Multidataset Independent Subspace Analysis With Application to Multimodal Fusion

Rogers F. Silva, *Member, IEEE*, Sergey M. Plis, Tülay Adalı, *Fellow, IEEE*, Marios S. Pattichis, *Senior Member, IEEE*, and Vince D. Calhoun, *Fellow, IEEE*

***Abstract*— Unsupervised latent variable models—blind source separation (BSS) especially—enjoy a strong reputation for their interpretability. But they seldom combine the rich diversity of information available in multiple datasets, even though multi-datasets yield insightful joint solutions otherwise unavailable in isolation. We present a direct, principled approach to multi-dataset combination that takes advantage of multidimensional subspace structures. In turn, we extend BSS models to capture the underlying modes of shared and unique variability across and within datasets. Our approach leverages joint information from heterogeneous datasets in a flexible and synergistic fashion. We call this method multidataset independent subspace analysis (MISA). Methodological innovations exploiting the Kotz distribution for subspace modeling, in conjunction with a novel combinatorial optimization for evasion of local minima, enable MISA to produce a robust generalization of independent component analysis (ICA), independent vector analysis (IVA), and independent subspace analysis (ISA) in a single unified model. We highlight the utility of MISA for multimodal information fusion, including sample-poor regimes ($N = 600$) and low signal-to-noise ratio, promoting novel applications in both unimodal and multimodal brain imaging data.**

***Index Terms*— BSS, MISA, multidataset, fusion, ICA, ISA, IVA, subspace, unimodal, multimodality, multiset data analysis, unify.**

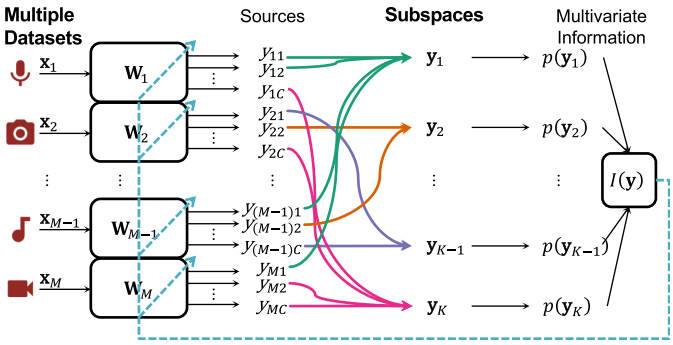


Fig. 1. **Subspace identification from multidatasets with MISA**. We consider the general case of M datasets/modalities ($\mathbf{x}_m$) *jointly* decomposed, without loss of generality, into C sources y_{mi} each, via linear transformations $\mathbf{W}_m$. Here, each $\mathbf{x}_m$ would be either audio or video streams, indicating fusion via the joint analysis of all datasets. Sources are combined into d_k-dimensional *subspaces* $\mathbf{y}_k$ and *all-order* statistics is utilized to gauge their associations and pursue subspace independence. Only a single correspondence "axis" is required, e.g., time, meaning there is a video frame for each audio sample in audio/video (a/v) data fusion, although the method is not limited to a/v, fusion, nor temporal synchrony specifically. Subspaces establish links among groups of sources across different datasets/modalities. Therefore, multidataset independent subspace analysis (MISA) blindly recovers hidden linked features of flexible dimensionality from multiple datasets and modalities. Code is available at https://github.com/rsilva8/MISA.

I. Introduction

BLIND source separation (BSS) [1], [2] is the recovery of unknown latent source signals from their observed mixtures without knowing the mixing process. It is widely adopted in signal, image, and video processing areas, including chemometrics [3], speech [4], multispectral imaging [5], [6], medical imaging [7], [8], and video processing [9], [10]. The "blind" property (unknown source *and* mixing) is highly effective, especially in applications lacking a precise model of the measured system(s) and with data confounded by noise of unknown or variable characteristics.

In our recent review [1], we introduced a unified multi-dataset multidiversity multidimensional framework for subspace modeling. It provided a fresh perspective on BSS, identifying both single-dataset multidimensional (SDM) and multidataset unidimensional (MDU) research as subproblems, and outlining a path to reconcile them. In turn, a new class of multidataset multidimensional (MDM) problems became apparent, emphasizing the potential benefits of general latent subspace correspondence across datasets.

Models designed for MDM problems are extremely flexible. A single *joint* model not only encodes higher complexity through features of flexible dimensionality (the subspaces $\mathbf{y}_k$) but also accommodates arbitrary links among these features over multiple datasets/modalities ($\mathbf{x}_m$). To illustrate (Fig. 1), we consider a multivariate information functional $I(\mathbf{y})$ that operates simultaneously on the joint probability

Manuscript received November 7, 2019; revised May 8, 2020; accepted September 16, 2020. Date of publication October 8, 2020; date of current version November 25, 2020. This work was supported by NIH grants R01EB006841, 5P20RR021938/P20GM103472, R01MH118695, RF1AG063153 and NSF grants 1539067, 1631838, and 1618551. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Shuicheng Yan. *(Corresponding author: Rogers F. Silva.)*

Rogers F. Silva, Sergey M. Plis, and Vince D. Calhoun are with the Tri-Institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), Georgia State University, Georgia Institute of Technology, and Emory University, Atlanta, GA 30303 USA, and also with The Mind Research Network, Albuquerque, NM 87106 USA (e-mail: rsilva@gsu.edu; splis@gsu.edu; vcalhoun@gsu.edu).

Tülay Adalı is with the Department of CSEE, University of Maryland at Baltimore County, Baltimore, MD 21250 USA (e-mail: adali@umbc.edu).

Marios S. Pattichis is with the Department of ECE, The University of New Mexico, Albuquerque, NM 87131 USA (e-mail: pattichi@unm.edu).

This article has supplementary downloadable material available at https://ieeexplore.ieee.org, provided by the authors.

Digital Object Identifier 10.1109/TIP.2020.3028452

density function (pdf) of all subspaces $p(\mathbf{y}_k)$. It captures the association modes underlying multidatasets while adaptively learning multiple linear transformations $\mathbf{W}_m$ (dashed lines). When datasets represent modalities, this *directly* leverages multimodal joint information and lets it guide the decompositions naturally. Combining different multimodal views of the *same* system, this generalized approach to multimodal fusion offers broader, unique insights into its underlying properties and behavior.

Aiming at generality, we pursue *statistical independence* among subspaces $\mathbf{y}_k$ to achieve joint BSS for MDM. Initial investigation of this approach [11]–[13] indicated the presence of critical issues. These included premature convergence to local minima, rigid hard-coded subspace distribution parameters, and a restricted orthogonal regularization for $\mathbf{W}_m$.

Here, we propose a vastly improved expansion to address these issues. We use combinatorial optimization to search over subspace configurations $\mathbf{P}$ (Fig. 2) and escape local minima, all-order statistics (i.e., both second- *and* higher-order statistics—SOS and HOS, respectively) to model $p(\mathbf{y}_k)$ via the more general Kotz distribution [14], and a scale-controlled formulation for numerical stability. We also generalize usage to non-orthogonal $\mathbf{W}_m$ sans data reduction. We refer to this robust, performant approach simply as multidataset independent subspace analysis (MISA) (Fig. 1). In the formulation below, $p(\mathbf{y})$ represents the joint pdf of all sources, and $p(\mathbf{y}_k)$ the pdf of the k-th subspace.

Let $I(\mathbf{y})$ be the Kullback-Leibler (KL) divergence, an information functional useful for comparing two pdfs $p(\mathbf{y})$ and $q(\mathbf{y})$, where, here, $q(\mathbf{y}) = \prod_{k=1}^{K} p(\mathbf{y}_k)$ is the desired *factor* pdf of $p(\mathbf{y})$. Then let $h(\cdot)$ be the joint differential entropy, $h(\mathbf{z}) = -\mathbb{E}[\ln p(\mathbf{z})]$, for a random vector $\mathbf{z}$ with pdf $p(\mathbf{z})$, $\mathbb{E}[\cdot]$ being the expected value operator, and let $\mathbf{P}_k$ be the subset of $\mathbf{P}$ assigning specific sources into subspace k. Consequently,

$$\begin{aligned} I(\mathbf{y}) &= -h(\mathbf{y}) + \sum_{k=1}^{K} h(\mathbf{y}_k) \\ &= -h(\mathbf{W}\mathbf{x}) + \sum_{k=1}^{K} h(\mathbf{P}_k \mathbf{W}\mathbf{x}). \end{aligned} \tag{1}$$

We propose to estimate a collection of linear transformations $\mathbf{y} = \mathbf{W}\mathbf{x}$ simultaneously from all datasets by solving:

$$\min_{\mathbf{W},\mathbf{P}} \; I(\mathbf{y}), \tag{2}$$

for any $\mathbf{W}$, subspace assignments $\mathbf{P}$, and data streams $\mathbf{x}$. This convenient formulation, which gives mutual information (MI) when the random vector $\mathbf{y}$ is two-dimensional, only attains its lower bound of $I(\mathbf{y}) = 0$ when $p(\mathbf{y}) = q(\mathbf{y})$, implying that the identified subspaces are indeed statistically independent. A sketch of the convergence proof for this approach is provided as supplemental material.

With MISA, direct study of the interactions and associations among multiple datasets and modalities becomes feasible, in a truly synergistic way. Consequently, joint sources $\mathbf{y}_k$ emerge naturally as a direct result of the shared variability estimated from all-order statistical dependences among datasets.

TABLE I
FREQUENTLY USED ACRONYMS

SDU single dataset unidimensional	SOS second-order statistics
SDM single dataset multidimensional	HOS higher-order statistics
MDU multidataset unidimensional	pdf probability density function
MDM multidataset multidimensional	GP greedy permutations
ICA independent component analysis	PCA principal component analysis
ISA independent subspace analysis	IVA independent vector analysis

Breaking from the limited, rigid paradigm of MDU models dominating current multimodal research [15, Ch. 8], it allows general subspace associations and even absent features in specific datasets. As a unifying toolkit, MISA can execute many general unconventional BSS tasks as well as classical special cases such as independent component analysis (ICA) [16], independent subspace analysis (ISA) [17], and independent vector analysis (IVA) [18]. Also, it outperforms several algorithms in each of these tasks, successfully achieving generalized subspace identification from multidatasets. This uniform implementation yields user accessibility and intuition thanks to the umbrella formulation and methodologies introduced here.

In the current paper, we demonstrate that MISA (our proposed method) outperforms algorithms such as Infomax [19], [20], Laplace IVA (IVA-L) [18], and Gaussian-Laplace IVA (IVA-GL) [21] in challenging experiments and realistic scenarios satisfying the requisites outlined in [22]. MISA's remarkable performance and stability in certain extremely noisy cases (signal-to-noise ratio (SNR) of 0.0043dB) highlights the benefit of careful multidataset subspace dependence modeling with all-order statistics. Likewise, MISA with greedy permutations (MISA-GP) clearly outperforms joint blind diagonalization with SOS (JBD-SOS) [23] and EST_ISA [24] even at low SNR levels (SNR of 3dB). This shows the benefit of combinatorial optimization to escape local minima in subspace analyses.

Hybrid data results on representative biomedical imaging features and realistic data dimensionality further support the high estimation quality and flexibility of MISA. These include novel applications in high-temporal-resolution functional magnetic resonance imaging (MRI), and multimodal fusion of heterogeneous neurobiological images and signals. The latter also demonstrates feasibility of data fusion even at low SNR and sample-poor regimes (number of observations $N = 600$), with examples involving functional, structural, and diffusion MRI, as well as electroencephalography (EEG) data. Subspace analysis in its general MDM form has not yet been conducted in a multimodal fusion setting. To the best of our knowledge, MISA is the only approach which can directly investigate this use-case using all-order statistics. Original code and data are available at https://github.com/rsilva8/MISA, with examples to accompany the descriptions in supplemental material (Sections II-B and II-D therein), and detailed derivation of the gradients.

In the following, Section II states the general MDM problem. Section III puts our contributions in context with related works, followed by our methodology description in Section IV. Finally, Sections V and VI present our results and conclusions, respectively. Frequently used acronyms are listed in Table I.

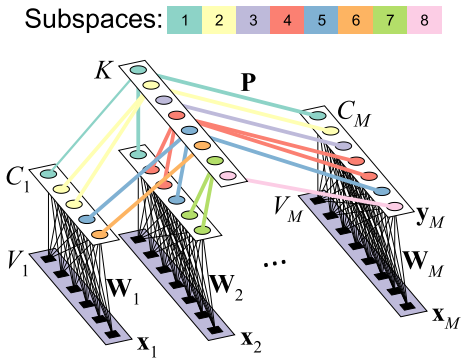


Fig. 2. **General architecture of linear MDM problems**. The lower layer corresponds to one $V_m \times 1$ observation of each input data stream $\mathbf{x}_m$. The middle layer represents the C_m sources. The top layer establishes the K subspaces $\mathbf{y}_k$, which are collections of *statistically dependent* sources (indicated by same-colored connections), following the compositions laid out in the assignment matrix $\mathbf{P}$. This architecture suggests a natural hierarchy among models [1], [15, Ch. 8] in accordance with the number of datasets comprising $\mathbf{x}$ and the occurrence of multidimensional sources within any single dataset. MDU and SDM problems include the simpler SDU case, and the most general MDM problem contains all others as special cases.

II. Background

The MDM problem can be formally stated as follows. Given N observations of $M \geq 1$ datasets, identify an unobservable latent source random vector $\mathbf{y} = \left[\mathbf{y}_1^\top \cdots \mathbf{y}_M^\top\right]^\top$, with $\mathbf{y}_m = [y_{m1} \cdots y_{mC_m}]^\top$ (C_m sources per dataset), from an observed random vector $\mathbf{x} = \left[\mathbf{x}_1^\top \cdots \mathbf{x}_M^\top\right]^\top$, with $\mathbf{x}_m = [x_1 \cdots x_{V_m}]^\top$ (V_m-dimensional datasets), generated via a mixture *vector function* $\mathbf{f}(\mathbf{y}, \boldsymbol{\theta})$ with unknown parameters $\boldsymbol{\theta}$. The m-th $V_m \times N$ data matrix containing N observations of $\mathbf{x}_m$ along its columns is denoted $\mathbf{X}_m$, and the $\bar{V} \times N$ matrix concatenating all $\mathbf{X}_m$ is denoted simply as $\mathbf{X}$ (likewise for $\mathbf{Y}$ and $\mathbf{Y}_m$). Both $\mathbf{y}$ and $\mathbf{f}(\mathbf{y}, \boldsymbol{\theta})$ have to be learned *blindly*, i.e., without knowledge of either of them. For tractability, assume:

1) the number of latent sources C_m, which may differ in each dataset, is known to the experimenter;
2) $\mathbf{f}(\mathbf{y}, \boldsymbol{\theta}) = \mathbf{A}\mathbf{y}$ is a *linear* transformation, with $\boldsymbol{\theta} = \mathbf{A}$;
3) $\mathbf{A}$ is a $\bar{V} \times \bar{C}$ block diagonal matrix with M blocks, describing a *separable layout* structure [1] representing $\mathbf{x}_m = \mathbf{A}_m\mathbf{y}_m$, $m = 1 \ldots M$, where $\bar{C} = \sum_{m=1}^{M} C_m$, $\bar{V} = \sum_{m=1}^{M} V_m$, each block $\mathbf{A}_m$ is $V_m \times C_m$, and V_m is the intrinsic dimensionality of each dataset;
4) some latent sources $y_{mi} \in \mathbf{y}$ are statistically related to each other, and this *dependence* is undirected (non-causal), occurring within and/or across datasets;
5) related sources establish d_k-dimensional subspaces [1] $\mathbf{y}_k$, $k = 1 \ldots K$, with K and the subspace compositions laid out by the experimenter in sparse assignment matrices $\mathbf{P}_k \in \{0,1\}^{d_k \times \bar{C}}$, such that $\mathbf{P} = \left[\mathbf{P}_1^\top \cdots \mathbf{P}_k^\top \cdots \mathbf{P}_K^\top\right]^\top$ is a permutation matrix;
6) subspaces do not relate to each other, i.e., either $p(\mathbf{y}) = \prod_{k=1}^{K} p(\mathbf{y}_k)$ or the cross-correlations $\rho_{k,k'} = 0$, $k \neq k'$.

Under these assumptions, recovering sources $\mathbf{y}$ amounts to finding a linear transformation $\mathbf{W}$ for the unmixing vector function $\mathbf{y} = \mathbf{W}\mathbf{x}$. This occurs when $\mathbf{W} = \mathbf{A}^-$, the pseudo-inverse of $\mathbf{A}$, implying $\mathbf{W}$ is *also* block diagonal and satisfies $\mathbf{y}_m = \mathbf{W}_m\mathbf{x}_m$. The experimenter's priors on the subspace structure within/between one or more datasets, plus the type of statistics describing within/between subspace relation, determines how $\mathbf{P}$ is set and, thus, whether and how the model simplifies to the classical special cases [1]. Our focus will be on MDM models driven by *statistical independence* among subspaces and *dependence* within subspaces, namely MISA, in the case of an overdetermined system with $V_m \geq C_m$, *without* implying $\mathbf{W}$ is square via the typical principal component analysis (PCA). Table II summarizes our key notations.

[1]The subspace terminology stems from [17] in which the columns of $\mathbf{A}$ corresponding to $\mathbf{y}_k$ form a linear (sub)space.

TABLE II
Key Notations

M, m	Number of datasets/modalities, counter
N, n	Number of observations, counter
K, k	Number of subspaces, counter
C, c	Number of sources, counter
V_m	intrinsic dimension of m-th dataset
d_k	dimension of k-th subspace
d_{mk}	dimension of k-th subspace in m-th dataset
$\mathbf{X}, \mathbf{x}$	Data matrix, and vector
$\mathbf{Y}, \mathbf{y}, y$	Source matrix, vector, and element
$\mathbf{y}_m$	Sources from m-th dataset
y_{mc}	c-th source in m-th dataset
$\mathbf{y}_k$	k-th subspace
$\mathbf{A}, \mathbf{A}_m$	mixing matrix, and its m-th block
$\mathbf{W}, \mathbf{W}_m$	unmixing matrix, and its m-th block
$\mathbf{P}, \mathbf{P}_k$	full and k-th subset assignment matrices
$\boldsymbol{\Sigma}_k^{\mathbf{y}}, \mathbf{R}_k^{\mathbf{y}}$	covariance and correlation matrices of $\mathbf{y}_k$
Patterns for Special Cases	
SDU:	$M = 1,\ K = C,\ d_k = 1\ \forall k$
SDM:	$M = 1,\ K < C,\ 1 \leq d_k < C\ \forall k$
MDU:	$M > 1,\ K = C_m\ \forall m,\ d_{mk} = 1\ \forall k, m$

In multimodal brain imaging research, various types of data can be utilized. MRI scans (e.g., structural, diffusion, functional, etc.) typically consist of 3D images, sometimes with an extra dimension. EEGs record the temporal evolution of scalp electric potentials, typically dozens of electrodes at the same time. After collecting two or more such modalities on the same subject, the information is often summarized to a single 3D image and/or time series for each modality. These summary features are obtained from multiple subjects and jointly analyzed with data fusion. Usually, only in-brain signal is considered from 3D images. Those in-brain voxels (volume pixels) are stacked into a single 1D vector prior to fusion. Other modalities, data preparation, and feature generation approaches exist but will not be discussed in this work.

III. Related Work

A. Applications

MDM problems permeate many fields and yet are largely undeveloped. In multimodal fusion of heterogeneous data [25], [26], robust identification of flexible joint features ($\mathbf{y}_k$) originating from *all* data modalities ($\mathbf{x}_m$) can yield one-of-a-kind views into a system's properties. This is a prominent direction in mental health research for biomarker identification and early diagnosis, with potential to convey new strategies for

disease severity assessment and translation into personalized treatments [1]. In classification, the association/dependence inherent to multimodal features $\mathbf{y}_k$ means that good separability in one dataset promotes features with similar property in other datasets, and vice-versa.

The benefits of model flexibility are also notable in various multiset analyses. In the case of multisubject unimodal data ($\mathbf{x}_m$) [27]–[32], it would better preserve subject specificity. In analyses that combine multi-site datasets ($\mathbf{x}_m$) from different scanners/devices, it could naturally mitigate harmonization issues [33], [34] since site/device-variability would seldom explain multidataset associations. In sensor fusion [5], [25], [35], [36], where noise characteristics can be similar if multiple sensors ($\mathbf{x}_m$) share the same environment, it would allow better detection (and potential removal) of noise. For hyperspectral imaging [37]–[39], hyperspectral features ($\mathbf{y}_k$) of higher complexity could be identified in time-lapse studies. For domain-adaptive image recognition [40]–[44], enhanced common and unique representations ($\mathbf{y}_k$) could be identified across image domains ($\mathbf{x}_m$). For multi-view image and video processing [9], [45], [46], objects with complex temporal patterns could be better characterized using (unimodal) higher-dimensional $\mathbf{y}_k$, not to mention potential fusion with audio features [47]–[50] via multimodal $\mathbf{y}_k$.

B. Methods

Our review of BSS in brain imaging [1] studied the underlying strategies of many methods. It offered a general, broad view of how different methods relate to each other by defining a common hierarchical taxonomy to accurately describe them. The unified framework introduced in that work provided a clear path for general MDM model development, which we adopted here to break from current MDU paradigms [15, Ch. 8]. However, it did not consider any of the issues addressed here, including combinatorial optimization, scale control, and non-orthogonal $\mathbf{W}_m$. These were also missing from our early investigations in [11]–[13]. Besides the vastly expanded methodology—which also introduces the general Kotz distribution for MISA—the current work presents a large number of new experiments and realistic applications.

Notably, the Kotz distribution was first introduced for BSS in [51] but applications were limited to MDU problems (IVA specifically). Consequently, that work cannot be applied to cases where $d_{mk} > 1$. In addition, its implementation treated the iteratively updated subspace covariances $\mathbf{\Sigma}_k^{\mathbf{y}}$ as constant with respect to $\mathbf{W}$ (previously, [30], [52] had hard-coded $\mathbf{\Sigma}_k^{\mathbf{y}} = \mathbf{I}$). The gradients derived for our MISA implementation do not make that assumption and, thus, yield a different search direction than [51] at each step during optimization, even for the IVA case. Also, the optimization approach in [51] was based on simple line search, which is rather different from the interior-point barrier optimization (with bounds and option for non-linear constraints) we utilize here. We also note the use of our novel scale control formulation for numerical stability.

Another work [53] also explores identification of subspace structures in the general MDM setting. However, it is limited to subspaces with Gaussian distribution and, thus, can only leverage SOS to identify subspaces. In contrast to our approach with the Kotz distribution, the approach in [53] cannot leverage HOS for subspace identification. Moreover, our option for the Kotz distribution implies that it suffices to set the parameters in (4) to $\boldsymbol{\psi}_G$ (Section IV-B) and our model simplifies to the same model in [53], highlighting the generality of MISA. The same argument applies to [23], [54].

Finally, premature convergence to local minima due to the mis-assignment of sources to subspaces is a known challenge for SDM model fitting [55]. However, general MDM problems have drastically more intricate within- and cross-dataset subspace-to-subspace interactions. When subspaces span multiple datasets, a combinatorially higher amount of possible local minima (upwards of $\prod_{k=1}^{K} \binom{\bar{C}-\sum_{l=0}^{k-1} d_l}{d_k} = 6 \cdot 10^{19}$ in Section V-B.4) undermines the numerical optimization performance (here, $d_0 \triangleq 0$). While combinatorial issues are common in other research areas [56]–[58], they have been largely neglected in BSS literature because of how simple (and often irrelevant) they are for ICA.

In Sections IV-D and IV-E we propose novel combinatorial optimization algorithms for evasion of local minima in the numerical optimization of (1). To the best of our knowledge, this is the first attempt at disentangling these permutation ambiguities in the general MDM case. In contrast to [59], our approach serves only to move a particular solution out of a local minima so that the numerical optimization may resume. Plus, the structural subspace priors contained in $\mathbf{P}_k$ guide our combinatorial procedures without relying on ancillary objective functions to determine residual source dependences.

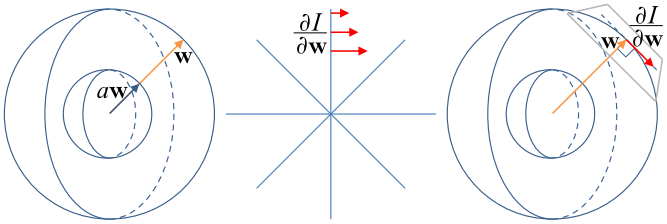
IV. Methodology

A. Scale Control

An inherent property of independence is invariance to arbitrary scaling of each or any source (i.e., multiplication by a non-zero scalar value), which is why ICA sources have *scale ambiguity*. This has an important implication on the geometry of the resulting objective function we seek to optimize. First, visualize the elements of $\mathbf{W}$ into a $\bar{D}$-dimensional vector ($\bar{D} = \bar{V}\bar{C}$) $\mathbf{w} = \mathrm{vec}(\mathbf{W})$ as would be done in a typical numerical optimization setting. Due to scale invariance, evaluation of the objective function on either $\mathbf{w}$ or $a\mathbf{w}$, where a is a non-zero scalar, yields the same value.

Since the objective function evaluates to the same values along the line[2] spanned by $\mathbf{w}$, only certain changes in the *direction* of $\mathbf{w}$ incur changes in the objective function. Consequently, it suffices to look for a solution on the surface of the *hypersphere* associated with a given a, since the landscape of objective function values would be identical across concentric (hyper) shells (Fig. 3 (a)). Moreover, scale invariance induces a "star" shape to the contour lines of the objective function in this scenario (Fig. 3 (b)). Since gradients are orthogonal to contour lines, they also ought to be orthogonal to $\mathbf{w}$ and lie on the *tangent* hyperplane of any given hypersphere (Fig. 3 (c)).

[2]Strictly speaking, this line is only a portion of the entire hyper surface (polyhedron) of ambiguity.



(a) Concentric hyperspheres (b) Contour lines (c) Tangent plane

Fig. 3. **Geometry of the independence-driven objective function in SDU problems**. (a) Due to the scale invariance property of statistical independence, evaluation of the objective function on either $\mathbf{w}$ or $a\mathbf{w}$, where a is a non-zero scalar, yields the same value. Only certain changes in the *direction* of $\mathbf{w}$ incur changes in the objective function. Thus, the solution space of independence-driven SDU problems lies on a hypersphere. (b) Scale invariance induces a "star" shape on the contour lines. (c) Consequently, the gradient of a scale invariant function must lie on the tangent hyperplane of the hypersphere associated with a given $\mathbf{w}$.

The main implication is that stepping in the (negative) direction of the gradient towards a local minimum will likely *inflate* $\mathbf{w}$ and lead the search direction in an outward spiral with respect to $\mathbf{w}$. This can be a problem if the norm of $\mathbf{w}$ grows indefinitely and eventually becomes numerically unstable. More importantly, as the norm of $\mathbf{w}$ increases toward outer shells, the landscape of the objective function starts to stretch (because its values are kept the same while the surface area of the hypersphere grows). Consequently, the *gradient grows shorter regardless of its proximity to any local minimum*. The smaller gradient will then lead to shorter step lengths, likely yielding very little improvement at latter stages of the numerical optimization and deterring convergence.

This issue is often disregarded in the literature (incidentally, the Infomax algorithm [19] is free of this issue) and should be addressed *prior* to evaluation of the efficient relative gradient [2, Ch. 4]. One simple approach to address it is to constrain the norm of $\mathbf{w}$. While direct, implementing this approach can be quite inefficient. Rather, since any scale is equally acceptable (at least in theory), we propose to control the estimated source scales by fixing them in the *model*. Specifically, this is accomplished by assigning the estimated subspace correlation matrix $\mathbf{R}_k^{\mathbf{y}}$ as the model dispersion matrix $\mathbf{D}_k$ in the Kotz distribution, effectively making the objective function *scale selective* rather than scale invariant (Section IV-B). Therefore, whenever the source estimates from the data do not support the model variances associated with this choice of $\mathbf{D}_k = \mathbf{R}_k^{\mathbf{y}}$, the mismatch induces changes in $\mathbf{W}$ that lead their variances towards the prescribed ones. In summary, the proposed scale selective formulation eliminates scaling issues without the need for a formal constraint.

B. Objective Function

Equation (1) admits some simplifications following a few manipulations. First, we note that $h(\mathbf{y}) = h(\mathbf{W}\mathbf{x}) = h(\mathbf{x}) + \ln|\det(\mathbf{W})|$, and $h(\mathbf{x})$ can be discarded since it is constant with respect to $\mathbf{W}$. Second, $\ln|\det(\mathbf{W})| = \sum_{m=1}^{M} \ln|\det(\mathbf{W}_m)|$ since $\mathbf{W}$ is block diagonal. Finally, when $V_m \neq C_m$, for any m, the determinant of $\mathbf{W}_m$ is undefined. In order to circumvent this issue, we propose to substitute the determinant by the product of the singular values of $\mathbf{W}_m$, i.e., $\prod_{i=1}^{C_m} \sigma_{mi}$, where σ_{mi} are the diagonal elements of $\mathbf{\Lambda}_m = \mathbf{U}_m^\top \mathbf{W}_m \mathbf{V}_m$ originating from the singular value decomposition $\mathbf{W}_m = \mathbf{U}_m \mathbf{\Lambda}_m \mathbf{V}_m^\top$. We note that $|\det \mathbf{W}| = \prod_{i=1}^{C} |\sigma_{mi}|$ when $\mathbf{W}$ is non-singular and square. Altogether, we can recast (1) as:

$$\check{I}(\mathbf{y}) = -\sum_{m=1}^{M} J_{D_m} - \sum_{k=1}^{K} \mathbb{E}\left[\ln p(\mathbf{y}_k)\right], \tag{3}$$

where $J_{D_m} = \sum_{i=1}^{C_m} \ln|\sigma_{mi}|$, and $\mathbf{y}_k = \mathbf{P}_k \mathbf{W}\mathbf{x}$.

This formulation is still incomplete because $p(\mathbf{y}_k)$ is undefined. Here we choose to model each subspace pdf as a multivariate Kotz distribution [14], [60]:

$$p(\mathbf{y}_k) = \frac{\beta_k \lambda_k^{\nu_k} \Gamma\left(\frac{d_k}{2}\right) \left(\mathbf{y}_k^\top \mathbf{D}_k^{-1} \mathbf{y}_k\right)^{\eta_k - 1}}{\pi^{\frac{d_k}{2}} \left(\det \mathbf{D}_k\right)^{\frac{1}{2}} \Gamma\left(\nu_k\right)} e^{-\lambda_k \left(\mathbf{y}_k^\top \mathbf{D}_k^{-1} \mathbf{y}_k\right)^{\beta_k}} \tag{4}$$

where d_k is the subspace dimensionality, $\beta_k > 0$ controls the shape of the pdf, $\lambda_k > 0$ the kurtosis (i.e., the degree of peakedness), and $\eta_k > \frac{2-d_k}{2}$ the hole size, while $\nu_k \triangleq \frac{2\eta_k + d_k - 2}{2\beta_k} > 0$ and $\alpha_k \triangleq \frac{\Gamma\left(\nu_k + \beta_k^{-1}\right)}{\lambda_k^{\beta_k^{-1}} d_k \Gamma\left(\nu_k\right)}$ for brevity. $\Gamma(\cdot)$ denotes the gamma function. The positive definite dispersion matrix $\mathbf{D}_k$ is related to the covariance matrix $\mathbf{\Sigma}_k^{\mathbf{y}}$ by $\mathbf{D}_k = \alpha_k^{-1} \mathbf{\Sigma}_k^{\mathbf{y}}$.

This is a good choice of pdf since it includes the multivariate power exponential family, particularly the classical multivariate Gaussian and multivariate Laplace distributions when the parameter set $\boldsymbol{\psi}_k = [\beta_k, \lambda_k, \eta_k]$ is set to $\boldsymbol{\psi}_G = [1, \frac{1}{2}, 1]$ and $\boldsymbol{\psi}_L = [\frac{1}{2}, 1, 1]$, respectively.

Minimizing (3) is equivalent to maximizing the (log-) likelihood of $\mathbf{y}_k$. In the following, we estimate $\mathbf{\Sigma}_k^{\mathbf{y}}$ from the data. This is appealing because the sample average $\bar{\mathbf{\Sigma}}^{\mathbf{x}}$ is readily available and can be conveniently combined with $\mathbf{W}$ to produce an approximation of $\mathbf{\Sigma}_k^{\mathbf{y}}$ for substitution in $\mathbf{D}_k$. This simple choice permits the reparameterization of $\mathbf{\Sigma}_k^{\mathbf{y}}$ as a function of $\mathbf{W}$, specifically $\bar{\mathbf{\Sigma}}_k^{\mathbf{y}} = \frac{1}{N-1} \mathbf{P}_k \mathbf{W}\mathbf{X}\mathbf{X}^\top \mathbf{W}^\top \mathbf{P}_k^\top$.

Two well-conceived dispersion matrix parameter choices are proposed for the Kotz distribution, one emphasizing invariance to source scales and the other not, resulting in two useful objective functions. Firstly, we let $\mathbf{Y}_k = \mathbf{P}_k \mathbf{W}\mathbf{X}$ and use n to index each of the N observations used in the sample mean approximation of the expected value $\mathbb{E}[\cdot]$ in (3). Secondly, based on the log-likelihood $\ln p(\mathbf{y}_k)$, we define $J_{C_k} = \ln\det \mathbf{D}_k$, $J_{F_k} = \ln\left(\mathbf{y}_k^\top \mathbf{D}_k^{-1} \mathbf{y}_k\right)$, and $J_{E_k} = \left(\mathbf{y}_k^\top \mathbf{D}_k^{-1} \mathbf{y}_k\right)^{\beta_k}$. Then, we let $\mathbf{D}_k = \alpha_k^{-1} \bar{\mathbf{\Sigma}}_k^{\mathbf{y}}$ for the standard *scale invariant* case:

$$\begin{aligned}\check{I}(\mathbf{y}) = &-\sum_{m=1}^{M} J_{D_m} + \frac{1}{2}\sum_{k=1}^{K} J_{C_k} - f(K, \beta_k, \lambda_k, \eta_k, d_k, \nu_k) \\ &- \sum_{k=1}^{K} \frac{\eta_k - 1}{N} \sum_{n=1}^{N} J_{F_{kn}} + \sum_{k=1}^{K} \frac{\lambda_k}{N} \sum_{n=1}^{N} J_{E_{kn}},\end{aligned} \tag{5}$$

where

$$f(K,\beta_k,\lambda_k,\eta_k,d_k,\nu_k)=\sum_{k=1}^{K}\left[\ln\beta_k+\nu_k\ln\lambda_k+\ln\Gamma\left(\frac{d_k}{2}\right)-\frac{d_k}{2}\ln\pi-\ln\Gamma\left(\nu_k\right)\right],$$

with gradient given by:

$$\nabla\check{I}(\mathbf{W})_{mi_k}=\left[B_k+\left[I-B_k\mathbf{Y}_k^\top\right]A_k\right]\mathbf{X}_m^\top-\left(\mathbf{W}_m^-\right)^\top \quad (6)$$
$$\forall k\in\{1,\ldots,K\},\forall m\in\{1,\ldots,M\}$$

where i_k represents all source indices (rows of $\nabla\check{I}(\mathbf{W})_m$) assigned to subspace k, $\circ$ is the Hadamard product, and

$$\begin{aligned}A_k&=\left[\bar{\mathbf{\Sigma}}_k^{\mathbf{y}}\right]^{-1}\mathbf{Y}_k\\B_k&=A_k\mathrm{diag}(\mathbf{t}_k)\\\mathbf{t}_k&=\left(\frac{2\beta_k\lambda_k}{N}\mathbf{z}_k^{\beta_k}+\frac{2\left(1-\eta_k\right)}{N}\right)\circ\mathbf{z}_k^{-1}\\\mathbf{z}_k&=[z_{k1},z_{kn},\ldots,z_{kN}]\\z_{kn}&=\mathbf{y}_{kn}^\top\left[\alpha_k^{-1}\mathbf{\Sigma}_k^{\mathbf{y}}\right]^{-1}\mathbf{y}_{kn}.\end{aligned}$$

For the scale-controlled approach, we let $\mathbf{D}_k=\mathbf{R}_k^{\mathbf{y}}$, and the correlation matrix $\mathbf{R}_k^{\mathbf{y}}\triangleq\boldsymbol{\gamma}_k^\top\mathbf{\Sigma}_k^{\mathbf{y}}\boldsymbol{\gamma}_k$, and $\boldsymbol{\gamma}_k\triangleq\left(\mathbf{I}_{d_k}\circ\mathbf{\Sigma}_k^{\mathbf{y}}\right)^{-\frac{1}{2}}$. In this case, only correlations are estimated from the data, while variances are fixed at α_k. The advantage of this choice is that it controls the scale of the sources rather than letting them be arbitrarily large/small.

In the *scale-controlled* case, $\check{I}(\mathbf{y})$ is identical to (5), except $J_{C_k}=\ln\det\left(\boldsymbol{\gamma}_k\mathbf{\Sigma}_k^{\mathbf{y}}\boldsymbol{\gamma}_k^\top\right)$, $J_{F_{kn}}=\ln(\mathbf{y}_{kn}^\top\left[\boldsymbol{\gamma}_k\mathbf{\Sigma}_k^{\mathbf{y}}\boldsymbol{\gamma}_k^\top\right]^{-1}\mathbf{y}_{kn})$ and $J_{E_{kn}}=\left(\mathbf{y}_{kn}^\top\left[\boldsymbol{\gamma}_k\mathbf{\Sigma}_k^{\mathbf{y}}\boldsymbol{\gamma}_k^\top\right]^{-1}\mathbf{y}_{kn}\right)^{\beta_k}$, with gradient:

$$\nabla\check{I}(\mathbf{W})_{mi_k}=\left[\bar{\boldsymbol{\gamma}}_k^{-1}B_k+\left[\bar{\boldsymbol{\gamma}}_kG_k-B_kA_k^\top+\left[\mathbf{Z}_{\mathbf{\Sigma}}^{-1}-\bar{\boldsymbol{\gamma}}_k^2\right]\right]\mathbf{Y}_k\right]\mathbf{X}_m^\top-\left(\mathbf{W}_m^-\right)^\top \quad (7)$$
$$\forall k\in\{1,\ldots,K\},\forall m\in\{1,\ldots,M\}$$

where

$$\begin{aligned}\bar{\boldsymbol{\gamma}}_k&=\left(\mathbf{I}\circ\mathbf{Z}_{\mathbf{\Sigma}}\right)^{-\frac{1}{2}}\\\mathbf{Z}_{\mathbf{\Sigma}}&=\mathbf{P}_k\mathbf{W}\mathbf{X}\mathbf{X}^\top\mathbf{W}^\top\mathbf{P}_k^\top\\G_k&=\mathbf{I}\circ\left(B_k\mathbf{Y}_k^\top\right)\\B_k&=A_k\mathrm{diag}(\mathbf{t}_k)\\A_k&=\mathbf{Z}_{\mathbf{\Sigma}}{}^{-1}\bar{\boldsymbol{\gamma}}_k^{-1}\mathbf{Y}_k\\\mathbf{t}_k&=\left(\frac{2\beta_k\lambda_k}{N}\mathbf{z}_k^{\beta_k}+\frac{2\left(1-\eta_k\right)}{N}\right)\circ\mathbf{z}_k^{-1}\\\mathbf{z}_k&=[z_{k1},\ldots,z_{kn},\ldots,z_{kN}]\\z_{kn}&=\mathbf{y}_{kn}^\top\left[\boldsymbol{\gamma}_k\bar{\mathbf{\Sigma}}_k^{\mathbf{y}}\boldsymbol{\gamma}_k^\top\right]^{-1}\mathbf{y}_{kn}.\end{aligned}$$

While the equations presented above are general and support any choice of subspace specific parameters $\boldsymbol{\psi}_k$, in the examples presented here, we opted to use the same set $\boldsymbol{\psi}_k=\boldsymbol{\psi}_L$ for all subspaces, modeling subspaces as multivariate Laplace distributions *with* correlation estimation. The derivation of the gradients can be found in *supplemental material* along with a description of the relative gradient update $\nabla\check{I}(\mathbf{W})\mathbf{W}^\top\mathbf{W}$ [2, Ch. 4] [61] we used together with the L-BFGS algorithm with bounds (L-BFGS-B) [62], [63] available in the non-linear constraint optimization function fmincon of MATLAB's Optimization Toolbox. Nonlinear constraints such as those shown next can be easily incorporated in fmincon's interior-point barrier method [64, Ch. 19] [65].

C. Pseudoinverse Reconstruction Error

In the overdetermined case, i.e., when $V_m>C_m$ and $\mathbf{W}$ is wide, it is necessary to constrain $\mathbf{W}$ in order to evade ill-conditioned solutions. The error incurred by $\mathbf{W}$ in reconstructing the data samples can indirectly guide and constrain $\mathbf{W}$. The mean squared error (MSE) between $\mathbf{x}$ and $\hat{\mathbf{x}}$ gives the following formulation of the reconstruction error (RE):

$$E=\mathbb{E}\left[\|\hat{\mathbf{x}}-\mathbf{x}\|_2^2\right]\approx\frac{1}{N}\sum_{n=1}^{n=N}\|\hat{\mathbf{x}}_n-\mathbf{x}_n\|_2^2. \quad (8)$$

Firstly, the optimal *linear* estimator of $\mathbf{x}$ based on $\mathbf{y}$ for a system with estimation error $\mathbf{e}$', such as $\mathbf{y}=\mathbf{W}\mathbf{x}+\mathbf{e}$', is $\hat{\mathbf{A}}\mathbf{y}$, where $\hat{\mathbf{A}}$ is the minimizer of MSE:

$$\hat{\mathbf{A}}=\mathbf{\Sigma}^{\mathbf{x}}\mathbf{W}^\top\left(\mathbf{W}\mathbf{\Sigma}^{\mathbf{x}}\mathbf{W}^\top+\mathbf{\Sigma}^{\mathbf{e}'}\right)^{-1}, \quad (9)$$

and $\mathbf{\Sigma}^{\mathbf{x}}$ is the data covariance. In the high SNR regime, $\mathrm{diag}\left(\mathbf{W}\mathbf{\Sigma}^{\mathbf{x}}\mathbf{W}^\top\right)\gg\mathrm{diag}\left(\mathbf{\Sigma}^{\mathbf{e}'}\right)$ element-wise and, as discussed in [66], yields

$$\hat{\mathbf{A}}=\mathbf{\Sigma}^{\mathbf{x}}\mathbf{W}^\top\left(\mathbf{W}\mathbf{\Sigma}^{\mathbf{x}}\mathbf{W}^\top\right)^{-1}=\mathbf{\Sigma}^{\mathbf{x}}\mathbf{W}^\top\mathbf{\Sigma}^{\mathbf{y}\,-1}. \quad (10)$$

This choice of $\hat{\mathbf{A}}$ *always* minimizes the error no matter how far $\mathbf{W}$ is from the true $\mathbf{W}_\star$ and serves little as a constraint.

Assuming unit source variances and data whitened such that $\mathbf{\Sigma}^{\mathbf{x}}=\mathbb{E}\left[\mathbf{x}\mathbf{x}^\top\right]=\mathbf{I}$, in ICA problems $\mathbf{W}_\star$ must be *row* orthonormal, i.e., $\mathbf{W}_\star\mathbf{W}_\star^\top=\mathbf{I}$. Our previous work [12] utilized $\hat{\mathbf{A}}=\mathbf{W}^\top$ to reconstruct $\mathbf{x}$ as $\hat{\mathbf{x}}=\mathbf{W}^\top\mathbf{W}\mathbf{x}$ instead. Under the whitening assumption, this can be implemented in (8) as a soft regularizer provably equivalent to regularization by either the Frobenius norm $\left\|\mathbf{W}^\top\mathbf{W}-\mathbf{I}\right\|_{\mathrm{F}}^2$ or $\left\|\mathbf{W}\mathbf{W}^\top-\mathbf{I}\right\|_{\mathrm{F}}^2$, when the regularizer constant approaches infinity [67]. Therefore, this approach effectively penalizes non-orthogonal $\mathbf{W}$.

Here, our investigation of the singular value decomposition (SVD) of $\mathbf{W}$ reveals that, if the matrix has orthonormal rows, then its singular values are all 1 and $\mathbf{W}=USV^\top=UV^\top$, where $S=\mathbf{I}$, U are the left singular vectors of $\mathbf{W}$, and V its right singular vectors. Therefore, $\mathbf{W}^\top\mathbf{W}=VU^\top UV^\top=VV^\top$. Since $\mathbf{W}$ is wide, V is tall, which implies $VV^\top\neq\mathbf{I}$, in general. Thus, using $\hat{\mathbf{x}}_n=\mathbf{W}^\top\mathbf{W}\mathbf{x}_n$, the RE simplifies as:

$$E_\top\approx\frac{1}{N}\sum_{n=1}^{n=N}\left\|(VV^\top-\mathbf{I})\mathbf{x}_n\right\|_2^2. \quad (11)$$

This clearly shows that RE with $\hat{\mathbf{A}}=\mathbf{W}^\top$ implicitly acts as a constraint on the right singular vectors of $\mathbf{W}$, selecting those whose outer product approximates the identity matrix $\mathbf{I}$.

If not orthonormal, $\mathbf{W}^\top\mathbf{W}=VS^2V^\top$ since $S\neq\mathbf{I}$. Thus, we propose to use the pseudoinverse $\mathbf{W}^-=\mathbf{W}^\top(\mathbf{W}\mathbf{W}^\top)^{-1}$ in lieu of $\mathbf{W}^\top$, with $\hat{\mathbf{x}}_n=\mathbf{W}^-\mathbf{W}\mathbf{x}_n$. Then, this pseudoinverse RE (PRE) (E_-) also simplifies as (11). This result

follows from the SVD of the pseudoinverse $\mathbf{W}^- = VS^{-1}U^\top$ and $\mathbf{W}^-\mathbf{W} = VS^{-1}U^\top USV^\top = VV^\top$. Unlike before, this formulation effectively constrains V in the general case. Note that since $\mathbf{\Sigma^x} = \mathbf{I}$ in the case of white data, the optimal estimator (10) simplifies to $\hat{\mathbf{A}} = \mathbf{W}^\top \left(\mathbf{W}\mathbf{W}^\top\right)^{-1} = \mathbf{W}^-$, i.e., the *pseudoinverse gives the least error when the data is white* (if the SNR is high), regardless of the values contained in $\mathbf{W}$. Thus, for white data, we conclude that the RE formulation ($E_\top$) is more appropriate than PRE (E_-). Our experience, however, suggests that $\mathbf{W}$ is far more likely non-orthogonal in real noisy, non-white data, justifying our preference for E_-.

Furthermore, we introduce a normalization term, dividing E_- by $\mathbf{x}_{\text{norm}}$, the average power in the data, and we get *the proportion of power missed*:

$$E \approx \frac{1}{\mathbf{x}_{\text{norm}}} \frac{1}{N} \sum_{n=1}^{n=N} \left\| \mathbf{W}^\top \left(\mathbf{W}\mathbf{W}^\top\right)^{-1} \mathbf{W}\mathbf{x}_n - \mathbf{x}_n \right\|_2^2, \quad (12)$$

where $\mathbf{x}_{\text{norm}} \approx \frac{1}{N} \sum_{n=1}^{n=N} \|\mathbf{x}_n\|_2^2$. Its gradient has the form:

$$\nabla E(\mathbf{W}) = C - C\mathbf{W}^-\mathbf{W} \quad (13)$$

where

$$C = \frac{2}{\mathbf{x}_{\text{norm}}\ N} \left[\mathbf{W}^-\right]^\top B$$
$$B = \mathbf{X}Z^\top + Z\mathbf{X}^\top$$
$$Z = \mathbf{W}^-\mathbf{W}\mathbf{X} - \mathbf{X}.$$

Since $\mathbf{X}$ and $\mathbf{W}$ are block-diagonal, these operations can be computed separately on each dataset by replacing $\mathbf{X}$ with $\mathbf{X}_m$ and $\mathbf{W}$ with $\mathbf{W}_m$. This can be used both as a data reduction approach or a nonlinear constraint for optimization.

Finally, in MDU problems, when there is prior knowledge supporting *linear* dependence (i.e., correlation) within subspaces, then one useful and popular approach is to use *group* PCA projection to initialize all blocks of $\mathbf{W}$ [68]. It works by performing a single data reduction step on datasets concatenated along the V dimension. We have investigated this approach in a separate work [69], offering efficient algorithms to enable this procedure when the number of datasets is very large ($M > 10000$). For comparison purposes, we also considered the use of group PCA (gPCA) as an alternate initialization approach for $\mathbf{W}$ in our experiments.

D. MISA With Greedy Permutations (SDM Case)

We present a greedy optimization approach to counter local minima resulting from arbitrary source permutations. To illustrate, consider a single dataset and assume $\mathbf{P}_k$ is a user-specified prior. Using abbreviated notation throughout, suppose $\mathbf{P}_1 = [1\ 1\ 1\ 0\ 0]$ and $\mathbf{P}_2 = [0\ 0\ 0\ 1\ 1]$ define a partitioning of five sources into two subspaces: $p(\mathbf{y}) = p(\mathbf{y}_{k=1})p(\mathbf{y}_{k=2}) = p(y_1, y_2, y_3)p(y_4, y_5)$, where $p(\cdot)$ is a joint pdf. It would be equally acceptable if the data supported either $p(\mathbf{y}) = p(y_4, y_5)p(y_1, y_2, y_3)$ (entire subspace permutation) or $p(\mathbf{y}) = p(y_1, y_3, y_2)p(y_5, y_4)$ (within-subspace permutation) or even some combination of these two cases. However, if the data supported $p(\mathbf{y}) = p(y_{\mathbf{1}}, y_4)p(y_2, y_3, y_{\mathbf{5}})$, then that would not be equivalently acceptable, denoting a local minimum.

Algorithm 1 Greedy Permutations GP

Require: dataset $\mathbf{X} \in \mathbb{R}^{V \times N}$, subspace assignment matrix $\mathbf{P} \in \{0,1\}^{K \times C}$, unmixing matrix $\mathbf{W} \in \mathbb{R}^{C \times V}$

$K, C = \dim(\mathbf{P})$
for $c = 1$ **to** C **do** ▷ loop over sources
  $\mathsf{kurrent} = \mathsf{find}(\mathbf{P}[:, c])$ ▷ index of current subspace
  $p = \mathsf{find}(\mathbf{P}[\mathsf{kurrent}, :])$ ▷ source indices
  $\mathbf{P}[:, p] = 0$ ▷ erase source assignments to subspace
  $\mathsf{vals} = [\ ]$ ▷ cost values array
  for $k = 1$ **to** $K + 1$ **do** ▷ loop over subspaces
    if $k > 1$ **then**
      $\mathbf{P}[k-1, p] = 0$ ▷ undo previous assignment
    end if
    $\mathbf{P}[k, p] = 1$ ▷ assign sources to subspace k
    $\mathbf{P}_{nu} = \mathsf{remove\_empty\_rows}(\mathbf{P})$
    $\mathsf{vals}[k] = \mathsf{cost}(\mathbf{X}, \mathbf{P}_{nu}, \mathbf{W}, \mathsf{scale\_control} = \mathsf{False})$ ▷ evaluate Equation (5)
  end for
  $\mathbf{P}[k, p] = 0$ ▷ undo previous assignment
  $k = \text{argmin}(\mathsf{vals})$ ▷ assignment with lowest cost
  if $k \neq \mathsf{kurrent}$ **and** $|\mathsf{vals}[k] - \mathsf{vals}[\mathsf{kurrent}]| < \sqrt{\mathsf{eps}}$ **then**
    $k = \mathsf{kurrent}$ ▷ ignore tiny change improvements
  end if
  $\mathbf{P}[k, p] = 1$
  $\mathbf{P} = \mathsf{remove\_empty\_rows}(\mathbf{P})$
end for
return $\mathbf{P}$

When these occur, the numerical optimization in Section IV-B stops early, at the newly found local minimum. At that point, we propose to check whether another permutation of sources would attain a lower objective value. This entails two challenges: 1) given the combinatorial nature of the task, even mild numbers of sources lead to huge numbers of candidate permutations, and 2) when the optimization stops early, most sources are still mixed and there is not enough *refinement* to establish which sources are dependent and belong in the same subspace. The low refinement precludes the combinatorial problem since it hinders the ability to distinguish between dependent and independent sources in the first place.

Firstly, therefore, we propose to transform the single-dataset multidimensional (SDM) ISA task into single-dataset unidimensional (SDU) ICA. We do that by temporarily voiding and replacing subspaces of size $d_k \geq 2$ by multiple sources (each with $d_k = 1$), and then restarting the numerical optimization from the current $\mathbf{W}$ estimate (local minimum). This pushes all sources towards being independent from each other. However, dependent sources will only be *as independent as possible* and will retain some of their dependence. Partly motivated by [59], this approach secures enough refinement to distinguish among subspaces. Thus, given sources that are as independent as possible, we propose a greedy search for any residual dependence among them. The greedy solution is valid because the specific ordering within subspaces is irrelevant. Unlike [59], our approach does not require accessory objective functions to detect dependent sources. Instead, it uses the same scale invariant objective defined in (5).

Algorithm 2 MISA-GP for SDM Problems $\mathsf{MISA\text{-}GP_{SDM}}$

Require: dataset $\mathbf{X} \in \mathbb{R}^{V\times N}$, user-defined (UD) subspace assignment matrix $\mathbf{P}_{\mathrm{UD}} \in \{0,1\}^{K\times C}$, initial unmixing matrix $\mathbf{W}_0 \in \mathbb{R}^{C\times V}$, maximum number of greedy iterations T

$\mathbf{W} = \mathsf{MISA}(\mathbf{X}, \mathbf{P}_{\mathrm{UD}}, \mathbf{W}_0, \mathsf{scale\_control} = \mathsf{True})$
$\mathsf{vals}[0] = \mathsf{cost}(\mathbf{X}, \mathbf{P}_{\mathrm{UD}}, \mathbf{W}, \mathsf{scale\_control} = \mathsf{True})$
$\mathbf{W}_{\mathrm{opt}}[0] = \mathbf{W}$; $t = 1$; $\mathsf{vals}[t] = \mathrm{Inf}$
while $t \leq T$ **and** $\mathsf{vals}[t] \neq \mathsf{vals}[t-1]$ **do**
  $\mathbf{P} = \mathbf{I}$ ▷ switch to SDU model
  $\mathbf{W}_{\mathrm{SDU}} = \mathsf{MISA}(\mathbf{X}, \mathbf{P}, \mathbf{W}, \mathsf{scale\_control} = \mathsf{True})$ ▷ solve SDU
  $\mathbf{P} = \mathsf{GP}(\mathbf{X}, \mathbf{P}, \mathbf{W}_{\mathrm{SDU}})$ ▷ Algorithm 1
  $\mathsf{ix} = \mathsf{match}(\mathbf{P}, \mathbf{P}_{\mathrm{UD}})$ ▷ find source ordering best ▷ matching prescribed $\mathbf{P}_{\mathrm{UD}}$
  $\mathbf{W} = \mathbf{W}[\mathsf{ix}, :]$ ▷ reorder sources (escape local min)
  $\mathbf{W} = \mathsf{MISA}(\mathbf{X}, \mathbf{P}_{\mathrm{UD}}, \mathbf{W}, \mathsf{scale\_control} = \mathsf{True})$ ▷ restart SDM
  $\mathsf{vals}[t] = \mathsf{cost}(\mathbf{X}, \mathbf{P}_{\mathrm{UD}}, \mathbf{W}, \mathsf{scale\_control} = \mathsf{True})$
  $\mathbf{W}_{\mathrm{opt}}[t] = \mathbf{W}$
  $t = t + 1$
end while
$t = \mathrm{argmin}(\mathsf{vals})$ ▷ retrieve best solution
return $\mathbf{W}_{\mathrm{opt}}[t]$

Algorithm 3 MISA-GP for MDM Problems MISA-GP

Require: dataset $\mathbf{X} = \{\mathbf{X}_m \in \mathbb{R}^{V_m\times N} : m \in M\}$, user-defined subspace assignment matrices $\mathbf{P}_{\mathrm{UD}} = \{\mathbf{P}_{\mathrm{UD},m} \in \{0,1\}^{K\times C_m} : m \in M\}$, initial unmixing matrix $\mathbf{W}_0 = \{\mathbf{W}_{0,m} \in \mathbb{R}^{C_m\times V_m} : m \in M\}$, maximum number of greedy iterations T

$\mathbf{W} = \mathsf{MISA}(\mathbf{X}, \mathbf{P}_{\mathrm{UD}}, \mathbf{W}_0, \mathsf{scale\_control} = \mathsf{True})$
$\mathsf{vals}[0] = \mathsf{cost}(\mathbf{X}, \mathbf{P}_{\mathrm{UD}}, \mathbf{W}, \mathsf{scale\_control} = \mathsf{True})$
$\mathbf{W}_{\mathrm{opt}}[0] = \mathbf{W}$; $t = 1$; $\mathsf{vals}[t] = \mathrm{Inf}$
while $t \leq T$ **and** $\mathsf{vals}[t] \neq \mathsf{vals}[t-1]$ **do**
  for $m = 1$ **to** M **do** ▷ For each dataset
    $\mathbf{P}_m = \mathbf{I}$ ▷ switch to SDU model
    $\mathbf{W}_{\mathrm{SDU}} = \mathsf{MISA}(\mathbf{X}_m, \mathbf{P}_m, \mathbf{W}_m, \mathsf{scale\_control} = \mathsf{True})$ ▷ solve SDU
    $\mathbf{P}_m = \mathsf{GP}(\mathbf{X}_m, \mathbf{P}_m, \mathbf{W}_{\mathrm{SDU}})$ ▷ Algorithm 1
    $\mathsf{ix} = \mathsf{match}(\mathbf{P}_m, \mathbf{P}_{\mathrm{UD},m})$ ▷ find source ordering ▷ best matching prescribed $\mathbf{P}_{\mathrm{UD},m}$
    $\mathbf{W}_m = \mathbf{W}_m[\mathsf{ix}, :]$ ▷ reorder sources (escape local min)
  end for
  $\mathbf{W} = \mathsf{subspace\_perm}(\mathbf{X}, \mathbf{P}_{\mathrm{UD}}, \mathbf{W}, \mathsf{scale\_control} = \mathsf{False})$
  $\mathbf{W} = \mathsf{MISA}(\mathbf{X}, \mathbf{P}_{\mathrm{UD}}, \mathbf{W}, \mathsf{scale\_control} = \mathsf{True})$ ▷ restart MDM
  $\mathsf{vals}[t] = \mathsf{cost}(\mathbf{X}, \mathbf{P}_{\mathrm{UD}}, \mathbf{W}, \mathsf{scale\_control} = \mathsf{False})$
  $\mathbf{W}_{\mathrm{opt}}[t] = \mathbf{W}$
  $t = t + 1$
end while
$t = \mathrm{argmin}(\mathsf{vals})$ ▷ retrieve best solution
return $\mathbf{W}_{\mathrm{opt}}[t]$

The procedure is 1) switch to the ICA model (effectively, make $\mathbf{P} = \mathbf{I}$), 2) numerically optimize it, 3) reassign sources into subspaces one at a time. In the latter, as indicated in Algorithm 1 (GP), each source is assigned sequentially to each subspace (if two or more are assigned to the same subspace, they are reassigned together thereafter). Thus, the model changes with every assignment, and simple evaluation of the objective $\mathsf{cost}(\cdot)$ (without numerical optimization) produces a value for each particular assignment. *The scale invariant formulation ensures source variances do not influence the estimation*. The assignment minimizing the objective function determines to which subspace a source belongs. Here, assume that $k = K+1$ inserts one more row in $\mathbf{P}$ for a new subspace; $[:, p]$ are the contents of columns indexed by p (conversely for rows); $\mathsf{find}(\cdot)$ recovers the indexes of all non-zero elements; $\mathsf{remove\_empty\_rows}(\mathbf{P})$ removes rows from $\mathbf{P}$ containing only zero entries; eps is the machine's precision.

After repeating this procedure for all sources, in an attempt to solve the original model, we order the identified subspaces so as to match the original prescribed subspace structure $\mathbf{P}$ as closely as possible. This final sorting ($\mathsf{match}(\cdot)$) defines a specific permutation of the sources, which we then use to reorder the rows of the local minimum solution $\mathbf{W}$ for the original ISA problem, effectively moving that solution out of the local minimum. After that, we resume the numerical optimization of the original ISA problem until another minimum is found. In our experiments, repeating this procedure just twice in a row ($T = 2$) and taking the best out of three solutions sufficed to drastically improve results. In Algorithm 2 ($\mathsf{MISA\text{-}GP_{SDM}}$), $\mathsf{MISA}(\cdot)$ represents the numerical optimization (Section IV-B).

A direct benefit of this approach is that more dependence tends to be retained within subspaces as compared to [59]. That is a desirable property because it leaves room for further post-processing and investigation. Another advantage of our approach is that it can match source assignments to user-prescribed subspace priors ($\mathbf{P}$) when they are available.

E. MISA With Greedy Permutations (MDM Case)

The previous approach addresses cross-subspace interference issues due to incorrect allocation of the *sources* and, therefore is appropriate for SDM problems. However, it is not sufficient to perform such procedure in MDM problems since ambiguities may also occur at the *subspace* level, i.e., incorrect allocation of the dataset-specific subspaces.

Consider the following example for a model with three subspaces spanning two datasets, each dataset containing five sources. Assume the correct assignment of sources is as follows: $p_1(y_{11}, y_{21}, y_{22})p_2(y_{12}, y_{13}, y_{23})p_3(y_{14}, y_{15}, y_{24}, y_{25})$, where the notation y_{mi} refers to source i from dataset m, and $p_k(\cdot)$ is the joint pdf of subspace k. Since $\mathsf{MISA\text{-}GP_{SDM}}$ is designed for single datasets, at best, it produces $p_1(y_{11})p_2(y_{12}, y_{13})p_3(y_{14}, y_{15})$ for $m = 1$ and $p_1(y_{21}, y_{22})p_2(y_{23})p_3(y_{24}, y_{25})$ for $m = 2$. Then, from a global perspective, these solutions would yield the correct subspace assignment above, thus solving the MDM problem. However, it is equally acceptable for SDM solvers to produce either $p_1(y_{11})p_2(y_{1\mathbf{4}}, y_{1\mathbf{5}})p_3(y_{1\mathbf{2}}, y_{1\mathbf{3}})$ for $m = 1$ or $p_1(y_{2\mathbf{4}}, y_{2\mathbf{5}})p_2(y_{23})p_3(y_{2\mathbf{1}}, y_{2\mathbf{2}})$ for $m = 2$ if the datasets are evaluated separately (notice the bold subscripts). Together they imply $p_1(y_{11}, y_{2\mathbf{4}}, y_{2\mathbf{5}})p_2(y_{1\mathbf{4}}, y_{1\mathbf{5}}, y_{23})p_3(y_{1\mathbf{2}}, y_{1\mathbf{3}}, y_{2\mathbf{1}}, y_{2\mathbf{2}})$, which does not match the correct assignment and, thus, fails to produce a solution for the MDM problem. What we have illustrated here is that within-dataset permutations of *equal-sized* subspaces may induce mismatches across datasets if the datasets are processed separately. Another complicating factor are subspaces absent from a particular dataset.

Borrowing from the ideas in Section (IV-D), we propose three approaches to address these issues. The first, extends the greedy search to *all* datasets by sequentially assigning each source (in every dataset) to every subspace and accepting the assignments that reduce the objective function. This would yield a complexity of at least $O(\bar{C}K)$, and $O(\bar{C}^2)$ in the (unlikely) worst case of $K = \bar{C}$. The second, processes each dataset separately (as in the previous example) and then applies the same greedy strategy at the level of subspaces instead. Effectively, this approach cycles through each subspace sequentially, trying to determine which of them can be combined to form a larger subspace. This yields a complexity of $O(C_m K M) + O(K^2 M)$. The final approach is to test all possible permutations of subspaces with the same size, after processing each dataset separately, which yields $O((K!)^M)$. While this can quickly become computationally prohibitive, it can also identify better solutions since it evaluates all subspace permutations of interest. In this work, we elected to use the third approach when the number of sources is small and the second when that number becomes larger (subspace_perm$(\cdot)$). Full procedures are indicated in Algorithm 3 (MISA-GP).

V. Results

We present results on multiple experiments satisfying the requisites outlined in [22], including a summary of various controlled simulations on carefully crafted synthetic data, as well as hybrid data and comparisons with several algorithms.

A. General Simulation Setup and Evaluation

In the following, we consider the problem of identifying statistically independent subspaces. Thus, in all experiments, each subspace $\mathbf{y}_k$ is a random sample with N observations from Laplace distribution. Subspace observations are linearly mixed via a random $\mathbf{A}$ as $\mathbf{x} = \mathbf{A}\mathbf{y} + \mathbf{e}$, where $\mathbf{e}$ is additive sensor white noise. $\mathbf{A}$ is generated from a standard Gaussian distribution. Its singular values are then adjusted to yield the condition number cond$(\mathbf{A})$ prescribed in Table III. Also, the white Gaussian noise $\mathbf{e}$ (zero mean and unit variance) is multiplied by a scalar value in order to attain the SNR prescribed in Table III. The SNR is the power ratio between the noisy signal $\mathbf{x}$ and the noise $\mathbf{e}$. The equality $\mathrm{SNR} = 10^{\frac{\mathrm{SNRdB}}{10}}$ permits decibel (dB) specifications.

The quality of results is evaluated using the normalized multidataset Moreau-Amari intersymbol interference (MISI) (14), which extends the ISI [70], [71] to multiple datasets.

$$\mathrm{MISI}(\mathbf{H}) = \frac{0.5}{K(K-1)} \left[\sum_{i=1}^{K} \left(-1 + \sum_{j=1}^{K} \frac{|h_{ij}|}{\max\limits_{k} |h_{ik}|} \right) + \sum_{j=1}^{K} \left(-1 + \sum_{i=1}^{K} \frac{|h_{ij}|}{\max\limits_{k} |h_{kj}|} \right) \right] \quad (14)$$

where $\mathbf{H}$ is a matrix with elements $h_{ij} = \mathbf{1}^\top \left| \mathbf{P}_i \hat{\mathbf{W}} \mathbf{A} \mathbf{P}_j^\top \right| \mathbf{1}$, with $(i, j) = 1 \ldots K$, i.e., the sum of absolute values from all elements of the interference matrix $\hat{\mathbf{W}}\mathbf{A}$ corresponding to subspaces i and j, and $\hat{\mathbf{W}}$ is the solution being evaluated.

TABLE III

Summary of Simulation Results. $(\mathbf{a}, \mathbf{b})$ Median (Over 10 Dataset Instances) of Best MISI (Over 10 Initializations Per Dataset). $(\mathbf{c}, \mathbf{d})$ Median MISI (Over 10 Initializations, 1 Dataset Instance)

SNRdB		30	10	3	0.46	0.0043
ICA1	PRE+Infomax	0.0222	0.0292	0.0685	**0.0928**	0.2576
	PRE+MISA	**0.0145**	**0.0165**	**0.0261**	0.1932	0.2743
IVA2	PRE+IVA-L	0.0111	0.0136	0.0197	0.0277	0.5158
	PRE+MISA	0.0059	0.0088	0.0113	0.0151	0.0338
	gPCA+IVA-L	0.0081	0.0090	0.0095	0.0094	0.1271
	gPCA+MISA	**0.0044**	**0.0045**	**0.0049**	**0.0065**	**0.0205**
ISA3	PRE+JBD-SOS	0.2700	0.2804	0.2996	0.3255	0.3712
	PRE+MISA	0.1153	0.1275	0.1320	0.1495	0.3404
	PRE+MISA-GP	**0.0366**	**0.0670**	**0.0794**	**0.1140**	**0.3404**

(a) Varying SNRdB, Fixed cond$(\mathbf{A}) = 7$

cond$(\mathbf{A})$		1	3	7	15
ICA1	PRE+Infomax	**0.0804**	0.0188	0.0216	0.0493
	PRE+MISA	0.1934	**0.0148**	**0.0161**	**0.0267**
IVA2	PRE+IVA-L	0.1923	0.1013	0.0749	0.0505
	PRE+MISA	**0.0052**	**0.0045**	**0.0049**	0.0078
	gPCA+IVA-L	0.0090	0.0086	0.0092	0.0095
	gPCA+MISA	**0.0052**	**0.0045**	**0.0049**	**0.0067**
ISA3	PRE+JBD-SOS	0.2905	0.2792	0.2815	0.2962
	PRE+MISA	0.1008	0.1065	0.1202	0.1351
	PRE+MISA-GP	**0.0395**	**0.0330**	**0.0612**	**0.0743**

(b) Fixed SNRdB $= 3$, Varying cond$(\mathbf{A})$

$\rho_{k,max}$	0	0.1	0.23	0.39	0.5	0.65
IVA-GL	0.4767	0.0361	0.0114	0.0199	0.0184	0.0186
MISA	**0.0273**	**0.0098**	**0.0072**	**0.0062**	**0.0061**	**0.0049**

(c) IVA1: Increasing max. subspace correlation $\rho_{k,max}$

	ISA1 ($\rho_k = 0$)		ISA2 ($\rho_k > 0.2$)	
	$d_k = k$	$d_k = 4$	$d_k = k$	$d_k = 4$
EST_ISA	–	.0.7557	–	0.7766
JBD-SOS	0.2600	0.3496	0.2826	0.3739
MISA	**0.0239**	**0.0162**	**0.0369**	**0.0326**

(d) Varying vs Fixed subspace dimensionality d_k

For fairness, all algorithms are initialized with the same $\mathbf{W}_0$. See optimization parameters in supplemental material.

B. Summary of Synthetic Data Simulations

The performance of MISA in a series of synthetic data experiments with different properties is summarized below (Table III). Complete details are available as supplemental material online.

1) ICA 1 ($\bar{V} > N$): effects of additive noise (a) and condition number (b) are assessed in a moderately large ICA problem ($\bar{C} = 75$, $M = 1$) with rectangular mixing matrix $\mathbf{A}$ ($\bar{V} = 8000$) at a fairly small sample size regime ($N = 3500$). Under low SNR (b), MISA outperforms Infomax when $\mathrm{cond}\,(\mathbf{A}) \neq 1$. At high SNR (a), MISA outperforms Infomax more often than not.

2) IVA 1 ($V_m < N$, $V_m = C_m$): MISA performance is assessed in an IVA problem (c), in which subspaces span all of $M = 10$ datasets. Specifically, we study the case when no data reduction is required (i.e., $V_m = C_m = 16$), noise is absent, and observations are abundant ($N = 32968$). The striking feature observed here is that the performance of IVA-GL [21]

is much more variable than that from MISA, especially with high correlation within the subspaces. MISA performs well even at low within-subspace correlation levels and is highly stable when these correlations are larger than 0.2.

3) IVA 2 ($V_m < N$): Effects of additive noise (a) and condition number (b) are assessed in a larger IVA problem ($C_m = 75$, $M = 16$) with rectangular mixing matrix $\mathbf{A}$ ($V_m = 250$) and an abundant number of observations $N = 32968$. Data reduction with either group PCA (gPCA) or pseudoinverse RE (PRE) produced equivalent results in this large N scenario. Under low SNR, increasing the condition number had a fairly small detrimental effect on the performance of both IVA-L [18] and MISA. More importantly, while both IVA-L and MISA performed very well at mild-to-high SNR levels, the performance of MISA on extremely noisy scenarios ($\mathrm{SNRdB} = 0.0043$) is remarkable ($0.1 < \mathrm{MISI} < 0.01$), irrespective of using PRE or gPCA.

4) ISA 1 and 2 ($\bar{V} < N$, $\bar{V} = \bar{C}$): MISA performance is assessed in ISA problems (d), in which subspaces are multidimensional, with $M = 1$. Specifically, we study the case when no data reduction is required (i.e., $\bar{V} = \bar{C} = 28$), noise is absent, and the number of observations N is abundant. Fixed and varying configurations of $K = 7$ subspaces are considered, at two subspace correlation ρ_k settings. The striking feature observed here is that the performance of both JBD-SOS [23] and EST_ISA [24] is very poor in all cases, even when within-subspace correlations are present. MISA-GP is the only method with good performance, highlighting the large benefit of our approach for evasion of local minima.

5) ISA 3 ($\bar{V} > N$): Effects of additive noise (a) and condition number (b) are assessed in a mildly large ISA problem ($\bar{C} = 51$, $M = 1$) with variable subspace dimensionalities d_k, rectangular mixing matrix $\mathbf{A}$ ($\bar{V} = 8000$) at a fairly small sample size regime ($N = 5250$). Under a challenging SNR, JBD-SOS and MISA fail in virtually all cases ($\mathrm{MISI} > 0.1$). Inclusion of combinatorial optimization enables MISA-GP to perform quite well at mild-to-high SNR levels ($\mathrm{SNRdB} \geq 3$).

Execution times for Table III (a-b) are reported in Table IV. The timings were recorded on a Linux server (Ubuntu 16.04) with an Intel Xeon E5-2630v4 (10-core, 20-thread, 3.5GHz) CPU, 256GB RAM (DDR4, 2.4GHz). The code was executed in native Matlab without any optimizations.

The timings are higher in Table IV (a) than in Table IV (b) for PRE-based ICA1 and IVA2 experiments. This is consistent with a corresponding MISI reduction, which was due to a less strict stopping condition for the PRE gradient norm. This suggests that allowing more noise to leak from the PRE step not only yields poorer MISI performance but also significantly slows down convergence (about 3-4 times slower than comparable experiments in Table IV (b)).

In ICA1, Infomax is 1-2 orders of magnitude faster, owing to its inherently different stochastic optimization strategy and gradient implementation, which is optimized for a single dataset. The difference, however, is not due to a difference in algorithmic complexity. Importantly, Infomax is limited and cannot generalize beyond SDU problems like MISA.

In IVA2, MISA takes at least twice as long to converge than IVA-L but attains better results in terms of MISI. Note that the maximum number of iterations in IVA-L was set to four times the total number of iterations until convergence for MISA on the same problem, from the same starting point.

In ISA3, MISA-GP timings are comparable to those of JBD-SOS. However, MISA-GP attains about one order of magnitude better results in terms of MISI.

Overall, the reported timings support that the computational cost of MISA is tractable, especially given it enables universal application to different problems.

TABLE IV

TIMING SUMMARY. $(\mathbf{a}, \mathbf{b})$ MEAN (OVER 10 DATASET INSTANCES) OF MEDIAN TIME (OVER 10 INITIALIZATIONS PER DATASET). TIMES ARE REPORTED IN SECONDS

SNRdB		30	10	3	0.46	0.0043
ICA1	PRE+Infomax	22.6	20.5	4.4	3.3	2.7
	PRE+MISA	743.3	1392.3	2397.5	1543.3	52.0
IVA2	PRE+IVA-L	3128.7	3230.1	3094.7	3794.6	3375.4
	PRE+MISA	11645.4	11087.1	9391.6	11323.9	8134.5
	gPCA+IVA-L	2352.3	2170.1	2080.5	2369.5	3167.1
	gPCA+MISA	4610.2	4432.4	4408.5	5068.8	6354.5
ISA3	PRE+JBD SOS	2965.3	3007.2	3088.1	3016.6	2934.2
	PRE+MISA	222.8	503.5	729.4	897.6	655.8
	PRE+MISA+GP	2811.4	3479.7	3998.7	4290.0	1429.1

(a) Varying SNRdB, Fixed $\mathrm{cond}(\mathbf{A}) = 7$

$\mathrm{cond}(\mathbf{A})$		1	3	7	15
ICA1	PRE+Infomax	4.2	25.1	25.1	7.2
	PRE+MISA	1582.8	391.9	740.1	1889.0
IVA2	PRE+IVA-L	2120.3	1784.0	1929.4	2376.6
	PRE+MISA	4112.9	3895.4	4238.5	6205.9
	gPCA+IVA-L	2803.7	2393.1	2318.0	2548.4
	gPCA+MISA	5689.4	4635.0	4544.7	5267.6
ISA3	PRE+JBD SOS	2999.9	2935.2	2934.6	3021.1
	PRE+MISA	1145.7	483.6	780.3	1012.5
	PRE+MISA+GP	3307.3	1354.9	3413.6	3893.3

(b) Fixed $\mathrm{SNRdB} = 3$, Varying $\mathrm{cond}(\mathbf{A})$

C. Hybrid Data Experiments

We present three major results on novel applications of BSS to brain image analysis, open sourcing realistic hybrid data standards (https://github.com/rsilva8/MISA) that test estimation limits at small sample size. The first pushes the conditions of experiment ICA 1 and emulates a single-subject temporal ICA of functional MRI (fMRI). The second investigates the use of IVA with $V_m > N$ for multimodal fusion of brain MRI-derived data. Finally, the last experiment evaluates the value of MDM models without data reduction for fusion of functional MRI (fMRI) and EEG neural signals.

Given the real features from prior publications utilized here, our experiments indeed reflect the usual size of fMRI, sMRI, and EEG datasets in neuroimaging multimodal fusion. Typically, studies combine 2-4 modalities (here, 2-3) with intrinsic dimensionality V_m of 15k-300k voxels, and 600 timepoints. The last example also illustrates how MISA can recover sources even without data reduction of the V_m dimension. Moreover, we illustrate source estimation with 600-1000 subjects, which is 3-10 times bigger than typical multimodal fusion datasets. Furthermore, the typical number of sources in multimodal fusion ranges from 4 to 30 (our experiments are 4 to 20). Lastly, to the best of our knowledge, no other

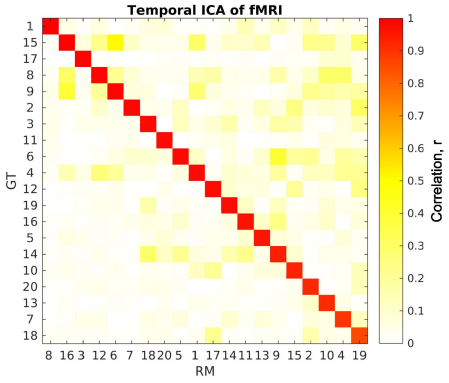


Fig. 4. **Correlation with the ground-truth (hybrid temporal ICA)**. The correlation between the spatial map estimates from MISA with PRE (RM) and the ground-truth (GT) is very high with little residual similarity across sources, suggesting the analysis was successful.

work has attempted general subspace estimation ($d_{mk} > 1$) in multimodal fusion, which is feasible with MISA, as we demonstrate in our last experiment.

1) Single-Subject Temporal ICA of fMRI: Here we consider temporal ICA of fast acquisition fMRI. The dimensionality of the data is $\bar{V} = \text{voxels} \approx 60\text{k}$ and $N = \text{time points} \approx 1300$. In order to better assess the performance of MISA in a realistic scenario, we propose to set the mixing matrix $\mathbf{A}$ as the *real* part of the data. First, we let $\bar{C} = 20$ sources. Then, $\mathbf{A}$ must be a $60\text{k} \times 20$ matrix. In order to have it correspond to real data, we assign to it the first twenty well-established aggregate spatial maps (3D volumes) published in [72].

For the synthetic part of the data, we propose to simulate a 20×1334 matrix of timecourses $\mathbf{y}$ by generating realistic autocorrelated samples that mimic observed fMRI timecourses to a good extent. Sampling 20 such timecourses that retain independence with respect to each other is challenging because independently sampled autocorrelated time series tend to be correlated with one another. Building on the simulation principles outlined in [22], we seek to avoid randomly correlated timecourses (sources) in order to prevent mismatches to the underlying ICA model we wish to test. In the same spirit, we also wish to have sources sampled from the same distribution used in the model, here a Laplace distribution. We developed the following steps in order to meet all these requirements:

1) Design a joint *autocorrelation* matrix $\mathbf{R}^{\mathbf{yy}}$ for all sources. For the example above, this means a $\bar{C}N \times \bar{C}N$ block-diagonal correlation matrix ($\bar{C}N = 26680$) with $\bar{C}$ blocks of size $N \times N$. Each block is designed with an exponentially decaying autocorrelation function with an autocorrelation around 0.85 between time point n and $n-1$, and around 0.2 between n and $n-10$. This structure retains autocorrelation within each N-long section of an observation while retaining uncorrelation/independence among sections.
2) Generate 50k $\bar{C}N$-dimensional observations using a Gaussian copula [73] and the autocorrelation matrix $\mathbf{R}^{\mathbf{yy}}$ from step 1. Using copulas enables transformation of the marginal distributions while retaining their correlation/dependence.

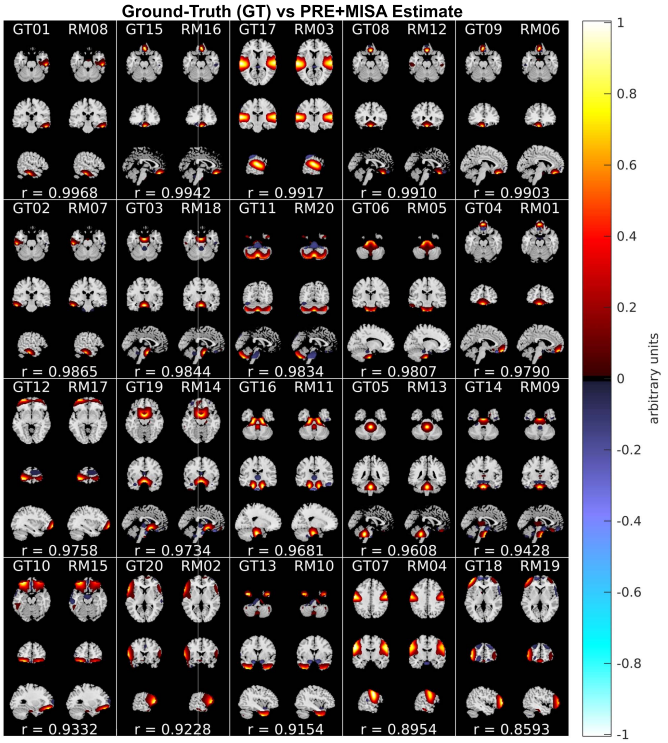


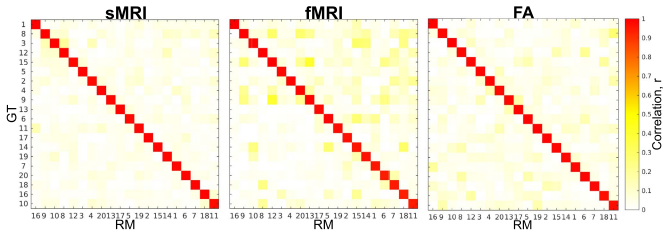
Fig. 5. **Side-by-side comparison with the ground-truth (hybrid temporal ICA)**. The clear resemblance to the ground-truth maps suggests a successful recovery of the mixing matrix $\mathbf{A}$. The sample correlation r is shown below each matched pair. Maps are sorted from highest to lowest correlation.

3) For each of the 50k copula-sampled observations, transform the sample into a Laplace distribution.
4) For each of the 50k transformed $\bar{C}N$-dimensional observations, reshape them into a $\bar{C} \times N$ matrix and compute the resulting $\bar{C} \times \bar{C}$ $\mathbf{R}^{\mathbf{y}}$ correlation matrix.
5) Compute the median correlation matrix $\mathbf{R}^{\mathbf{y}}_{med}$ over the 50k observed $\mathbf{R}^{\mathbf{y}}$.
6) Retain the transformed observation whose $\mathbf{R}^{\mathbf{y}}$ is closest to $\mathbf{R}^{\mathbf{y}}_{med}$ and reject the rest.

This type of rejection sampling effectively produces the desired outcome. Finally, Gaussian noise is added to the mixture for a *low* $\text{SNRdB} = 3$. The condition number of $\mathbf{A}$ was 4.59.

In the results, the data was reduced using PRE and then processed with MISA to obtain independent *timecourses*. The correlation between ground-truth (GT) and PRE+MISA spatial map estimates (RM) is presented in Fig. 4, and the spatial maps (estimating $\mathbf{A}$ from $\hat{\mathbf{W}}^{-}$) in Fig. 5. $\text{MISI} = 0.0365$.

2) Multimodal IVA of sMRI, fMRI, and FA: In this multimodal fusion of structural MRI (sMRI), fMRI, and Fractional Anisotropy (FA) diffusion MRI data, the dimensionalities are $V_1 = \text{voxels} \approx 300\text{k}$, $V_2 = \text{voxels} \approx 67\text{k}$, $V_3 = \text{voxels} \approx 15\text{k}$, respectively, and $N = \text{subjects} = 600$ (each modality measured on the same subject). We pursue a hybrid setting where only the mixing matrices $\mathbf{A}_m$ are taken from real datasets to overcome typically small N in patient population studies. First, we let $C_m = 20$ sources in each dataset. Then, $\mathbf{A}_1$, $\mathbf{A}_2$, and $\mathbf{A}_3$ must be $300\text{k} \times 20$, $67\text{k} \times 20$, and $15\text{k} \times 20$,



(a) Structural MRI (b) Functional MRI (c) Fractional Anisotropy

Fig. 6. **Correlation with the ground-truth (multimodal IVA).** The correlation between the spatial map estimates from MISA with PRE (RM) and the ground-truth (GT) is very high in all modalities, with little residual similarity across sources, suggesting the analysis was successful.

respectively. To each, we assign the first twenty aggregate 3D spatial maps published in [74], [72], [75], respectively.

For the simulated part of the data, we generate three 20×600 matrices of subject expression levels $\mathbf{y}$. $K = 20$ subspaces, each with $d_k = 3$ and $N = 600$ observations, were sampled independently from a Gaussian copula, using an inverse exponential autocorrelation function with maximal correlation varying from 0.65 to 0.85 for each subspace. These were transformed to Laplace distribution marginals (*not* multivariate Laplace) so as to induce a *controlled mismatch* between the data (only SOS dependence) and the model subspace distributions (multivariate Laplace—all-order dependence). Finally, Gaussian noise was added separately in each dataset for a low $\mathrm{SNRdB} = 3$. The condition numbers of $\mathbf{A}_1$, $\mathbf{A}_2$, and $\mathbf{A}_3$ were 1.52, 4.59, 1.63, respectively.

In the results, the data was reduced using PRE and then processed with MISA to obtain independent *subject expression levels*. Per-modality correlation between ground-truth and PRE+MISA spatial maps are presented in Fig. 6, and spatial maps (estimating $\mathbf{A}$ from $\hat{\mathbf{W}}^{-}$) in Fig. 7. $\mathrm{MISI} = 0.0273$.

3) Multimodal MISA of fMRI, and EEG: We show the value of MDM models without data reduction for fusion of EEG event-related potentials (ERP) and fMRI datasets, with dimensionality $V_1 = \text{time points} \approx 600$, $V_2 = \text{voxels} \approx 67\text{k}$, respectively, and $N = \text{subjects} = 1001$. Let $C_1 = 4$ and $C_2 = 6$ sources in the ERP and fMRI datasets, respectively, organized into $K = 4$ subspaces (y_{mi} represents source i from dataset m):

$k = 1$: IVA-type, sources y_{11} and y_{21} ($d_k = 2$);
$k = 2$: MISA-type, sources y_{12}, y_{22} and y_{23} ($d_k = 3$);
$k = 3$: MISA-type, sources y_{13}, y_{14} and y_{24} ($d_k = 3$);
$k = 4$: ISA-type, sources y_{25} and y_{26} ($d_k = 2$).

Utilizing real spatial maps and timecourses, $\mathbf{A}_1$ and $\mathbf{A}_2$ must be 600×4 and $67\text{k} \times 6$, respectively, this time ensuring they form column-orthogonal mixings (with Gram-Schmidt).

For the simulated part of the data, we generate 4×1001 and 6×1001 matrices of subject expression levels for ERP and fMRI datasets, respectively. A total of $K = 4$ d_k-dimensional subspaces with $N = 600$ observations each were sampled from a multivariate Laplace distribution, using an inverse exponential autocorrelation function with maximal correlation of 0.65 for each subspace. Noise was absent in both datasets. The condition number was 1.00 for both $\mathbf{A}_1$ and $\mathbf{A}_2$.

Fig. 8 shows the results obtained from constrained MISA-GP, i.e., with $\hat{\mathbf{A}} = \mathbf{W}^{\top}$ RE constraint using (8). No

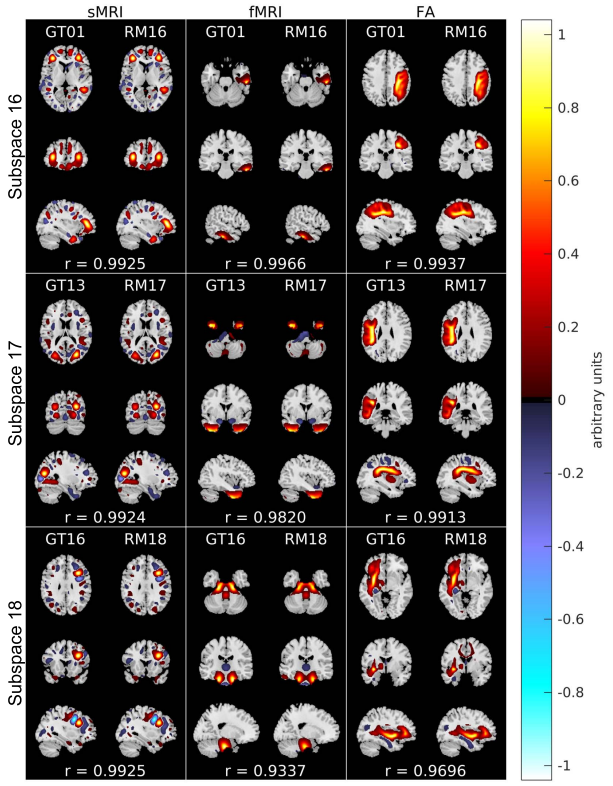


Fig. 7. **Summary of multimodal IVA maps**. In each panel, ground-truth (GT) maps are presented on the left and maps estimated from MISA with PRE (RM) on the right. Each subspace represents the multimodal set of maps (joint features) with highest, median, and minimum correlation with the GT, from top to bottom, respectively. *No IVA-L comparison available since it failed to converge, likely due to the small sample size ($N = 600$) or inability to detect SOS dependence.*

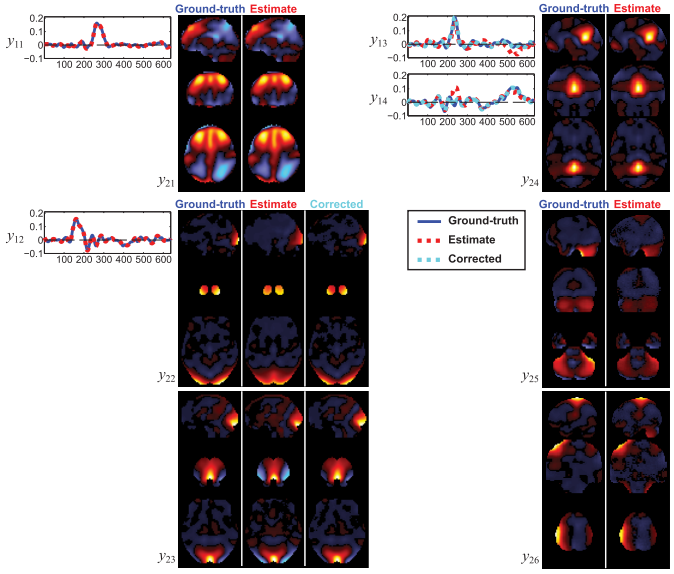


Fig. 8. **Multimodal MISA of fMRI and ERP**. GT maps are presented on the left of each panel, MISA-GP estimates in the middle, and corrected MISA-GP estimates on the right. GT ERPs are presented in blue, MISA-GP ERPs in dashed red, and corrected MISA-GP ERPs in dashed cyan.

data reduction was performed on the data. The spatial fMRI maps and ERP timecourses were produced by estimating $\mathbf{A}$ from $\hat{\mathbf{W}}^{\top}$. Since subspace independence is invariant to linear transformations (arbitrary basis) within any subspace [17], the estimation yields timecourses (red) and maps (middle) that do

not match the GT exactly. In an attempt to correct for that, we performed additional within-modality ICAs on the columns of $\mathbf{A}_m$ corresponding to subspaces. This effectively selected for a particular basis within each subspace (right maps and cyan timecourses). The ability to choose a particular representation demonstrates the kinds of post-processing enabled by MDM models. Overall, this result validates and illustrates the benefit of a constrained optimization approach.

VI. Conclusion

We have presented MISA, an approach that solves multiple BSS problems (including ICA, IVA, ISA, and more) under the same framework, with remarkable performance and improved robustness even at low SNR. In particular, we have derived a general formulation that controls for source scales, leveraging the flexible Kotz distribution in an interior point non-linear constraint optimization, with PRE as a general and flexible formulation for either direct subspace estimation or dimensionality reduction, in conjunction with combinatorial optimization for evasion of local minima, permitting self-correction to the closest subspace structures supported by the data (MISA-GP). Altogether, the proposed methods permit all-order statistics linkage across multidatasets as well as features of higher complexity to be identified and fully exploited in a direct, principled, and synergistic way, even at sample sizes as low as $N = 600$.

Flexible approaches like MISA are key to meet the growing complexity of multidataset tasks. These complexities are incorporated in the hybrid dataset standards we open source here, built from relevant results published in the brain imaging BSS literature. Generalizations building on this work could be easily developed exploring other divergence families. Future work will focus on compiling real multimodal datasets to validate MISA's ability to capture reliable modes of shared and unique variability across and within modalities.

It is also worth noting the natural trade-off that exists between flexibility and complexity. In practice, given some problem specification and prior information, a dedicated algorithm offers the simplest solution. However, the lack of flexibility therein often limits its utility to explore different scenarios. Our work considers a more general case, where one general solution is easily simplified by taking the domain information into account for a given problem. The complexity is unchanged in comparison to a dedicated algorithm. But the general algorithm makes it very easy to switch between models and explore different solutions.

We suggest that further optimization for computational efficiency is certainly possible.

References

[1] R. F. Silva, S. M. Plis, J. Sui, M. S. Pattichis, T. Adalı, and V. D. Calhoun, "Blind source separation for unimodal and multimodal brain networks: A unifying framework for subspace modeling," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 7, pp. 1134–1149, Oct. 2016.

[2] P. Comon and C. Jutten, *Handbook of Blind Source Separation*, 1st ed. Oxford, U.K.: Academic Press, 2010.

[3] L. Yi *et al.*, "Chemometric methods in data processing of mass spectrometry-based metabolomics: A review," *Anal. Chim. Acta*, vol. 914, pp. 17–34, Mar. 2016.

[4] S. Saito, K. Oishi, and T. Furukawa, "Convolutive blind source separation using an iterative least-squares algorithm for non-orthogonal approximate joint diagonalization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2434–2448, Dec. 2015.

[5] A. A. Nielsen, "Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data," *IEEE Trans. Image Process.*, vol. 11, no. 3, pp. 293–305, Mar. 2002.

[6] R. Ammanouil, A. Ferrari, C. Richard, and D. Mary, "Blind and fully constrained unmixing of hyperspectral images," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5510–5518, Dec. 2014.

[7] V. D. Calhoun, J. Liu, and T. Adalı, "A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data," *NeuroImage*, vol. 45(Suppl. 1), no. 1, pp. S163–S172, Mar. 2009.

[8] V. D. Calhoun and J. Sui, "Multimodal fusion of brain imaging data: A key to finding the missing link(s) in complex mental illness," *Biol. Psychiatry: Cognit. Neurosci. Neuroimag.*, vol. 1, no. 3, pp. 230–244, May 2016.

[9] S. Bhinge, Y. Levin-Schwartz, and T. Adalı, "Data-driven fusion of multi-camera video sequences: Application to abandoned object detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 1697–1701.

[10] M. A. Nicolaou, V. Pavlovic, and M. Pantic, "Dynamic probabilistic CCA for analysis of affective behavior and fusion of continuous annotations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1299–1311, Jul. 2014.

[11] R. F. Silva, S. M. Plis, T. Adalı, and V. D. Calhoun, "Multidataset independent subspace analysis," in *Proc OHBM*, Hamburg, Germany, 2014, p. 3506.

[12] R. F. Silva, S. M. Plis, T. Adalı, and V. D. Calhoun, "Multidataset independent subspace analysis extends independent vector analysis," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Paris, France, Oct. 2014, pp. 2864–2868.

[13] R. F. Silva, S. M. Plis, M. S. Pattichis, T. Adalı, and V. D. Calhoun, "Incorporating second-order statistics in multidataset independent subspace analysis," in *Proc OHBM*, Honolulu, HI, USA, 2015, p. 3743.

[14] S. Kotz, "Multivariate distributions at a cross road," in *A Modern Course on Statistical Distributions in Scientific Work*. Calgary, AB, Canada: Springer, 1974, pp. 247–270.

[15] R. F. Silva and S. M. Plis, *How to Integrate Data From Multiple Biological Layers in Mental Health?*. Cham, Switzerland: Springer, 2019, pp. 135–159.

[16] P. Comon, "Independent component analysis, a new concept?" *Signal Process.*, vol. 36, no. 3, pp. 287–314, Apr. 1994.

[17] J.-F. Cardoso, "Multidimensional independent component analysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, vol. 4, Seattle, WA, USA, May 1998, pp. 1941–1944.

[18] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *Proc ICA*, vol. 3889. Charleston, SC, USA, 2006, pp. 165–172.

[19] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, no. 6, pp. 1129–1159, Nov. 1995.

[20] S.-I. Amari, "Natural gradient works efficiently in learning," *Neural Comput.*, vol. 10, no. 2, pp. 251–276, Feb. 1998.

[21] M. Anderson, T. Adalı, and X. L. Li, "Joint blind source separation with multivariate Gaussian model: Algorithms and performance analysis," *IEEE Trans. Signal Process.*, vol. 60, no. 4, pp. 1672–1683, Apr. 2012.

[22] R. F. Silva, S. M. Plis, T. Adalı, and V. D. Calhoun, "A statistically motivated framework for simulation of stochastic data fusion models applied to multimodal neuroimaging," *NeuroImage*, vol. 102, Part 1, pp. 92–117, Nov. 2014.

[23] D. Lahat, J.-F. Cardoso, and H. Messer, "Second-order multidimensional ICA: Performance analysis," *IEEE Trans. Signal Process.*, vol. 60, no. 9, pp. 4598–4610, Sep. 2012.

[24] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 3361–3368.

[25] D. Lahat, T. Adalı, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proc. IEEE*, vol. 103, no. 9, pp. 1449–1477, Sep. 2015.

[26] K. L. Miller *et al.*, "Multimodal population brain imaging in the UK biobank prospective epidemiological study," *Nature Neurosci.*, vol. 19, no. 11, pp. 1523–1536, Nov. 2016.

[27] V. D. Calhoun and T. Adalı, "Multisubject independent component analysis of fMRI: A decade of intrinsic networks, default mode, and neurodiagnostic discovery," *IEEE Rev. Biomed. Eng.*, vol. 5, pp. 60–73, 2012.

[28] A.-K. Seghouane and A. Iqbal, "Sequential dictionary learning from correlated data: Application to fMRI data analysis," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 3002–3015, Jun. 2017.

[29] A.-R. Mohammadi-Nejad, G.-A. Hossein-Zadeh, and H. Soltanian-Zadeh, "Structured and sparse canonical correlation analysis as a brain-wide multi-modal data fusion approach," *IEEE Trans. Med. Imag.*, vol. 36, no. 7, pp. 1438–1448, Jul. 2017.

[30] J.-H. Lee, T.-W. Lee, F. A. Jolesz, and S.-S. Yoo, "Independent vector analysis (IVA): Multivariate approach for fMRI group study," *NeuroImage*, vol. 40, no. 1, pp. 86–109, Mar. 2008.

[31] S. Bhinge, R. Mowakeaa, V. D. Calhoun, and T. Adalı, "Extraction of time-varying spatiotemporal networks using parameter-tuned constrained IVA," *IEEE Trans. Med. Imag.*, vol. 38, no. 7, pp. 1715–1725, Jul. 2019.

[32] M. Pakravan and M. B. Shamsollahi, "Extraction and automatic grouping of joint and individual sources in multisubject fMRI data using higher order cumulants," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 2, pp. 744–757, Mar. 2019.

[33] M. Yu *et al.*, "Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data," *Hum. Brain Mapping*, vol. 39, no. 11, pp. 4213–4227, 2018.

[34] H. Mirzaalian *et al.*, "Inter-site and inter-scanner diffusion MRI data harmonization," *NeuroImage*, vol. 135, pp. 311–323, Jul. 2016.

[35] F. Alam, R. Mehmood, I. Katib, N. N. Albogami, and A. Albeshri, "Data fusion and IoT for smart ubiquitous environments: A survey," *IEEE Access*, vol. 5, pp. 9533–9554, 2017.

[36] N. E. D. Elmadany, Y. He, and L. Guan, "Information fusion for human action recognition via Biset/Multiset globality locality preserving canonical correlation analysis," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5275–5287, Nov. 2018.

[37] M. Uzair, A. Mahmood, and A. Mian, "Hyperspectral face recognition with spatiospectral information fusion and PLS regression," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 1127–1137, Mar. 2015.

[38] J. Yao, D. Meng, Q. Zhao, W. Cao, and Z. Xu, "Nonconvex-sparsity and nonlocal-smoothness-based blind hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2991–3006, Jun. 2019.

[39] A. Villa, J. A. Benediktsson, J. Chanussot, and C. Jutten, "Hyperspectral image classification with independent component discriminant analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 12, pp. 4865–4876, Dec. 2011.

[40] H. Xu, J. Zheng, A. Alavi, and R. Chellappa, "Cross-domain visual recognition via domain adaptive dictionary learning," 2018, *arXiv:1804.04687*. [Online]. Available: http://arxiv.org/abs/1804.04687

[41] J. Fan *et al.*, "HD-MTL: Hierarchical deep multi-task learning for large-scale visual recognition," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1923–1938, Apr. 2017.

[42] H. Lu, C. Shen, Z. Cao, Y. Xiao, and A. van den Hengel, "An embarrassingly simple approach to visual domain adaptation," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3403–3417, Jul. 2018.

[43] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer joint matching for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1410–1417.

[44] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, "Visual domain adaptation: A survey of recent advances," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 53–69, May 2015.

[45] Z. Cai, L. Wang, X. Peng, and Y. Qiao, "Multi-view super vector for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 596–603.

[46] L. Tang, Z.-X. Yang, and K. Jia, "Canonical correlation analysis regularization: An effective deep multiview learning baseline for RGB-D object recognition," *IEEE Trans. Cognit. Develop. Syst.*, vol. 11, no. 1, pp. 107–118, Mar. 2019.

[47] M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp, "Audiovisual synchronization and fusion using canonical correlation analysis," *IEEE Trans. Multimedia*, vol. 9, no. 7, pp. 1396–1403, Nov. 2007.

[48] L. Gao, R. Zhang, L. Qi, E. Chen, and L. Guan, "The labeled multiple canonical correlation analysis for information fusion," *IEEE Trans. Multimedia*, vol. 21, no. 2, pp. 375–387, Feb. 2019.

[49] P. Narvor, B. Rivet, and C. Jutten, "Audiovisual speech separation based on independent vector analysis using a visual voice activity detector," in *Proc LVA/ICA*, Grenoble, France, 2017, pp. 247–257.

[50] F. Nesta, S. Mosayyebpour, Z. Koldovsky, and K. Palecek, "Audio/video supervised independent vector analysis through multimodal pilot dependent components," in *Proc. 25th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2017, pp. 1150–1164.

[51] M. Anderson, G.-S. Fu, R. Phlypo, and T. Adalı, "Independent vector analysis, the kotz distribution, and performance bounds," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 3243–3247.

[52] A. Hyvärinen and U. Köster, "FastISA: A fast fixed-point algorithm for independent subspace analysis," in *Proc ESANN*, 2006, pp. 371–376.

[53] D. Lahat and C. Jutten, "Joint independent subspace analysis using second-order statistics," *IEEE Trans. Signal Process.*, vol. 64, no. 18, pp. 4891–4904, Sep. 2016.

[54] M. Anderson, X.-L. Li, and T. Adalı, "Nonorthogonal independent vector analysis using multivariate Gaussian model," in *Proc LVA/ICA* (Lecture Notes in Computer Science). Paris, France: Springer, 2010, vol. 6365, pp. 354–361.

[55] A. Hyvärinen, J. Hurri, and P. Hoyer, *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision* (Computational Imaging and Vision). 1st ed. London, U.K.: Springer, 2009, vol. 39.

[56] T. Bouwmans, S. Javed, H. Zhang, Z. Lin, and R. Otazo, "On the applications of robust PCA in image and video processing," *Proc. IEEE*, vol. 106, no. 8, pp. 1427–1457, Aug. 2018.

[57] J. Pont-Tuset, P. Arbelaez, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping for image segmentation and object proposal generation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 128–140, Jan. 2017.

[58] N. Y. El-Zehiry and L. Grady, "Contrast driven elastica for image segmentation," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2508–2518, Jun. 2016.

[59] Z. Szabó, B. Póczos, and A. Lőrincz, "Separation theorem for independent subspace analysis and its consequences," *Pattern Recognit.*, vol. 45, no. 4, pp. 1782–1791, Apr. 2012.

[60] S. Nadarajah, "The kotz-type distribution with applications," *Statistics*, vol. 37, no. 4, pp. 341–358, Jul. 2003.

[61] J.-F. Cardoso and B. H. Laheld, "Equivariant adaptive source separation," *IEEE Trans. Signal Process.*, vol. 44, no. 12, pp. 3017–3030, Dec. 1996.

[62] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM J. Scientific Comput.*, vol. 16, no. 5, pp. 1190–1208, Sep. 1995.

[63] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, "Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization," *ACM Trans. Math. Softw.*, vol. 23, no. 4, pp. 550–560, 1997.

[64] J. Nocedal and S. Wright, *Numerical Optimization*, 2nd ed. New York, NY, USA: Springer, 2006.

[65] R. A. Waltz, J. L. Morales, J. Nocedal, and D. Orban, "An interior algorithm for nonlinear optimization that combines line search and trust region steps," *Math. Program.*, vol. 107, no. 3, pp. 391–408, Jul. 2006.

[66] S. Haufe *et al.*, "On the interpretation of weight vectors of linear models in multivariate neuroimaging," *NeuroImage*, vol. 87, pp. 96–110, Feb. 2014.

[67] Q. Le, A. Karpenko, J. Ngiam, and A. Ng, "ICA with reconstruction cost for efficient overcomplete feature learning," in *Proc NIPS*, Granada, Spain, 2011, pp. 1017–1025.

[68] MIALAB. (2015). *Group ICA of fMRI Toolbox (GIFT)*. [Online]. Available: http://trendscenter.org/trends/software/gift/index.html

[69] S. Rachakonda, R. F. Silva, J. Liu, and V. D. Calhoun, "Memory efficient PCA methods for large group ICA," *Frontiers Neurosci.*, vol. 10, p. 17, Feb. 2016.

[70] S.-I. Amari, A. Cichocki, and H. H. Yang, "A new learning algorithm for blind signal separation," in *Proc NIPS*, vol. 8, 1996, pp. 757–763.

[71] O. Macchi and E. Moreau, "Self-adaptive source separation by direct or recursive networks," in *Proc ICDSP*, Cyprus, Middle East, 1995, pp. 122–129.

[72] E. A. Allen *et al.*, "A baseline for the multivariate comparison of resting-state networks," *Frontiers Syst. Neurosci.*, vol. 5, p. 2, Feb. 2011.

[73] R. Nelsen, *An Introduction to Copulas* (Springer Series in Statistics), vol. 1, 2nd ed. New York, NY, USA: Springer, 2006.

[74] J. M. Segall *et al.*, "Correspondence between structure and function in the human brain at rest," *Frontiers Neuroinform.*, vol. 6, p. 10, Mar. 2012.

[75] L. Wu, V. D. Calhoun, R. E. Jung, and A. Caprihan, "Connectivity-based whole brain dual parcellation by group ICA reveals tract structures and decreased connectivity in schizophrenia," *Hum. Brain Mapping*, vol. 36, no. 11, pp. 4681–4701, Nov. 2015.

Rogers F. Silva (Member, IEEE) received the B.Sc. degree in electrical engineering from the Pontifical Catholic University (PUCRS), Porto Alegre, Brazil, in 2003, and the M.S. degree in computer engineering (with minors in statistics and in mathematics) and the Ph.D. degree (with distinction) in computer engineering from The University of New Mexico, Albuquerque, NM, USA, in 2011 and 2017, respectively. He is currently a Research Scientist with the Tri-Institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), Georgia Institute of Technology, Georgia State University, and Emory University. Previously, he was a Postdoctoral Fellow with The Mind Research Network, a Data Scientist with Datalytic Solutions, and worked as an Engineer, a Lecturer, and a Consultant. As a multidisciplinary scientist, he develops algorithms for statistical and machine learning, image analysis, numerical optimization, memory efficient large scale data reduction, and distributed analyses, focusing on multimodal, multi-subject neuroimaging data from thousands of subjects. His research interests include multimodal data fusion, statistical and machine learning, image, video, and data analysis, multiobjective, combinatorial and constrained optimization, signal processing, and neuroimaging.

Sergey M. Plis received the Ph.D. degree in computer science from The University of New Mexico, Albuquerque, NM, USA, in 2007. He is currently an Associate Professor of computer science with the Georgia State University and the Director of machine learning core with the Tri-Institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS). His research interests include developing novel and applying existing techniques and approaches to analyzing large scale datasets in multimodal brain imaging and other domains. He develops tools that fall within the fields of machine learning and data science. One of his key goals is to take advantage of the strengths of imaging modalities and infer structure and patterns that are hard to obtain non-invasively and/or that are unavailable for direct observation. In the long term this amounts to developing methods capable of revealing mechanisms used by the brain to form task-specific transient interaction networks and their cognition-inducing interactions via multimodal fusion at features and interaction levels. His ongoing work is focused on inferring multimodal probabilistic and causal descriptions of these function-induced networks based on fusion of fast and slow imaging modalities. This includes feature estimation via deep learning-based pattern recognition and learning causal graphical models.

Tülay Adalı (Fellow, IEEE) received the Ph.D. degree in electrical engineering from North Carolina State University, Raleigh, NC, USA, in 1992. She joined Faculty with the University of Maryland, Baltimore County (UMBC), Baltimore, MD, USA, in 1992, where she is currently a Distinguished University Professor with the Department of Computer Science and Electrical Engineering. Her current research interests include statistical signal processing, machine learning, and their applications, with an emphasis on medical image analysis and fusion. She has been active in conference and workshop organizations. She was the General or a Technical Co-Chair of the IEEE Machine Learning for Signal Processing (MLSP) and Neural Networks for Signal Processing Workshops 2001–2008, and helped organize a number of conferences, including the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). She has served or currently serving on numerous editorial boards and technical committees of the IEEE Signal Processing Society. She was the Chair of the MLSP Technical Committee from 2003 to 2005 and from 2011 to 2013, the Technical Program Co-Chair for ICASSP 2017, and the Special Sessions Chair for ICASSP 2018 and ICASSP 2024. She is currently serving as the Vice President for Technical Directions of the IEEE Signal Processing Society. She is a Fellow of the AIMBE, a Fulbright Scholar, and an IEEE Signal Processing Society Distinguished Lecturer. She was a recipient of a 2020 Humboldt Research Award, the 2010 IEEE Signal Processing Society Best Paper Award, the 2013 University System of Maryland Regents' Award for Research, and the NSF CAREER Award.

Marios S. Pattichis (Senior Member, IEEE) received the B.Sc. degree (Hons.) in computer sciences, the B.A. degree (Hons.) in mathematics, the M.S. degree in electrical engineering, and the Ph.D. degree in computer engineering from The University of Texas at Austin, Austin, in 1991, 1993, and 1998, respectively. He is currently a Professor with the Department of Electrical and Computer Engineering, University of New Mexico (UNM), Albuquerque. His current research interests include digital image and video processing, video communications, dynamically reconfigurable computer architectures, biomedical, space, and educational image processing applications. At UNM, he holds the Gardner Zemke Professorship in Teaching at ECE. He is a fellow of the Center for Collaborative Research and Community Engagement with the College of Education and Human Sciences. He was a recipient of the 2016 Lawton-Ellis and the 2004 Distinguished Teaching Awards from the Department of Electrical and Computer Engineering, UNM. For his development of the digital logic design laboratories at UNM, he was recognized by Xilinx Corporation, in 2003 and by the UNM School of Engineering's Harrison Faculty Excellent Award, in 2006. He was a Founding Co-PI of the Configurable Space Microsystems Innovations and Applications Center (COSMIAC), UNM, where he is currently the Director of ivPCL. He was the General Chair of the 2008 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI). He was a General Co-Chair of the 2020 IEEE SSIAI. He is currently a Senior Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING. He has also served as a Senior Associate Editor for the IEEE SIGNAL PROCESSING LETTERS. He has been an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING and the IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS. He has served as a Guest Associate Editor for the IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE.

Vince D. Calhoun (Fellow, IEEE) received the B.S. degree in electrical engineering from the University of Kansas, Lawrence, KS, USA, in 1991, the M.S. degrees in biomedical engineering and information systems from The Johns Hopkins University, Baltimore, MD, USA, in 1993 and 1996, respectively, and the Ph.D. degree in electrical engineering from the University of Maryland at Baltimore County, Baltimore, in 2002. He is the Founding Director of the Tri-Institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS) and a Georgia Research Alliance Eminent Scholar in Brain Health and Image Analysis, where he holds appointments at the Georgia Institute of Technology, Georgia State University, and Emory University. He was previously the President of The Mind Research Network and a Distinguished Professor of Electrical and Computer Engineering with the University of New Mexico. He is the author of more than 800 full journal articles and over 850 technical reports, abstracts, and conference proceedings. His research interests include the development of flexible methods to analyze functional magnetic resonance imaging data, such as independent component analysis (ICA), deep learning for neuroimaging, data fusion of multimodal imaging and genetics data, neuroinformatics tools, and the identification of biomarkers for disease. His research was funded by NIH and NSF among other funding agencies. He is a fellow of the American Association for the Advancement of Science, the American Institute of Biomedical and Medical Engineers, the American College of Neuropsychopharmacology, and the International Society of Magnetic Resonance in Medicine. He served as the Chair for the Organization for Human Brain Mapping from 2018 to 2019 and is the Past Chair of the IEEE Machine Learning for Signal Processing Technical Committee. He currently serves on the IEEE BISP Technical Committee and is also a member of the IEEE Data Science Initiative Steering Committee.