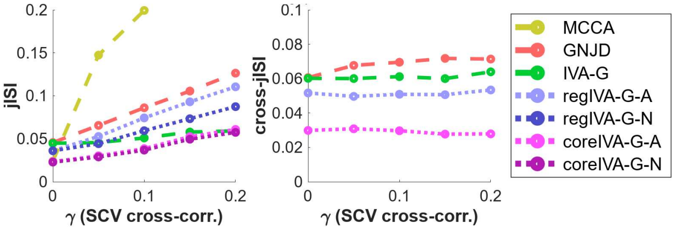

Fig. 9. jISI and cross-jISI w.r.t. varying γ, the SCV cross-correlation. regIVA-G methods overlap in cross-jISI. coreIVA-G methods overlap in jISI and cross-jISI.

beaten by IVA-G in jISI when σ is high.

Fig. 8 plots the algorithms' average jISI and cross-jISI performances with varying ζ, the level of additive noise in the SCVs. Increasing ζ decreases the total level of correlation in the SCVs and results in a harder JBSS problem. All algorithms perform worse with greater noise, with the relationship between ζ and jISI appearing to be linear. Interestingly, the cross-jISI of all algorithms is not as affected by greater noise, aside from GNJD which has a exponential plot very similar to its cross-jISI plot observed for increasing β. An increase in β or ζ both correspond to all off-diagonal values in each $\mathbf{C}_{\mathbf{s}_n}$ becoming closer together, which may predict performance issues for GNJD. Like the previous simulations, coreIVA-G is among the best performing of all tested algorithms.

Fig. 9 plots the algorithms' average jISI and cross-jISI performances with varying γ, the cross-correlation between separate SCVs. $\gamma = 0$ corresponds to uncorrelated SCVs, whereas increasing γ presents a harder JBSS problem. We first note that this is the only case of changing the generative model's variables where we observed a difference in jISI between the analytic and numerical methods of regIVA-G and coreIVA-G: when $\gamma > 0$, regIVA-G-N outperformed regIVA-G-A in jISI, and coreIVA-G-N (slightly) outperformed coreIVA-G-A in jISI. We anticipate that this is because the numerical methods' cost functions per each ith remaining dataset are a function of all demixing vectors $\mathbf{w}_n^{[i]}$, whereas the analytic methods' objective functions are a function of a single demixing vector at a time. In particular, the $\log\mid\det\left(\mathbf{W}^{[k]}\right)\mid$ term in the numerical methods may lead to better performance with correlated SCVs (otherwise $\log\mid\det\left(\mathbf{W}^{[k]}\right)\mid = 0$ when SCVs are uncorrelated and the datasets are prewhitened). However we note that with coreIVA-G, the difference in jISI between coreIVA-G-N and coreIVA-G-A is observed to be very small, which justifies coreIVA-G-A as practical method not just in time complexity but also in separation performance.

Next, we study the performance of the JBSS algorithms in the context of a resting-state fMRI data experiment.

B. fMRI data experiment

One common application of JBSS is for analyzing medical imaging datasets, particularly with fMRI data [1]–[3]. For many fMRI datasets, most SCVs estimated by JBSS are typically both low-rank and correlated to each other, with statistics like those modeled in Section VII.A. Thus, fMRI datasets typically have SCVs that are challenging to estimate for algorithms that only exploit source correlation. At the same time, fMRI datasets are also typically very large and thus also necessitate efficient JBSS algorithms, like those algorithms only exploiting source correlation.

Our experiments use the resting-state fMRI data from the bipolar-schizophrenia network on intermediate phenotypes (B-SNIP) [29], [30]. We used subject datasets available at multiple sites for a total of $K = 1175$ subjects. A single 5-minute resting fMRI scan was captured for each subject, who were instructed to maintain an open-eyed state, concentrate on a crosshair presented on a display screen, and remain still throughout the scanning process. At least $R \geq 97$ time points were obtained for each subject. We removed the first 3 time points to address the T_1 effect and each subject's data was preprocessed including motion correction and slicetime correction. Each subject image was masked, yielding a matrix $\tilde{\mathbf{X}}^{[k]} \in \mathbb{R}^{R\times T}$ where each of the R time point rows was a flattened observation vector of $T = 57878$ voxels. We then standardized and whitened these $\tilde{\mathbf{X}}^{[k]}$ using PCA, and the first N principal components were retained for the subsequent JBSS performance evaluations. The order $N = 80$ was chosen by selecting an adequate order analyzing post-analysis results, however, we also note that $N = 80$ was the largest possible order we could use given the HPCF maximum available memory (350 GB). We thus preprocessed $K = 1175$ subjects' fMRI datasets $\mathbf{X}^{[k]} \in \mathbb{R}^{80\times 57878}$, $k = 1, \ldots, K$, which were then used to perform JBSS.

Given the massive size of this fMRI data, which would be infeasible for most JBSS methods (including IVA-G, GJD methods, etc.), we perform JBSS on the data using only two methods: coreIVA-G-A and MCCA-SUMCORR. With $NK >> T$, we performed the MCCA-SUMCORR solution from a singular value decomposition (SVD) of $\mathbf{X} = [\mathbf{X}^{[1]^\top}, \ldots, \mathbf{X}^{[K]^\top}]^\top \in \mathbb{R}^{NK\times T}$, as estimating the MCCA-SUMCORR $\mathbf{W}^{[k]}$ from the N left singular vectors of $\mathbf{X}$ provides a significantly more CPU and memory efficient alternative to calculating these from the eigendecomposition of of $\hat{\mathbf{C}}_{\mathbf{x}} = \frac{1}{T-1}\,\mathbf{X}\,\mathbf{X}^\top \in \mathbb{R}^{NK\times NK}$. This method of performing MCCA-SUMCORR is mathematically equivalent to group-PCA [31], and thus can be seen as performing a group-level PCA on the K datasets to select N group level components whose corresponding weights in the PCA form the demixing matrices $\mathbf{W}^{[k]}$ [10]. We used $K_b = 40$ for coreIVA-G based on higher quality post-analysis results with this K_b.

As we do not have ground-truth sources with real data (and thus can't directly measure source separation with jISI), we use other performance measures on the results:

- total CPU-time to estimate all $\mathbf{W}^{[k]}$
- cross-jISI between $\mathbf{W}^{[k]}$ of runs with different initializations (measuring consistency in estimated sources)
- "mean PSR": average of the SCV's power spectral ratios (PSR). PSR is defined as the power ratio between low-frequency (< 0.1 Hz) and high-frequency (> 0.15 Hz) bands within estimated sources. Considering the frequencies of neural-activity related BOLD signals are generally below 0.15 Hz, high power ratio values typically indicate BOLD activity and low power ratio values typically associate with noisier estimates and artifacts [32].

To measure cross-jISI, we ran coreIVA-G 10 different times with 10 random initializations for the estimated $\mathbf{W}^{[k]}$. When plotting estimated fMRI sources, we retained the run that had the minimum cross-jISI between all other runs (the run "most similar" to all runs). As MCCA-SUMCORR is performed with SVD, which has an analytic solution (invariant to initialization), we report MCCA-SUMCORR's cross-jISI as 0 and use results from a single run.

Table 2 presents the performance measures of coreIVA-G and MCCA-SUMCORR (MCCA) on the fMRI data. We first note that MCCA-SUMCORR takes about 20% more

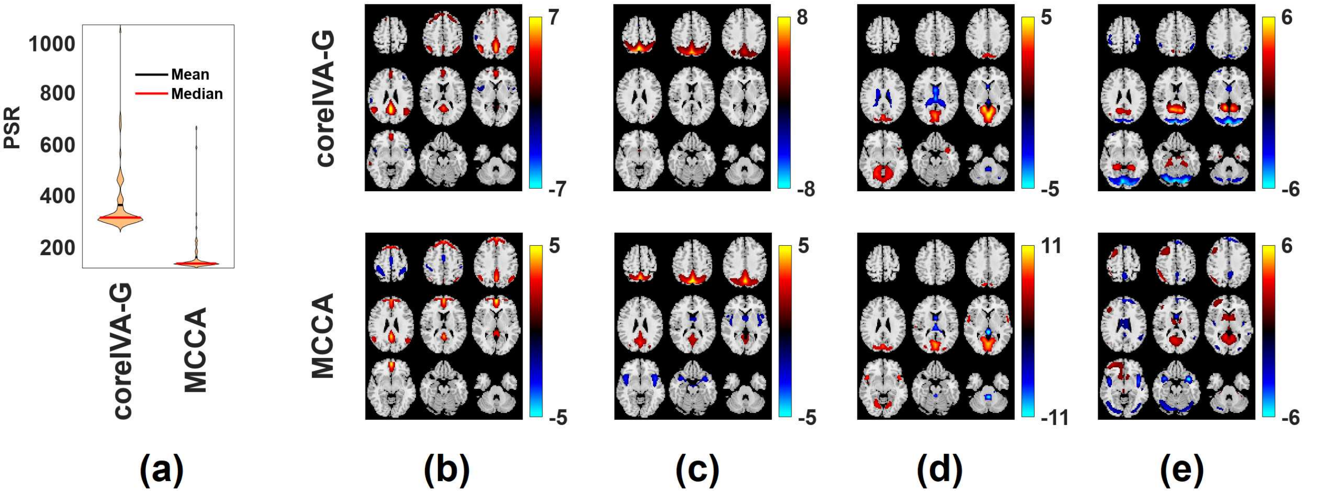

Fig. 10. Plots of JBSS estimated sources for coreIVA-G and MCCA-SUMCORR (MCCA) obtained from the fMRI data. (a) Distribution of power spectral ratio (PSR) values over the $N = 80$ SCVs, for both coreIVA-G and MCCA-SUMCORR. (b) (c) (d) (e) plot the principal right singular vector of four different SCVs (a way of summarizing the K sources in the corresponding SCV $\mathbf{Y}_n \in \mathbb{R}^{K \times T}$, accounting for possible sign ambiguities of the sources). (b) and (c) correspond to two SCVs representing the default mode network (DMN), (d) and (e) correspond to two SCVs representing the visual (VIS) networks.

Table 2. Performance measures of coreIVA-G and MCCA-SUMCORR (MCCA) on the $K = 1175$ fMRI datasets. The best performing algorithm per measure has its value in bold. coreIVA-G's CPU-time is averaged over 10 runs. Mean PSR is the average of all $N = 80$ SCV's power spectral ratios. We included only coreIVA-G and MCCA-SUMCORR as other methods (such as IVA-G, GJD, etc.) were computationally infeasible for this large dataset.

	CPU-time (hours)	cross-jISI	mean PSR
coreIVA-G	**40.88**	9.42e-9	**3.16**
MCCA	48.78	**0**	2.02

CPU-time. We anticipate the relative time difference between the algorithms would increase significantly with larger K due to the O(K) complexity of the coreIVA-G regression step, whereas MCCA-SUMCORR has a complexity with K of O(K^3). We then note that MCCA-SUMCORR has the advantage of an analytic solution and thus the estimated solution is invariant to the initialization (cross-jISI of 0), however, coreIVA-G's observed cross-jISI value of 9.42e-9 is low enough such that the difference between sources of any two runs is essentially negligable. We also note that coreIVA-G has higher mean PSR values averaged across the $N = 80$ SCVs, indicating that coreIVA-G generally estimates less noisy sources compared to MCCA-SUMCORR. Fig. 10 (a) presents violin plots visualizing the distribution of the $N = 80$ SCVs' PSR values, further demonstrating that on a whole the PSR values were significantly higher per SCV with coreIVA-G compared to MCCA-SUMCORR.

Fig. 10 also plots spatial maps for the algorithms' estimated SCVs (referred to as networks), two networks corresponding to default mode network (DMN) domains in (b) and (c), and two corresponding to visual domains (VIS) in (d) and (e). Each plot is of the principal right singular vector of that corresponding SCV $\mathbf{Y}_n \in \mathbb{R}^{K \times T}$, providing a way of summarizing the K sources $\mathbf{y}_n^{[k]}$ in $\mathbf{Y}_n$, accounting for possible sign ambiguities. When analyzing plots of estimated sources, we found these differences in coreIVA-G vs. MCCA-SUMCORR:

- "focal" activations correspond to activation peaks at the center of an activated region and a gradual decrease in magnitude away from the region, which is a desired quality in fMRI spatial maps. Sources estimated by MCCA-SUMCORR generally display more noise, particularly the blue plotted areas, and are generally less "focal" than the corresponding sources estimated by coreIVA-G.
- higher activation magnitude within a source may correspond to a better isolation of that source's functional network (FN), which may indicate a better demixing of sources. We observed overall higher activation magnitudes with coreIVA-G than with MCCA-SUMCORR.

We anticipate that these differences are largely due to coreIVA-G being an IVA-based method, which can generally perform better for preserving per-subject variability in SCVs [33]. This is opposed to MCCA-SUMCORR, which was shown in [9] to model SCVs as an effectively rank-1 matrix (a source shared across the datasets). This rank-1 model is expected to perform well when SCVs are highly homogeneous, but may otherwise be outperformed by IVA-based methods when more heterogeneity exists within SCVs.

VIII. Conclusion

This paper presents an efficient methodology for scaling IVA-G on a subset of datasets to a much larger set of datasets, called "regIVA-G". We proposed two such methods for regressing an additional dataset: a numerical solution minimizing the IVA-G cost, and a previously proposed analytic solution [16] with an objective function comparable with those of GJD-based methods. We then derived the regIVA-G methods' nonidentifiability conditions: conditions for which the regIVA-G methods are unable to uniquely identify the true sources. These conditions are highly general (assuming the subset's sources have been identified), highlighting the powerful estimation capabilities of both regIVA-G methods. Following this, we derived a novel tractable cost function for measuring the representativeness of a subset of datasets, comparable to discrepancy-based costs for coreset (representative subset) selection. We thus propose using this discrepancy, in conjunction with weighting the datasets to best minimize the discrepancy, as the "coreIVA-G" method building onto the regIVA-G method. Finally, we experimentally demonstrate that regIVA-G and coreIVA-G methods can significantly outperform other JBSS methods in terms of CPU-time, jISI, and cross-jISI, making these methods highly practical and highly generalizable to many different types of data.

The main limitation of the regIVA-G and coreIVA-G methods is that they only assume a multivariate Gaussian model, and thus only exploit source correlation to perform JBSS. Algorithms that exploit higher-order statistics are generally known for strong performance, and provide superior identifiability conditions when the data's SCVs are non-Gaussian. Thus, future work may generalize these methods to a general "regIVA" or "coreIVA" methodology modeling non-Gaussian distributions as well, in addition to other statistical properties of the data (such as sample dependence within the sources).

References

[1] T. Adalı, M. Anderson, and G.-S. Fu, "Diversity in Independent Component and Vector Analyses: Identifiability, Algorithms, and Applications in medical imaging," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 18–33, 2014.

[2] Y.-O. Li, T. Adalı, W. Wang, and V. D. Calhoun, "Joint Blind Source Separation by Multiset Canonical Correlation Analysis," *IEEE Transactions on Signal Processing*, vol. 57, no. 10, pp. 3918–3929, 2009.

[3] T. Adalı, Y. Levin-Schwartz, and V. D. Calhoun, "Multimodal data fusion using source separation: Application to medical imaging," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1494–1506, 2015.

[4] D. Sugumar, P. Vanathi, and S. Mohan, "Joint Blind Source Separation algorithms in the separation of non-invasive maternal and fetal ECG," in *2014 International Conference on Electronics and Communication Systems (ICECS)*. IEEE, 2014, pp. 1–6.

[5] A. A. Nielsen, "Multiset Canonical Correlations Analysis and multispectral, truly multitemporal remote sensing data," *IEEE transactions on image processing*, vol. 11, no. 3, pp. 293–305, 2002.

[6] T. Kim, T. Eltoft, and T.-W. Lee, "Independent Vector Analysis: An extension of ICA to multivariate components," in *International conference on independent component analysis and signal separation*. Springer, 2006, pp. 165–172.

[7] Z. Boukouvalas, D. C. Elton, P. W. Chung, and M. D. Fuge, "Independent Vector Analysis for data fusion prior to molecular property prediction with machine learning," *arXiv preprint arXiv:1811.00628*, 2018.

[8] M. Anderson, G.-S. Fu, R. Phlypo, and T. Adalı, "Independent Vector Analysis: Identification conditions and performance bounds," *IEEE Transactions on Signal Processing*, vol. 62, no. 17, pp. 4399–4410, 2014.

[9] J. R. Kettenring, "Canonical analysis of several sets of variables," *Biometrika*, vol. 58, no. 3, pp. 433–451, 1971.

[10] L. C. Parra, "Multi-set Canonical Correlation Analysis simply explained," *arXiv preprint arXiv:1802.03759*, 2018.

[11] X.-L. Li, T. Adalı, and M. Anderson, "Joint Blind Source Separation by generalized joint diagonalization of cumulant matrices," *Signal Processing*, vol. 91, no. 10, pp. 2314–2322, 2011.

[12] X.-F. Gong, X.-L. Wang, and Q.-H. Lin, "Generalized non-orthogonal joint diagonalization with LU decomposition and successive rotations," *IEEE Transactions on Signal Processing*, vol. 63, pp. 1322–1334, 2015.

[13] X.-F. Gong, L. Mao, Y.-L. Liu, and Q.-H. Lin, "A Jacobi generalized orthogonal joint diagonalization algorithm for Joint Blind Source Separation," *IEEE Access*, vol. 6, pp. 38 464–38 474, 2018.

[14] M. Anderson, T. Adalı, and X.-L. Li, "Joint Blind Source Separation with multivariate Gaussian model: Algorithms and performance analysis," *IEEE Transactions on Signal Processing*, vol. 60, no. 4, pp. 1672–1683, 2011.

[15] J. Vía, M. Anderson, X.-L. Li, and T. Adalı, "Joint blind source separation from second-order statistics: Necessary and sufficient identifiability conditions," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 2520–2523.

[16] B. Gabrielson, M. Sun, M. A. B. S. Akhonda, V. D. Calhoun, and T. Adali, "Independent Vector Analysis with Multivariate Gaussian Model: a Scalable Method by Multilinear Regression," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[17] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[18] S. A. Williamson and J. Henderson, "Understanding collections of related datasets using dependent MMD coresets," *Information*, vol. 12, no. 10, p. 392, 2021.

[19] B. Mirzasoleiman, J. Bilmes, and J. Leskovec, "Coresets for data-efficient training of machine learning models," in *International Conference on Machine Learning*. PMLR, 2020, pp. 6950–6960.

[20] R. Dwivedi and L. Mackey, "Generalized kernel thinning," *arXiv preprint arXiv:2110.01593*, 2021.

[21] Z. Karnin and E. Liberty, "Discrepancy, coresets, and sketches in machine learning," in *Conference on Learning Theory*. PMLR, 2019, pp. 1975–1993.

[22] F. Huszár and D. Duvenaud, "Optimally-weighted herding is Bayesian quadrature," *arXiv preprint arXiv:1204.1664*, 2012.

[23] B. Gabrielson, M. A. Akhonda, Z. Boukouvalas, S.-J. Kim, and T. Adalı, "ICA with Orthogonality Constraint: Identifiability And A New Efficient Algorithm," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3720–3724.

[24] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and applications*. Academic press, 2010.

[25] D. Olive, *Prediction and Statistical Learning*. Unpublished notes available from http://parker.ad.siu.edu/Olive/slrun.pdf, 2023, p. 362.

[26] M. T. Harandi, M. Salzmann, S. Jayasumana, R. Hartley, and H. Li, "Expanding the family of grassmannian kernels: An embedding perspective," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*. Springer, 2014, pp. 408–423.

[27] S. A. Amari *et al.*, "Advances in neural information processing systems," in *Advances in neural information processing systems*, vol. 8. Cambridge, MA: MIT Press, 1996, p. 757–763.

[28] Q. Long, C. Jia, Z. Boukouvalas, B. Gabrielson, D. Emge, and T. Adali, "Consistent run selection for independent component analysis: Application to fMRI analysis," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2581–2585.

[29] C. A. Tamminga, E. I. Ivleva, M. S. Keshavan, G. D. Pearlson, B. A. Clementz, B. Witte, D. W. Morris, J. Bishop, G. K. Thaker, and J. A. Sweeney, "Clinical phenotypes of psychosis in the Bipolar-Schizophrenia Network on Intermediate Phenotypes (B-SNIP)," *American Journal of psychiatry*, vol. 170, no. 11, pp. 1263–1274, 2013.

[30] C. A. Tamminga, G. Pearlson, M. Keshavan, J. Sweeney, B. Clementz, and G. Thaker, "Bipolar and schizophrenia network for intermediate phenotypes: outcomes across the psychosis continuum," *Schizophrenia bulletin*, vol. 40, no. Suppl_2, pp. S131–S137, 2014.

[31] S. M. Smith, A. Hyvärinen, G. Varoquaux, K. L. Miller, and C. F. Beckmann, "Group-PCA for very large fMRI datasets," *Neuroimage*, vol. 101, pp. 738–749, 2014.

[32] E. A. Allen, E. B. Erhardt, E. Damaraju, W. Gruner, J. M. Segall, R. F. Silva, M. Havlicek, S. Rachakonda, J. Fries, R. Kalyanam *et al.*, "A baseline for the multivariate comparison of resting-state networks," *Frontiers in systems neuroscience*, vol. 5, p. 2, 2011.

[33] Y. Du, D. Lin, Q. Yu, J. Sui, J. Chen, S. Rachakonda, T. Adali, and V. D. Calhoun, "Comparison of IVA and GIG-ICA in brain functional network estimation using fMRI data," *Frontiers in neuroscience*, vol. 11, p. 267, 2017.