DEEP INDEPENDENT VECTOR ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce Deep Independent Vector Analysis (DeepIVA), a multivariate latent variable model for learning linked and identifiable latent sources from multimodal data. DeepIVA jointly identifies sources within each modality and aligns sources across modalities. We propose a rigorous set of performance metrics measuring unimodal identifiability, cross-modal linkage, and cross-segment consistency. We validate DeepIVA using synthetic data and then apply it to a large multimodal neuroimaging dataset with structural and functional magnetic resonance imaging. In the synthetic data experiment, DeepIVA successfully recovers nonlinearly mixed multimodal sources, outperforming baseline methods. In the neuroimaging experiment, DeepIVA reveals linked imaging biomarkers associated with sex and age, shedding light on coupled structural-functional brain relationships.

1 INTRODUCTION

A fundamental problem in representation learning is to learn latent variables used to generate observed data. In blind source separation (BSS), independent component analysis (ICA) (Comon, 1994; Hyvärinen & Oja, 2000) aims to recover latent sources that are statistically independent from observed data, but there is no guarantee of identifiability in general without additional assumptions. Notably, the solution of a *linear* ICA problem is identifiable up to permutation and linear rescaling only when at most one of latent sources is Gaussian (Comon, 1994). The solution of a *nonlinear* ICA problem, on the other hand, is highly non-unique without additional restrictions (Hyvärinen & Pajunen, 1999; Locatello et al., 2019). If the learned sources are not identifiable, it is impossible to reveal the underlying structure of the data. Recent advances in nonlinear ICA have proposed to recover identifiable latent sources that are mixed nonlinearly up to trivial indeterminacies by leveraging additional information such as temporal structure (Hyvarinen & Morioka, 2016; Hälvä & Hyvarinen, 2020; Hyvarinen & Morioka, 2017) and auxiliary variables (Hyvarinen et al., 2019; Khemakhem et al., 2020). In particular, an identifiable variational autoencoder (iVAE) (Khemakhem et al., 2020) has been developed to recover nonlinearly mixed sources up to permutation and pointwise nonlinear transformation by utilizing auxiliary variables such as class labels (Hyvarinen et al., 2019).

Apart from identifiability, it is often important to learn linked representations from multiple data modalities, as each modality only captures partial information of the data-generating system. In magnetic resonance imaging, structural MRI (sMRI) can provide static anatomical structure of the brain in high resolution but cannot capture temporal dynamics, whereas functional MRI (fMRI) can measure blood-oxygenation-level-dependent (BOLD) signals over time at the cost of lower spatial resolution. Joint analysis of sMRI and fMRI can uncover spatiotemporal relationships that cannot be detected by a single imaging modality, providing new insights into structural and functional interactions in the brain and its disorders (Calhoun & Sui, 2016). To identify linked sources from multiple datasets, a unified framework called multidataset independent subspace analysis (MISA) (Silva et al., 2021) has been developed, encompassing multiple *linear* latent variable models, such as ICA, independent vector analysis (IVA) (Kim et al., 2006; Adali et al., 2014), and independent subspace analysis (ISA) (Cardoso, 1998; Hyvärinen & Hoyer, 2000). Built upon MISA, multimodal IVA (MMIVA) (Silva et al., 2024) and multimodal subspace IVA (MSIVA) (Li et al., 2026) have been recently developed to capture one-to-one and many-to-many latent multimodal associations, respectively. In both cases, the recovered linked sources are significantly associated with phenotype measures such as age, sex and psychosis from large-scale multimodal neuroimaging datasets including sMRI and fMRI. Although both MMIVA and MSIVA assume that sources undergo a linear mixing process, the true mixing process in neuroimaging data is likely *nonlinear*, considering

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

nonlinear transformations in modeling and preprocessing stages. Nonlinear methods such as deep neural networks (LeCun et al., 2015) have been increasingly applied for neuroimaging data analysis, capturing robust brain-phenotype relationships (Abrol et al., 2021).

Here, we ask the question: How can we learn *linked* and *identifiable* nonlinear representations from multimodal data? We develop Deep Independent Vector Analysis (DeepIVA), a multivariate latent variable model for learning *linked* and *identifiable* latent sources from multiple data modalities. In DeepIVA, we leverage the identifiability properties of iVAEs to recover sources from each modality and the MISA objective function to align sources across all modalities. We theoretically show that sources recovered from DeepIVA are identifiable up to permutation and pointwise transformation. We then empirically demonstrate that DeepIVA successfully recovers nonlinearly mixed sources from multiple synthetic datasets, and effectively recovers linked sources from a large multimodal neuroimaging dataset, compared to baseline methods.

2 METHODS

2.1 DEEP INDEPENDENT VECTOR ANALYSIS

2.1.1 GENERATIVE MODEL

IVA (Kim et al., 2006; Adali et al., 2014) is a multivariate latent variable model which extends the ICA problem from a single dataset to multiple datasets and captures statistical dependence across datasets. IVA aims to identify linked *vector* sources across N_m datasets ($N_m > 1$) where each observation $\mathbf{x}^{[m]}$ can be modeled as a mixture of statistically independent sources $\mathbf{s}^{[m]}$. In DeepIVA, we consider the following multimodal generative model:

$$\mathbf{x}^{[m]} = \mathbf{f}^{[m]}\left(\mathbf{s}^{[m]}\right) + \boldsymbol{\epsilon}^{[m]}, \quad m = 1, \ldots, N_m, \tag{1}$$

$$\begin{aligned} p_{\boldsymbol{\theta}}\left(\mathbf{x}^{[1]}, \ldots, \mathbf{x}^{[N_m]}, \mathbf{s}^{[1]}, \ldots, \mathbf{s}^{[N_m]} | \mathbf{u}\right) &= p_{\boldsymbol{\theta}_\mathbf{s}}(\mathbf{s}|\mathbf{u}) \prod_{m=1}^{N_m} p_{\mathbf{f}^{[m]}}\left(\mathbf{x}^{[m]}|\mathbf{s}^{[m]}\right), \\ &= p_{\boldsymbol{\theta}_\mathbf{s}}(\mathbf{s}|\mathbf{u}) \prod_{m=1}^{N_m} p_{\boldsymbol{\epsilon}^{[m]}}\left(\mathbf{x}^{[m]} - \mathbf{f}^{[m]}\left(\mathbf{s}^{[m]}\right)\right), \end{aligned} \tag{2}$$

and we define the conditionally independent latent model:

$$\begin{aligned} p_{\boldsymbol{\theta}_\mathbf{s}}(\mathbf{s}|\mathbf{u}) &= p_{\boldsymbol{\theta}_\mathbf{s}}\left(\mathbf{s}^{[1]}, \ldots, \mathbf{s}^{[N_m]} | \mathbf{u}\right) \\ &= \prod_{i=1}^{N_s} p_{\boldsymbol{\theta}_{\mathbf{s},i}}\left(s_i^{[1]}, \ldots, s_i^{[N_m]} | \mathbf{u}\right), \end{aligned} \tag{3}$$

where $m \in [\![1, N_m]\!]$ indexes the data modality, $\mathbf{x}^{[m]} \in \mathbb{R}^{N_x}$ is an observed random variable, $\mathbf{u} \in \mathbb{R}^{N_u}$ is an additional observed random variable (e.g., time indices or class labels (Hyvarinen et al., 2019)), $\mathbf{s}^{[m]} \in \mathbb{R}^{N_s}$ ($N_s \leq N_x$) is a latent variable, $\boldsymbol{\epsilon}^{[m]} \in \mathbb{R}^{N_x}$ is a modality-specific independent noise variable with the probability density function (pdf) $p_{\boldsymbol{\epsilon}^{[m]}}\left(\boldsymbol{\epsilon}^{[m]}\right)$, $\boldsymbol{\theta}^{[m]}$ is a set of model parameters, and $\mathbf{f}^{[m]} : \mathbb{R}^{N_s} \to \mathbb{R}^{N_x}$ is an injective nonlinear mixing function. Source component vector (SCV) is defined as a set of linked (statistically dependent) sources spanning N_m datasets, $\mathbf{s}_i = \left[s_i^{[1]}, \ldots, s_i^{[N_m]}\right]^\top$, $\mathbf{s}_i \in \mathbb{R}^{N_m}$. In neuroimaging data, observations are subjects and features are voxel intensities.

The properties of individual modalities can be characterized by integrating $p_{\boldsymbol{\theta}_\mathbf{s}}(\mathbf{s}|\mathbf{u})$ over $s_i^{m'}$, $\forall i$, $\forall m', m' \neq m$, implies the following conditionally independent unimodal latent model:

$$p_{\boldsymbol{\theta}_\mathbf{s}^{[m]}}\left(s_1^{[m]}, \ldots, s_{N_s}^{[m]} | \mathbf{u}\right) = \prod_{i=1}^{N_s} p_{\boldsymbol{\theta}_{s,i}^{[m]}}\left(s_i^{[m]} | \mathbf{u}\right). \tag{4}$$

Assuming the unimodal marginals $s_i^{[m]}|\mathbf{u}$ can be accurately modeled with a univariate exponential family distribution, we can write the following conditionally independent unimodal generative model:

$$\mathbf{x}^{[m]} = \mathbf{f}^{[m]}\left(\mathbf{s}^{[m]}\right) + \boldsymbol{\epsilon}^{[m]}, \quad m = 1, \ldots, N_m, \tag{5}$$

$$
\begin{aligned}
p_{\boldsymbol{\theta}^{[m]}}\left(\mathbf{x}^{[m]}, \mathbf{s}^{[m]} | \mathbf{u}\right) &= p_{\mathbf{f}^{[m]}}\left(\mathbf{x}^{[m]} | \mathbf{s}^{[m]}\right) p_{\boldsymbol{\theta}_{\mathbf{s}}^{[m]}}\left(\mathbf{s}^{[m]} | \mathbf{u}\right), \\
&= p_{\boldsymbol{\epsilon}^{[m]}}\left(\mathbf{x}^{[m]} - \mathbf{f}^{[m]}\left(\mathbf{s}^{[m]}\right)\right) p_{\boldsymbol{\theta}_{\mathbf{s}}^{[m]}}\left(\mathbf{s}^{[m]} | \mathbf{u}\right), \qquad (6) \\
p_{\boldsymbol{\theta}_{\mathbf{s}}^{[m]}}\left(\mathbf{s}^{[m]} | \mathbf{u}\right) &= p_{\mathbf{T}^{[m]}, \boldsymbol{\lambda}^{[m]}}\left(\mathbf{s}^{[m]} | \mathbf{u}\right) \\
&= \prod_{i=1}^{N_s} \frac{Q_i^{[m]}\left(s_i^{[m]}\right)}{Z_i^{[m]}(\mathbf{u})} \exp\left[\sum_{j=1}^{k} T_{i,j}^{[m]}\left(s_i^{[m]}\right) \lambda_{i,j}^{[m]}(\mathbf{u})\right], \qquad (7)
\end{aligned}
$$

where $\boldsymbol{\theta}^{[m]} = \left(\mathbf{f}^{[m]}, \mathbf{T}^{[m]}, \boldsymbol{\lambda}^{[m]}\right)$ is a set of parameters of the unimodal generative model, $Q_i^{[m]}$ is the base measure, $Z_i^{[m]}(\mathbf{u})$ is the normalizing constant, $\mathbf{T}_i^{[m]} = \left(T_{i,1}^{[m]}, \ldots, T_{i,k}^{[m]}\right)$ are the sufficient statistics, $\boldsymbol{\lambda}_i^{[m]}(\mathbf{u}) = \left(\lambda_{i,1}^{[m]}(\mathbf{u}), \ldots, \lambda_{i,k}^{[m]}(\mathbf{u})\right)$ are the parameters depending on $\mathbf{u}$, and k is the dimension of each sufficient statistic.

2.1.2 Identifiable variational autoencoder

In DeepIVA, we leverage the identifiability properties of identifiable variational autoencoders (iVAEs) (Khemakhem et al., 2020) to recover nonlinearly mixed sources by conditioning on auxiliary variables. Given a pair of observations $\left(\mathbf{x}^{[m]}, \mathbf{u}\right)$ sampled from the unimodal generative model defined by Eqs. 5 6 7, the iVAE objective aims to learn the encoder-decoder parameters $\left(\boldsymbol{\phi}^{[m]}, \boldsymbol{\theta}^{[m]}\right)$ that maximize the data generation likelihood by maximizing the evidence lower bound (ELBO):

$$
\mathcal{L}_{\text{iVAE}}^{[m]} = \max_{\boldsymbol{\theta}^{[m]}, \boldsymbol{\phi}^{[m]}} \log p_{\boldsymbol{\theta}^{[m]}}\left(\mathbf{x}^{[m]} | \mathbf{u}\right) - D_{\text{KL}}\left(q_{\boldsymbol{\phi}^{[m]}}\left(\mathbf{s}^{[m]} | \mathbf{x}^{[m]}, \mathbf{u}\right) \| p_{\boldsymbol{\theta}^{[m]}}\left(\mathbf{s}^{[m]} | \mathbf{x}^{[m]}, \mathbf{u}\right)\right), \quad (8)
$$

where $p_{\boldsymbol{\theta}^{[m]}}\left(\mathbf{x}^{[m]} | \mathbf{u}\right) = \int p_{\boldsymbol{\theta}^{[m]}}\left(\mathbf{x}^{[m]}, \mathbf{s}^{[m]} | \mathbf{u}\right) \mathrm{d}\mathbf{s}^{[m]}$ is the conditional marginal distribution of the observations, and $q_{\boldsymbol{\phi}^{[m]}}\left(\mathbf{s}^{[m]} | \mathbf{x}^{[m]}, \mathbf{u}\right)$ is the variational posterior of its true posterior $p_{\boldsymbol{\theta}^{[m]}}\left(\mathbf{s}^{[m]} | \mathbf{x}^{[m]}, \mathbf{u}\right)$.

2.1.3 Multidataset independent subspace analysis

Multidataset independent subspace analysis (MISA) (Silva et al., 2021) is a unified framework encompassing multiple *linear* BSS models including ICA, IVA, and ISA. In DeepIVA, the SCVs are conditionally independent given u, so the joint density factorizes across SCVs. As in MISA, we model each segment u's SCV with a multivariate Kotz distribution (Nadarajah, 2003; Kotz, 1975):

$$
p_\psi\left(\mathbf{s}_i^{[u]}\right) = \frac{\beta \xi^\nu \Gamma\left(\frac{N_m}{2}\right)\left(\mathbf{s}_i^{[u]\top} \mathbf{D}_i^{[u]-1} \mathbf{s}_i^{[u]}\right)^{\eta-1}}{\pi^{\frac{N_m}{2}}\left(\det\left(\mathbf{D}_i^{[u]}\right)\right)^{\frac{1}{2}} \Gamma(\nu)} e^{-\xi\left(\mathbf{s}_i^{[u]\top} \mathbf{D}_i^{[u]-1} \mathbf{s}_i^{[u]}\right)^\beta}, \qquad (9)
$$

where $\psi = [\beta, \xi, \eta]$ is the set of Kotz hyperparameters. The positive definite dispersion matrix $\mathbf{D}_i^{[u]}$ is given by $\mathbf{D}_i^{[u]} = \alpha^{-1} \boldsymbol{\Sigma}_{\mathbf{s}_i^{[u]}}$, where $\boldsymbol{\Sigma}_{\mathbf{s}_i^{[u]}}$ is the SCV covariance matrix. For brevity, we define $\nu \triangleq \frac{2\eta + N_m - 2}{2\beta} > 0$ and $\alpha \triangleq \frac{\Gamma\left(\nu + \beta^{-1}\right)}{\xi^{\beta^{-1}} N_m \Gamma(\nu)}$, where $\Gamma(\cdot)$ denotes the gamma function. The Kotz distribution is highly flexible, encompassing the multivariate Gaussian distribution $\left(\psi_G = \left[1, \frac{1}{2}, 1\right]\right)$ and the multivariate Laplace distribution $\left(\psi_L = \left[\frac{1}{2}, 1, 1\right]\right)$.

The MISA objective function (Silva et al., 2021) is defined as the KL divergence between the joint Kotz distribution of all SCVs $p_\psi\left(\mathbf{s}^{[u]}\right)$ and the product of all SCV distributions $\prod_{i=1}^{N_s} p_\psi\left(\mathbf{s}_i^{[u]}\right)$:

$$
\begin{aligned}
\mathcal{L}_{\text{MISA}}^{[u]} &= \min_{\{\mathbf{g}^{[m]}\}_{m=1}^{N_m}} D_{\text{KL}}\left(p_\psi\left(\mathbf{s}^{[u]}\right) \| \prod_{i=1}^{N_s} p_\psi\left(\mathbf{s}_i^{[u]}\right)\right) \\
&= \min_{\{\mathbf{g}^{[m]}\}_{m=1}^{N_m}} -\sum_{m=1}^{N_m} J_D^{[m,u]} + \frac{1}{2}\sum_{i=1}^{N_s} \ln\left|\det \mathbf{D}_i^{[u]}\right| - f - \sum_{i=1}^{N_s} \frac{\eta - 1}{N} \sum_{n=1}^{N} J_{F_{i,n}}^{[u]} + \sum_{i=1}^{N_s} \frac{\xi}{N} \sum_{n=1}^{N} J_{E_{i,n}}^{[u]}, \qquad (10)
\end{aligned}
$$

where $f = \sum_{i=1}^{N_s} \left[\ln \beta + \nu \ln \xi + \ln \Gamma\left(\frac{N_m}{2}\right) - \frac{N_m}{2} \ln \pi - \ln \Gamma(\nu)\right]$, $J_{F_{i,n}}^{[u]} = \ln\left(\mathbf{s}_{i,n}^{[u]\top} \mathbf{D}_i^{[u]^{-1}} \mathbf{s}_{i,n}^{[u]}\right)$, and $J_{E_{i,n}}^{[u]} = \left(\mathbf{s}_{i,n}^{[u]\top} \mathbf{D}_i^{[u]^{-1}} \mathbf{s}_{i,n}^{[u]}\right)^{\beta}$. In DeepIVA, we compute the Jacobian matrix $\mathbf{J}^{[m,u]}$ of the nonlinear transformation parameterized by each encoder $\mathbf{g}^{[m]}$. For computational efficiency, we approximate each Jacobian by the average Jacobian across samples, $\overline{\mathbf{J}}^{[m,u]} = \frac{1}{N_u} \sum_{n=1}^{N_u} \frac{\partial \mathbf{g}^{[m]}}{\partial \mathbf{x}_n^{[m,u]}}$, where N_u is the number of samples in each segment. The loss term is defined as $J_D^{[m,u]} = \ln \left|\det \overline{\mathbf{J}}^{[m,u]}\right|$ if $\overline{\mathbf{J}}^{[m,u]}$ is a square matrix; $J_D^{[m,u]} = \sum_{i=1}^{N_s} \ln \left|\sqrt{\sigma_i^{[m,u]}}\right|$ where $\left\{\sigma_i^{[m,u]}\right\}_{i=1}^{N_s}$ is the set of non-zero eigenvalues of $\overline{\mathbf{J}}^{[m,u]} \overline{\mathbf{J}}^{[m,u]\top}$ if $\overline{\mathbf{J}}^{[m,u]}$ is not a square matrix.

2.1.4 IDENTIFIABILITY OF DEEPIVA

We provide a conceptual sketch of the proof that the learned model parameters and linked sources from DeepIVA are identifiable up to permutation and pointwise nonlinear transformation (see Appx. A for the detailed proof).

Definition 2.1 ($\sim_{\mathbf{P}^{[m]}}$-identifiability)**.** Let $\mathcal{P} = \{p_{\boldsymbol{\theta}^{[m]}} : \boldsymbol{\theta}^{[m]} \in \Theta\}$ be a statistical model with parameter space Θ. Let $\sim_{\mathbf{P}^{[m]}}$ be the equivalence relation on Θ defined as:

$$\begin{aligned} &\left(\mathbf{f}^{[m]}, \mathbf{T}^{[m]}, \boldsymbol{\lambda}^{[m]}\right) \sim \left(\hat{\mathbf{f}}^{[m]}, \hat{\mathbf{T}}^{[m]}, \hat{\boldsymbol{\lambda}}^{[m]}\right) \Leftrightarrow \\ &\exists \mathbf{P}^{[m]}, \mathbf{c}^{[m]} \mid \mathbf{T}^{[m]}\left(\mathbf{f}^{[m]^{-1}}\left(\mathbf{x}^{[m]}\right)\right) = \mathbf{P}^{[m]} \hat{\mathbf{T}}^{[m]}\left(\hat{\mathbf{f}}^{[m]^{-1}}\left(\mathbf{x}^{[m]}\right)\right) + \mathbf{c}^{[m]}, \end{aligned} \tag{11}$$

where $\mathbf{P}^{[m]} \in \mathbb{R}^{kN_s \times kN_s}$ is a modality-specific permutation matrix [1] and $\mathbf{c}^{[m]} \in \mathbb{R}^{kN_s}$ is a modality-specific constant vector.

Since unimodal sources from iVAEs are identifiable up to permutation ambiguities, a cross-modal alignment step is necessary to ensure that sources are aligned across modalities.

Definition 2.2 (Cross-modal alignment)**.** Sources are *aligned* across N_m modalities if the permutation matrix $\mathbf{P}^{[m]}$ is identical for all N_m modalities:

$$\mathbf{P}^{[m]} = \mathbf{P}, \ \ \forall m \in [\![1, N_m]\!]. \tag{12}$$

Theorem 2.3 (Identifiability of iVAE (Khemakhem et al., 2020))**.** *Assume that the observed data are sampled from a generative model defined by Eqs. 5 6 7. The parameters $\boldsymbol{\theta}^{[m]} = \left(\mathbf{f}^{[m]}, \mathbf{T}^{[m]}, \boldsymbol{\lambda}^{[m]}\right)$ are $\sim_{\mathbf{P}}$-identifiable if the following assumptions hold:*

(i) The mixing function $\mathbf{f}^{[m]}$ in Eq. 5 is injective and has all second order cross derivatives.

(ii) The set $\left\{\mathbf{x}^{[m]} \in \mathcal{X} | \varphi_{\boldsymbol{\epsilon}^{[m]}}\left(\mathbf{x}^{[m]}\right) = 0\right\}$ has measure zero, where $\mathcal{X} \subset \mathbb{R}^{N_x}$ is the image of $\mathbf{f}^{[m]}$ and $\varphi_{\boldsymbol{\epsilon}^{[m]}}$ is the characteristic function of the density $p_{\boldsymbol{\epsilon}^{[m]}}$ defined in Eq. 6.

(iii) The sufficient statistics $T_{i,j}^{[m]}$ in Eq. 7 ($k \geq 2$) are twice differentiable almost everywhere, and $\left(T_{i,j}^{[m]}\right)_{1 \leq j \leq k}$ are linearly independent on any subset of $\mathcal{X}$ of measure greater than zero.

(iv) There exist $kN_s + 1$ distinct points $u_0, \ldots, u_{kN_s}$ such that the matrix $\mathbf{L} \in \mathbb{R}^{kN_s \times kN_s}$ is invertible, $\mathbf{L} = \left[\boldsymbol{\lambda}^{[m]}(u_1) - \boldsymbol{\lambda}^{[m]}(u_0), \ldots, \boldsymbol{\lambda}^{[m]}(u_{kN_s}) - \boldsymbol{\lambda}^{[m]}(u_0)\right]$.

Theorem 2.4 (Sklar's Theorem (Sklar, 1959))**.** *A multivariate joint pdf can be expressed as the product of its univariate marginal pdfs and a copula density, where the latter characterizes all of the dependence between variables.*

[1] A univariate model defined by Eqs. 5 6 7 can only represent the marginal distribution of the joint distribution defined by Eq. 3, and it cannot distinguish between $p_{\boldsymbol{\theta}_\mathbf{s}}\left(\mathbf{s}^{[1]}, \ldots, \mathbf{s}^{[N_m]} | \mathbf{u}\right)$ and $p_{\boldsymbol{\theta}_\mathbf{s}}\left(\mathbf{P}_j^{[1]} \mathbf{s}^{[1]}, \ldots, \mathbf{P}_j^{[N_m]} \mathbf{s}^{[N_m]} | \mathbf{u}\right)$ with arbitrary $\mathbf{P}_j^{[1]} \in \mathbb{R}^{N_s}, \ldots, \mathbf{P}_j^{[N_m]} \in \mathbb{R}^{N_s}$, $1 \leq j \leq k$. Therefore, there is no guarantee that $\mathbf{P}^{[m]}$ is identical across modalities.

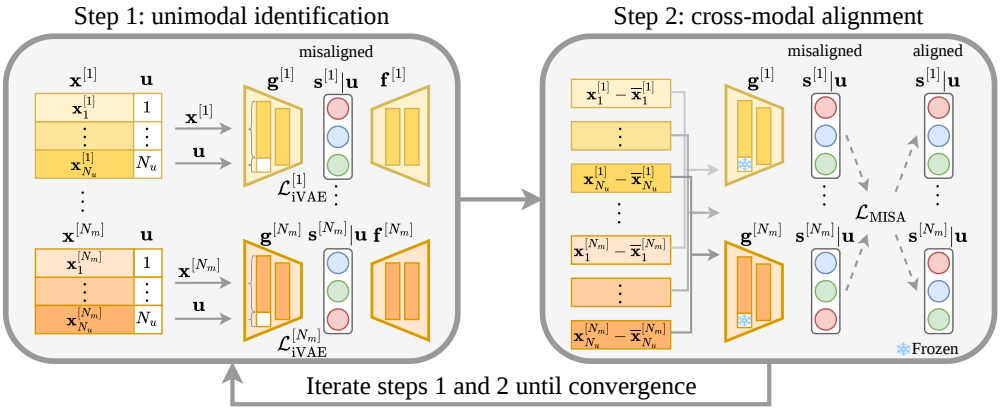


Figure 1: **Overview of the DeepIVA architecture.** Step 1 (unimodal identification): An iVAE is trained to identify sources up to permutation and pointwise nonlinear transformation for each of N_m data modalities. Step 2 (cross-modal alignment): The MISA objective function is applied to align sources across all N_m data modalities. Steps 1 and 2 are iterated until convergence.

Theorem 2.5 (Identifiability of DeepIVA)**.** *Assume that the observed data are sampled from a multimodal generative model defined by Eqs. 1-7 and the assumptions in Thm. 2.3 hold. The learned model parameters and linked sources from DeepIVA are $\sim_{\mathbf{P}}$-identifiable if the SCVs exhibit dependency diversity (non-zero, distinct cross-modal covariance matrices across SCVs).*

Proof sketch. The iVAE model is $\sim_{\mathbf{P}^{[m]}}$-identifiable for each modality, according to Thm. 2.3. Arbitrary pointwise nonlinear transformations (with respect to $\mathbf{T}^{[m]}$ and $\hat{\mathbf{T}}^{[m]}$) remain, but they do not alter the dependence structure between linked sources, according to Thm. 2.4. Thus, we can assume that, without loss of generality, the sufficient statistics $\mathbf{T}^{[m]}$ yield estimates that follow the marginal distribution of a Kotz distribution, so that the MISA objective can align sources across modalities. If the dependency diversity condition in Thm. 2.5 holds, the linked sources remain $\sim_{\mathbf{P}}$-identifiable and aligned across modalities after pointwise nonlinear transformations and alignment. □

2.1.5 Model Architecture

We propose Deep Independent Vector Analysis (DeepIVA), a multivariate latent variable model for learning *linked* and *identifiable* nonlinear representations from multimodal data (Fig. 1). In DeepIVA, an iVAE is initiated for each data modality and then a single MISA module is initiated across all data modalities. The iVAE objective aims to recover sources for each modality while the MISA module aims to identify linkage of sources across modalities. At each epoch, we alternate between training the modality-specific iVAEs and the cross-modal MISA. Specifically, we update the iVAE model parameters $\left(\boldsymbol{\theta}^{[m]}, \boldsymbol{\phi}^{[m]}\right)$ using all segments simultaneously, for each of the N_m modalities separately, following the iVAE loss (Eq. 8). We then process one demeaned segment (defined by the auxiliary variable) from all N_m modalities at a time and simultaneously update the encoder parameters for all modalities according to the MISA objective function (Eq. 10). In iVAE, the reparameterization trick (Kingma & Welling, 2014) is used to sample from a multivariate Gaussian distribution with a diagonal covariance, i.e., $q_{\boldsymbol{\phi}^{[m]}}\left(\mathbf{s}^{[m]}|\mathbf{x}^{[m]}, \mathbf{u}\right) = \mathcal{N}\left(\mathbf{s}^{[m]}|\mathbf{g}^{[m]}\left(\mathbf{x}^{[m]}, \mathbf{u}; \boldsymbol{\phi}_{\mathbf{g}^{[m]}}\right), \mathbf{I}\sigma^2\left(\mathbf{x}^{[m]}, \mathbf{u}; \boldsymbol{\phi}_{\sigma}\right)\right)$. In MISA, we set the Kotz distribution parameters to model the multivariate Laplace distribution, i.e., $\psi_L = \left[\frac{1}{2}, 1, 1\right]$, to capture higher-order statistical dependence (Lee et al., 2008). Since iVAE identifies sources up to unconstrained pointwise nonlinear transformation, we can choose this transformation, without loss of generality, to map the recovered sources onto a multivariate Laplace's marginal distribution, consistent with the distributional assumption MISA imposes.

An L-layer multilayer perceptron (MLP, $L = 2$ in this study) is used as the iVAE backbone. The input dimension of the first layer in the encoder is equal to the sum of the feature dimension and the auxiliary variable dimension $(N_x + N_u)$. The input and output dimensions of each intermediate layer are the same, which doubles the feature size $(2N_x)$. The output dimension of the last layer is again

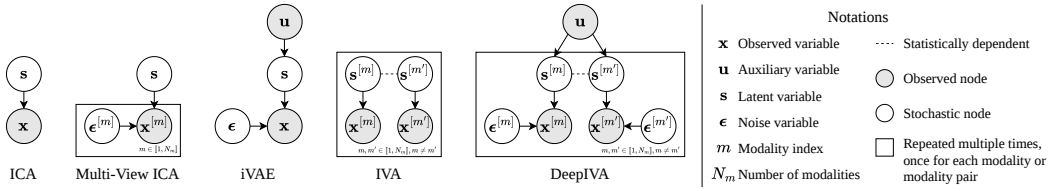


Figure 2: **Graphical model of data generating process.**

Table 1: **Comparison of DeepIVA to existing BSS models.** DeepIVA is the only approach that can identify nonlinearly mixed sources within each modality while also recovering cross-modal linkage.

Model	Identifiability	Multimodality	Nonlinearity	Linkage
Linear ICA (Comon, 1994)	√	×	×	×
Multi-View Linear ICA (Richard et al., 2020)	√	√	×	×
Multi-View Nonlinear ICA (Gresele et al., 2020)	√	√	√	×
iVAE (Khemakhem et al., 2020)	√	×	√	×
Linear IVA (Kim et al., 2006)	√	√	×	√
MISA (Silva et al., 2021)	√	√	×	√
DeepIVA (Ours)	√	√	√	√

equal to the feature dimension (N_x). In each layer, a Leaky ReLU (Maas et al., 2013) with a negative slope of 0.1 is used as the activation function.

In addition, since MISA is not designed to handle auxiliary information, we modify the original encoder architecture to distinguish between data features $\mathbf{x}^{[m]}$ and the auxiliary variable $\mathbf{u}$ such that the iVAE updates model parameters with respect to both $\mathbf{x}^{[m]}$ and $\mathbf{u}$ at the input layer, and that the MISA updates only those pertaining to $\mathbf{x}^{[m]}$ but *not* $\mathbf{u}$. The original iVAE model uses a single input layer taking the concatenated $\mathbf{x}^{[m]}$ and $\mathbf{u}$. In DeepIVA, we split this layer into two: one for data features $\mathbf{x}^{[m]}$ and another for the auxiliary variable $\mathbf{u}$. The parameters with respect to $\mathbf{u}$ are only updated at the iVAE training step but remain frozen at the MISA training step. Also, the inputs for the auxiliary variable are set to zeros during MISA training to ensure no influence from the frozen weights. Models are implemented in the PyTorch framework and trained with NVIDIA V100 GPUs.

2.2 Baseline comparison

We compare DeepIVA with six methods: (1) linear ICA (Comon, 1994), (2) multi-view linear ICA (mvICA) (Richard et al., 2020; 2021), (3) iVAE (Khemakhem et al., 2020), (4) iVAE on concatenated multimodal data along the sample dimension (denoted as joint iVAE, jiVAE), (5) iVAE on concatenated multimodal data along the feature dimension (denoted as group iVAE, giVAE), and (6) linear IVA model in MISA (Kim et al., 2006; Silva et al., 2021).

Fig. 2 illustrates the data generating processes of different models. Unimodal ICA assumes observed data is generated by a latent variable $\mathbf{s}$. Multi-view ICA assumes that different views $\left\{\mathbf{x}^{[m]}\right\}_{m=1}^{N_m}$ originate from a *shared* source variable $\mathbf{s}$ and distinct additive noise variables $\boldsymbol{\epsilon}^{[m]}$, capturing *second*-order statistical dependence. iVAE assumes conditional independence given the auxiliary variable $\mathbf{u}$ to achieve identifiability. IVA assumes that each modality is generated by a modality-specific source variable $\mathbf{s}^{[m]}$, and sources are *linked* across modalities. DeepIVA assumes that each data modality $\mathbf{x}^{[m]}$ is generated by a modality-specific source variable given an auxiliary variable $\mathbf{s}^{[m]}|\mathbf{u}$ and a noise variable $\boldsymbol{\epsilon}^{[m]}$, and sources are *linked* across modalities, following a multivariate Laplace distribution with *all*-order statistical dependence (Lee et al., 2008). Table 1 summarizes the properties of DeepIVA and related BSS models.

2.3 SYNTHETIC DATA EXPERIMENT

We generate synthetic datasets including multivariate Gaussian sources. Specifically, we simulate a dataset $\mathbf{X} \in \mathbb{R}^{N \times N_s \times N_m}$, where $N = N_o \times N_u$ is the number of total observations, N_o is the number of observations per segment, N_u is the number of segments, N_s is the number of sources, and N_m is the number of modalities. Here, we set $N_m = 2$, $N_s = 10$, $N_o = 130$, and $N_u = 22$. These configurations are chosen according to source identification performance in IVA tasks (Li et al., 2023). Appx. B describes the detailed synthetic data generating process.

We apply DeepIVA and six baseline methods (ICA, mvICA, iVAE, jiVAE, giVAE, and IVA module in MISA) to eight synthetic datasets, each generated with a different random seed. As for hyperparameters, we set an initial learning rate of 0.001 for iVAEs. When running the IVA module in MISA, its learning rate is set to the iVAE learning rate divided by the number of segments, since the MISA model is trained on each segment separately. When running DeepIVA, the MISA learning rate decays over training epochs, shifting from cross-modal alignment early on to unimodal separation later (Appx. C). A learning rate scheduler is used to reduce the learning rate by a factor of 0.1 if there is no improvement for 20 epochs. We use a batch size of 78, matching the number of observations per segment in the training set. During training, the model parameters are updated by the Adam optimizer (Kingma & Ba, 2015). Definitive screening design (DSD) (Jones & Nachtsheim, 2011) is used to identify optimal hyperparameters, including the MISA learning rate and Adam momentum coefficients β_1 and β_2 (Appx. C). Each model is trained for 1000 epochs until convergence.

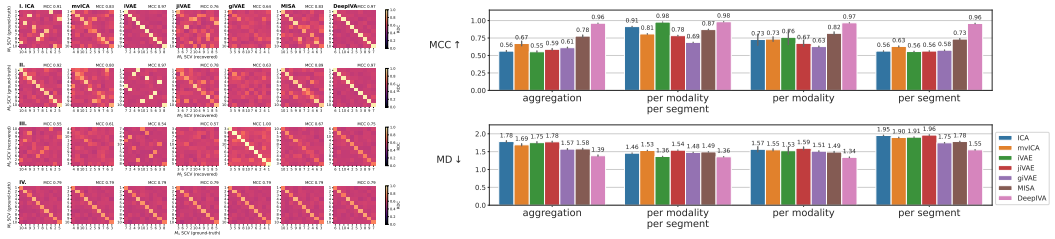
2.4 NEUROIMAGING DATA EXPERIMENT

We utilize the UK Biobank dataset (Miller et al., 2016) $\mathbf{X} \in \mathbb{R}^{N \times N_x \times N_m}$ including two imaging modalities T1-weighted sMRI and resting-state fMRI ($N_m = 2$) from 2907 subjects ($N = 2907$). We preprocess sMRI and fMRI to obtain the gray matter (GM) and mean-scaled amplitude of low frequency fluctuations (mALFF) feature maps, respectively (see Appx. D.1 for preprocessing details). Each GM or mALFF feature map includes 44318 voxels ($N_x = 44318$). Here, we use age and sex groups as auxiliary information, assuming that sources within each modality are conditionally independent given the age and sex group label. This choice is motivated by prior work showing that age and sex substantially influence both brain structure and function (Good et al., 2001; Raz et al., 2004; Ruigrok et al., 2014; Silva et al., 2024; Li et al., 2026). We stratify the neuroimaging data into 22 segments based on the sex and age groups and ensure that the number of samples across segments approximately follows a uniform distribution (Appx. D.2). For each segment, we split samples into training, validation, and test sets with a ratio of $60\% : 20\% : 20\%$.

We evaluate three information-theoretic criteria (Akaike, 1998; Cavanaugh, 1999; Rissanen, 1978) based on the singular values of the multimodal group principal component analysis (MGPCA) covariance matrix derived from the multimodal neuroimaging dataset to select the number of latent sources N_s (Li et al., 2007). We next apply MGPCA on two data modalities (sMRI and fMRI) to reduce the feature dimension from 44318 voxels to N_s common sources. After that, the transformation is applied separately to each data modality in order to obtain modality-specific reduction. We then apply methods to the reduced data $\mathbf{X}_\mathrm{r} \in \mathbb{R}^{N \times N_s \times N_m}$, respectively. MISA and the MISA module in DeepIVA process each segment separately, whereas the other methods use the full batch of 2907 samples. Since the neuroimaging data undergo linear dimension reduction (MGPCA), we assume the resulting sources are linearly mixed and skip the nonlinear activation in DeepIVA and iVAE for the neuroimaging experiment. All other hyperparameters in the neuroimaging experiment match those used in the synthetic data experiment.

2.5 EVALUATION METRICS

We utilize two metrics to evaluate model performance: (1) the trimmed mean correlation coefficient (MCC) between the 25th percentile and the 75th percentile, and (2) the minimum distance (MD). MCC computes the randomized dependence coefficient (RDC) matrix (Lopez-Paz et al., 2013) between two sets of sources, applies the optimal permutation, and then measures similarity along the matrix's diagonal. Unlike MCC, MD also accounts for the off-diagonal (dis)similarity, capturing not only how correlated matched sources are, but also how independent unmatched (off-diagonal) pairs remain. For each metric, we derive four types of coefficients: (1) a coefficient per modality, per



(a) Aggregated RDC matrices across segments (seed 5).

(b) MCC (↑) and MD (↓) between recovered sources and ground-truth sources. The bar plot shows the mean and 95% confidence interval across 8 seeds.

Figure 3: **Synthetic data experiment results** (10 **sources,** 22 **segments,** 78 **samples per segment, training set).** DeepIVA successfully recovers unimodal sources and cross-modal linkage.

segment; (2) an aggregated coefficient per modality; (3) an aggregated coefficient per segment; (4) a final aggregated coefficient across all modalities and segments. See Appx. E for more details.

3 Results

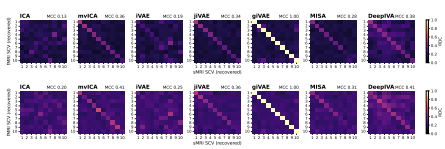
3.1 DeepIVA learns linked and identifiable sources from synthetic data

We identify the following optimal hyperparameters using the synthetic data: MISA initial learning rate $\text{lr}_0 = 0.8 \times 10^{-4}$, target learning rate $\text{lr}_t = 3.22 \times 10^{-5}$, learning rate decay coefficient $\text{T}_{\text{decay}} = 40$ epochs, Adam momentum coefficients $\beta_1 = 0.9812$ and $\beta_2 = 0.8$ (Appx. C). DeepIVA is then trained using these hyperparameters for both synthetic and neuroimaging data experiments.

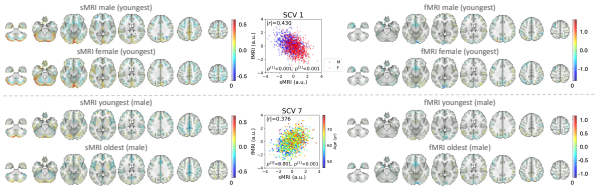
Fig. 3a shows the aggregated RDC matrices for a synthetic dataset with 10 sources, 22 segments, and 78 samples per segment, computed for DeepIVA and baseline methods. Rows I and II show the RDC matrices between the ground-truth and recovered sources for the first modality (M_1) and second modality (M_2), respectively. Successful source identification produces high RDC values along the main diagonal under the same column permutation for both modalities, with greater diagonal-to-off-diagonal contrast indicating better separation. Row III shows the RDC matrices for the recovered sources between modalities, while row IV shows the corresponding matrices for the ground-truth sources. Successful cross-modal alignment produces high RDC values along the main diagonal in row III, matching the pattern in row IV.

As shown in Fig. 3a, ICA recovers sources per modality with fairly good MCCs (M_1 MCC: 0.91; M_2 MCC: 0.92) but fails to align sources across modalities (MCC: 0.55). Multi-view ICA fails to fully recover unimodal sources (M_1 MCC: 0.83; M_2 MCC: 0.80) or align cross-modal sources (MCC: 0.61). iVAE successfully identifies sources within each modality (M_1 MCC: 0.97; M_2 MCC: 0.97) but fails to capture cross-modal linkage (MCC: 0.54). Neither jiVAE nor giVAE recovers unimodal sources or cross-modal structure. giVAE identifies the same set of sources for both modalities, which does not reflect the true cross-modal linkage pattern. By contrast, MISA exhibits stronger cross-modal dependence along the main diagonal, suggesting its ability to detect cross-modal linkage (MCC: 0.67). However, MISA does not fully recover unique unimodal sources (M_1 MCC: 0.83; M_2 MCC: 0.89). DeepIVA successfully recovers unimodal sources (M_1 MCC: 0.97; M_2 MCC: 0.97) and substantially outperforms all baselines in detecting cross-modal linkage (MCC: 0.75).

Fig. 3b presents average MCC and MD for eight synthetic datasets, each generated from a different random seed. For the per-modality per-segment metric, which measures identifiability within each modality and segment without aggregation across either, both iVAE and DeepIVA achieve the best performance, confirming their ability to recover segment-specific unimodal sources. The remaining aggregated metrics additionally capture cross-segment consistency and cross-modal linkage. On these, DeepIVA outperforms all other approaches, achieving an aggregated average MCC of 0.96 ($23.08\% - 74.55\%$ higher than baselines) and an aggregated average MD of 1.39 ($11.46\% - 21.91\%$ lower than baselines). Training set results are reported here, as the larger sample size yields more representative estimates; corresponding test set results are shown in Figs. A1 and A2.



(a) Aggregated RDC matrices between two neuroimaging modalities for the training set (top) and the test set (bottom).



(b) DeepIVA SCVs associated with sex (top) and age (bottom) for the training set.

Figure 4: **Neuroimaging experiment results.** (a) DeepIVA shows strong cross-modal dependence in both the training and test sets. (b) Spatial maps illustrate imaging biomarkers related to sex and age. The Pearson correlation coefficient (r) indicates the cross-modal source correlation. The p-value indicates group differences for sMRI ($p^{[1]}$) and fMRI ($p^{[2]}$) sources separately (two-sample t-test; Bonferroni-correction for 20 comparisons). The axial slices show the reconstructed segment-specific data for each modality, with voxel intensity mapped to both color hue and opacity.

3.2 DeepIVA recovers linked neuroimaging sources associated with sex and age

We next apply DeepIVA and baseline methods to the UKB multimodal neuroimaging dataset to evaluate whether they can detect cross-modal sources in real data. We recover 10 sources for each neuroimaging modality according to the order selected by the information critera (Appx. D.3).

Fig. 4a presents the RDC matrices between two neuroimaging modalities, aggregated across 22 segments, for both the training and test sets. As in the synthetic experiments, DeepIVA shows strong cross-modal dependence along the main diagonal in both the training set (MCC: 0.38) and test set (MCC: 0.41), suggesting that it better captures linked sources across the two imaging modalities (sMRI and fMRI). mvICA shows comparable performance, followed by jiVAE and MISA. giVAE outputs a shared set of sources for both modalities.

We further identify associations between the recovered SCVs from DeepIVA and phenotypes (Fig. 4b), observing significant sex-related SCVs (e.g., SCV 1; male vs. female: $p^{[1]} < 0.001$, $p^{[2]} < 0.001$) and age-related SCVs (e.g., SCV 8; younger vs. older: $p^{[1]} < 0.001$, $p^{[2]} < 0.001$). We present the sMRI and fMRI spatial maps for two representative SCVs (see Appx. D.4 for reconstruction details and test set result). For sex-related SCV 1, females relative to males show stronger positive intensity in the cerebellum and occipital pole, and stronger negative intensity in the inferior temporal gyrus in sMRI; they show similar negative activation in the subcallosal cortex and stronger negative activation in the occipital pole in fMRI. For age-related SCV 7, older relative to younger subjects show stronger negative intensity in the insular cortex and precuneus cortex in sMRI; they show increased negative activation in the subcallosal cortex and reduced negative activation in the occipital pole in fMRI.

4 Discussion

Contributions. We propose Deep Independent Vector Analysis (DeepIVA), a flexible multivariate latent variable model for learning linked and identifiable latent sources from multimodal data. We demonstrate that DeepIVA recovers linked and identifiable sources from multiple synthetic datasets, as well as biologically meaningful sources from a large multimodal neuroimaging dataset.

Limitations. DeepIVA assumes sources are conditionally independent given the auxiliary variable, which may not be available in practice; whether age and sex adequately capture the true neuroimaging data generating process remains unknown. It also assumes linear mixing after MGPCA reduction, leaving nonlinear mixing validation on real data for future work.

Future work. We plan to extend our method from the nonlinear IVA problem to the nonlinear ISA problem, where source dependence is captured across higher-dimensional subspaces. Future work could also explore approaches that do not require side information, such as utilizing a multimodal contrastive learning objective (von Kügelgen et al., 2021; Daunhawer et al., 2023) and constraining the Jacobian of the mixing function via column orthogonality (Gresele et al., 2021; Reizinger et al., 2022) or sparsity (Moran et al., 2022; Zheng et al., 2022; Zheng & Zhang, 2023).

AI USE STATEMENT

In this work, we used generative AI tools to correct grammar and improve clarity of the language. We did not use generative AI tools to write code or to generate the manuscript.

ETHICS STATEMENT

This study used the public and de-identified UK Biobank Resource under Application Number 34175 and did not require the Institutional Review Board (IRB) approval.

REPRODUCIBILITY STATEMENT

Code will be made publicly available upon acceptance.

REFERENCES

Anees Abrol, Zening Fu, Mustafa Salman, Rogers Silva, Yuhui Du, Sergey Plis, and Vince Calhoun. Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. *Nature Communications*, 12(1):353, 2021. doi: 10.1038/s41467-020-20655-6. URL https://doi.org/10.1038/s41467-020-20655-6.

Tülay Adali, Matthew Anderson, and Geng-Shen Fu. Diversity in independent component and vector analyses: Identifiability, algorithms, and applications in medical imaging. *IEEE Signal Processing Magazine*, 31(3):18–33, 2014. doi: 10.1109/MSP.2014.2300511.

Hirotogu Akaike. *Information Theory and an Extension of the Maximum Likelihood Principle*, pp. 199–213. Springer New York, New York, NY, 1998. ISBN 978-1-4612-1694-0. doi: 10.1007/978-1-4612-1694-0_15. URL https://doi.org/10.1007/978-1-4612-1694-0_15.

Matthew Anderson, Geng-Shen Fu, Ronald Phlypo, and Tülay Adalı. Independent vector analysis: Identification conditions and performance bounds. *IEEE Transactions on Signal Processing*, 62 (17):4399–4410, 2014. doi: 10.1109/TSP.2014.2333554.

John Ashburner, Gareth Barnes, Chun-Chuan Chen, Jean Daunizeau, Guillaume Flandin, Karl Friston, Stefan Kiebel, James Kilner, Vladimir Litvak, Rosalyn Moran, Will Penny, Adeel Razi, Klaas Stephan, Sungho Tak, Peter Zeidman, Darren Gitelman, Rik Henson, Chloe Hutton, Volkmar Glauche, Jérémie Mattout, and Christophe Phillips. Spm12 manual. *Wellcome Trust Centre for Neuroimaging, London, UK*, 2464(4), 2014.

Vince D. Calhoun and Jing Sui. Multimodal fusion of brain imaging data: A key to finding the missing link(s) in complex mental illness. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 1(3):230–244, 2016. ISSN 2451-9022. doi: 10.1016/j.bpsc.2015.12.005. URL https://www.sciencedirect.com/science/article/pii/S2451902216000598.

J.-F. Cardoso. Multidimensional independent component analysis. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, volume 4, pp. 1941–1944 vol.4, 1998. doi: 10.1109/ICASSP.1998.681443.

Joseph E Cavanaugh. A large-sample model selection criterion based on kullback's symmetric divergence. *Statistics & Probability Letters*, 42(4):333–343, 1999. ISSN 0167-7152. doi: 10.1016/S0167-7152(98)00200-4. URL https://www.sciencedirect.com/science/article/pii/S0167715298002004.

Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994. ISSN 0165-1684. doi: 10.1016/0165-1684(94)90029-9. URL https://www.sciencedirect.com/science/article/pii/0165168494900299.

David F. Crouse. On implementing 2d rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, 52(4):1679–1696, 2016. doi: 10.1109/TAES.2016.140952.

Imant Daunhawer, Alice Bizeul, Emanuele Palumbo, Alexander Marx, and Julia E Vogt. Identifiability results for multimodal contrastive learning. In *International Conference on Learning Representations (ICLR)*, 2023.

Catriona D. Good, Ingrid S. Johnsrude, John Ashburner, Richard N.A. Henson, Karl J. Friston, and Richard S.J. Frackowiak. A voxel-based morphometric study of ageing in 465 normal adult human brains. *NeuroImage*, 14(1):21–36, 2001. ISSN 1053-8119. doi: https://doi.org/10.1006/nimg.2001.0786. URL https://www.sciencedirect.com/science/article/pii/S1053811901907864.

Luigi Gresele, Paul K. Rubenstein, Arash Mehrjou, Francesco Locatello, and Bernhard Schölkopf. The incomplete rosetta stone problem: Identifiability results for multi-view nonlinear ica. In Ryan P. Adams and Vibhav Gogate (eds.), *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pp. 217–227. PMLR, 22–25 Jul 2020. URL https://proceedings.mlr.press/v115/gresele20a.html.

Luigi Gresele, Julius von Kügelgen, Vincent Stimper, Bernhard Schölkopf, and Michel Besserve. Independent mechanism analysis, a new concept? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 28233–28248. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/edc27f139c3b4e4bb29d1cdbc45663f9-Paper.pdf.

Ludovica Griffanti, Gholamreza Salimi-Khorshidi, Christian F. Beckmann, Edward J. Auerbach, Gwenaëlle Douaud, Claire E. Sexton, Enikő Zsoldos, Klaus P. Ebmeier, Nicola Filippini, Clare E. Mackay, Steen Moeller, Junqian Xu, Essa Yacoub, Giuseppe Baselli, Kamil Ugurbil, Karla L. Miller, and Stephen M. Smith. Ica-based artefact removal and accelerated fmri acquisition for improved resting state network imaging. *NeuroImage*, 95:232–247, 2014. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2014.03.034. URL https://www.sciencedirect.com/science/article/pii/S1053811914001815.

Hermanni Hälvä and Aapo Hyvarinen. Hidden markov nonlinear ica: Unsupervised learning from nonstationary time series. In Jonas Peters and David Sontag (eds.), *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pp. 939–948. PMLR, 03–06 Aug 2020. URL https://proceedings.mlr.press/v124/halva20a.html.

Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/d305281faf947ca7acade9ad5c8c818c-Paper.pdf.

Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ICA of Temporally Dependent Stationary Sources. In Aarti Singh and Jerry Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 460–469. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/hyvarinen17a.html.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 859–868. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/hyvarinen19a.html.

Aapo Hyvärinen and Patrik Hoyer. Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, 12(7):1705–1720, 07 2000. ISSN 0899-7667. doi: 10.1162/089976600300015312. URL https://doi.org/10.1162/089976600300015312.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4):411–430, 2000. ISSN 0893-6080. doi: https://doi.org/10.

1016/S0893-6080(00)00026-5. URL https://www.sciencedirect.com/science/article/pii/S0893608000000265.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999. ISSN 0893-6080. doi: https://doi.org/10.1016/S0893-6080(98)00140-3. URL https://www.sciencedirect.com/science/article/pii/S0893608098001403.

Bradley Jones and Christopher J. Nachtsheim. A class of three-level designs for definitive screening in the presence of second-order effects. *Journal of Quality Technology*, 43(1):1–15, 2011. doi: 10.1080/00224065.2011.11917841. URL https://doi.org/10.1080/00224065.2011.11917841.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2207–2217. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/khemakhem20a.html.

Taesu Kim, Torbjørn Eltoft, and Te-Won Lee. Independent vector analysis: An extension of ica to multivariate components. In Justinian Rosca, Deniz Erdogmus, José C. Príncipe, and Simon Haykin (eds.), *Independent Component Analysis and Blind Signal Separation*, pp. 165–172, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-32631-1. doi: 10.1007/11679363_21.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.

Samuel Kotz. Multivariate distributions at a cross road. In G. P. Patil, S. Kotz, and J. K. Ord (eds.), *A Modern Course on Statistical Distributions in Scientific Work*, pp. 247–270, Dordrecht, 1975. Springer Netherlands. ISBN 978-94-010-1842-5. doi: 10.1007/978-94-010-1842-5_20.

Dana Lahat and Christian Jutten. An alternative proof for the identifiability of independent vector analysis using second order statistics. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4363–4367, 2016. doi: 10.1109/ICASSP.2016.7472501.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. doi: 10.1038/nature14539. URL https://doi.org/10.1038/nature14539.

Jong-Hwan Lee, Te-Won Lee, Ferenc A. Jolesz, and Seung-Schik Yoo. Independent vector analysis (iva): Multivariate approach for fmri group study. *NeuroImage*, 40(1):86–109, 2008. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2007.11.019. URL https://www.sciencedirect.com/science/article/pii/S1053811907010622.

Xinhui Li, Daniel Khosravinezhad, Vince D Calhoun, and Rogers F Silva. Evaluating trade-offs in iva of multimodal neuroimaging using cross-platform multidataset independent subspace analysis. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, 2023. doi: 10.1109/ISBI53787.2023.10230492.

Xinhui Li, Peter Kochunov, Tulay Adali, Rogers F. Silva, and Vince D. Calhoun. Multimodal subspace independent vector analysis effectively captures latent relationships between brain structure and function. *Imaging Neuroscience*, 05 2026. ISSN 2837-6056. doi: 10.1162/IMAG.a.1266. URL https://doi.org/10.1162/IMAG.a.1266.

Yi-Ou Li, Tülay Adalı, and Vince D. Calhoun. Estimating the number of independent components for functional magnetic resonance imaging data. *Human Brain Mapping*, 28(11):1251–1266, 2007. doi: 10.1002/hbm.20359. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.20359.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/locatello19a.html.

David Lopez-Paz, Philipp Hennig, and Bernhard Schölkopf. The randomized dependence coefficient. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/aab3238922bcc25a6f606eb525ffdc56-Paper.pdf.

Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, volume 28, 2013. URL https://ai.stanford.edu/~amaas/papers/relu_hybrid_icml2013_final.pdf.

Karla L Miller, Fidel Alfaro-Almagro, Neal K Bangerter, David L Thomas, Essa Yacoub, Junqian Xu, Andreas J Bartsch, Saad Jbabdi, Stamatios N Sotiropoulos, Jesper L R Andersson, Ludovica Griffanti, Gwenaëlle Douaud, Thomas W Okell, Peter Weale, Iulius Dragonu, Steve Garratt, Sarah Hudson, Rory Collins, Mark Jenkinson, Paul M Matthews, and Stephen M Smith. Multimodal population brain imaging in the uk biobank prospective epidemiological study. *Nature Neuroscience*, 19(11):1523–1536, 2016. doi: 10.1038/nn.4393.

Gemma Elyse Moran, Dhanya Sridhar, Yixin Wang, and David Blei. Identifiable deep generative models via sparse decoding. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=vd0onGWZbE.

Saralees Nadarajah. The kotz-type distribution with applications. *Statistics*, 37(4):341–358, 2003. doi: 10.1080/0233188031000078060. URL https://doi.org/10.1080/0233188031000078060.

Klaus Nordhausen, Esa Ollila, and Hannu Oja. On the performance indices of ica and blind source separation. In *2011 IEEE 12th international workshop on signal processing advances in wireless communications*, pp. 486–490. IEEE, 2011. doi: 10.1109/SPAWC.2011.5990458.

Naftali Raz, Faith Gunning-Dixon, Denise Head, Karen M Rodrigue, Adrienne Williamson, and James D Acker. Aging, sexual dimorphism, and hemispheric asymmetry of the cerebral cortex: replicability of regional differences in volume. *Neurobiology of Aging*, 25(3):377–396, 2004. ISSN 0197-4580. doi: https://doi.org/10.1016/S0197-4580(03)00118-0. URL https://www.sciencedirect.com/science/article/pii/S0197458003001180.

Patrik Reizinger, Luigi Gresele, Jack Brady, Julius von Kügelgen, Dominik Zietlow, Bernhard Schölkopf, Georg Martius, Wieland Brendel, and Michel Besserve. Embrace the gap: Vaes perform independent mechanism analysis. In *Advances in Neural Information Processing Systems*, volume 35, pp. 12040–12057. Curran Associates, Inc., 2022. doi: 10.52202/068431-0875. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/4eb91efe090f72f7cf42c69aab03fe85-Paper-Conference.pdf.

Hugo Richard, Luigi Gresele, Aapo Hyvarinen, Bertrand Thirion, Alexandre Gramfort, and Pierre Ablin. Modeling shared responses in neuroimaging studies through multiview ica. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19149–19162. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/de03beffeed9da5f3639a621bcab5dd4-Paper.pdf.

Hugo Richard, Pierre Ablin, Bertrand Thirion, Alexandre Gramfort, and Aapo Hyvarinen. Shared independent component analysis for multi-subject neuroimaging. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29962–29971. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/fb508ef074ee78a0e58c68be06d8a2eb-Paper.pdf.

Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978. ISSN 0005-1098. doi: https://doi.org/10.1016/0005-1098(78)90005-5. URL https://www.sciencedirect.com/science/article/pii/0005109878900055.

Amber N.V. Ruigrok, Gholamreza Salimi-Khorshidi, Meng-Chuan Lai, Simon Baron-Cohen, Michael V. Lombardo, Roger J. Tait, and John Suckling. A meta-analysis of sex differences in human brain structure. *Neuroscience & Biobehavioral Reviews*, 39:34–50, 2014. ISSN 0149-7634. doi: https://doi.org/10.1016/j.neubiorev.2013.12.004. URL https://www.sciencedirect.com/science/article/pii/S0149763413003011.

Rogers F. Silva, Sergey M. Plis, Tülay Adalı, Marios S. Pattichis, and Vince D. Calhoun. Multidataset independent subspace analysis with application to multimodal fusion. *IEEE Transactions on Image Processing*, 30:588–602, 2021. doi: 10.1109/TIP.2020.3028452.

Rogers F. Silva, Eswar Damaraju, Xinhui Li, Peter Kochunov, Judith M. Ford, Daniel H. Mathalon, Jessica A. Turner, Theo G. M. van Erp, Tulay Adali, and Vince D. Calhoun. A method for multimodal iva fusion within a misa unified model reveals markers of age, sex, cognition, and schizophrenia in large neuroimaging studies. *Human Brain Mapping*, 45(17):e70037, 2024. doi: 10.1002/hbm.70037. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.70037.

Abe Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8(3):229–231, 1959.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 16451–16467. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/8929c70f8d710e412d38da624b21c3c8-Paper.pdf.

Yu-Feng Zang, Yong He, Chao-Zhe Zhu, Qing-Jiu Cao, Man-Qiu Sui, Meng Liang, Li-Xia Tian, Tian-Zi Jiang, and Yu-Feng Wang. Altered baseline brain activity in children with adhd revealed by resting-state functional mri. *Brain and Development*, 29(2):83–91, 2007. ISSN 0387-7604. doi: 10.1016/j.braindev.2006.07.002. URL https://www.sciencedirect.com/science/article/pii/S0387760406001549.

Na Zhao, Li-Xia Yuan, Xi-Ze Jia, Xu-Feng Zhou, Xin-Ping Deng, Hong-Jian He, Jianhui Zhong, Jue Wang, and Yu-Feng Zang. Intra- and inter-scanner reliability of voxel-wise whole-brain analytic metrics for resting state fmri. *Frontiers in Neuroinformatics*, Volume 12 - 2018, 2018. ISSN 1662-5196. doi: 10.3389/fninf.2018.00054. URL https://www.frontiersin.org/journals/neuroinformatics/articles/10.3389/fninf.2018.00054.

Yujia Zheng and Kun Zhang. Generalizing nonlinear ica beyond structural sparsity. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 13326–13355. Curran Associates, Inc., 2023. doi: 10.52202/075280-0585. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/2aebc17b683792a17dd4a24fcb038ba6-Paper-Conference.pdf.

Yujia Zheng, Ignavier Ng, and Kun Zhang. On the identifiability of nonlinear ica: Sparsity and beyond. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 16411–16422. Curran Associates, Inc., 2022. doi: 10.52202/068431-1194. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6801fa3fd290229efc490ee0cf1c5687-Paper-Conference.pdf.

A IDENTIFIABILITY OF DEEPIVA

Here we provide a proof that DeepIVA recovers sources that are identifiable up to permutation and pointwise nonlinear transformation within each modality and aligned across modalities.

A.1 GENERATIVE MODEL

We consider the following conditionally independent multimodal generative model:

$$\mathbf{x}^{[m]} = \mathbf{f}^{[m]}\left(\mathbf{s}^{[m]}\right) + \boldsymbol{\epsilon}^{[m]}, \quad m = 1, \ldots, N_m, \tag{A1}$$

$$\begin{aligned} p_{\boldsymbol{\theta}}\left(\mathbf{x}^{[1]}, \ldots, \mathbf{x}^{[N_m]}, \mathbf{s}^{[1]}, \ldots, \mathbf{s}^{[N_m]} | \mathbf{u}\right) &= p_{\boldsymbol{\theta}_\mathbf{s}}(\mathbf{s}|\mathbf{u}) \prod_{m=1}^{N_m} p_{\mathbf{f}^{[m]}}\left(\mathbf{x}^{[m]}|\mathbf{s}^{[m]}\right), \\ &= p_{\boldsymbol{\theta}_\mathbf{s}}(\mathbf{s}|\mathbf{u}) \prod_{m=1}^{N_m} p_{\boldsymbol{\epsilon}^{[m]}}\left(\mathbf{x}^{[m]} - \mathbf{f}^{[m]}\left(\mathbf{s}^{[m]}\right)\right), \end{aligned} \tag{A2}$$

where we define the conditionally independent latent model:

$$\begin{aligned} p_{\boldsymbol{\theta}_\mathbf{s}}(\mathbf{s}|\mathbf{u}) &= p_{\boldsymbol{\theta}_\mathbf{s}}\left(\mathbf{s}^{[1]}, \ldots, \mathbf{s}^{[N_m]} | \mathbf{u}\right) \\ &= \prod_{i=1}^{N_s} p_{\boldsymbol{\theta}_{\mathbf{s},i}}\left(s_i^{[1]}, \ldots, s_i^{[N_m]} | \mathbf{u}\right) \\ &= \prod_{i=1}^{N_s} p_{\boldsymbol{\theta}_{\mathbf{s},i}}\left(\mathbf{s}_i | \mathbf{u}\right), \end{aligned} \tag{A3}$$

where $m \in [\![1, N_m]\!]$ indexes the data modality, $\mathbf{x}^{[m]} \in \mathbb{R}^{N_x}$ is an observed random variable, $\mathbf{u} \in \mathbb{R}^{N_u}$ is an additional observed random variables (e.g., time indices or class labels (Hyvarinen et al., 2019)), $\mathbf{s}^{[m]} \in \mathbb{R}^{N_s}$ ($N_s \leq N_x$) is a latent random variable, $\boldsymbol{\epsilon}^{[m]} \in \mathbb{R}^{N_x}$ is a modality-specific independent noise variable with probability density function (pdf) $p_{\boldsymbol{\epsilon}^{[m]}}\left(\boldsymbol{\epsilon}^{[m]}\right)$, $\boldsymbol{\theta}$ is a set of parameters of the generative model, $\mathbf{f}^{[m]} : \mathbb{R}^{N_s} \to \mathbb{R}^{N_x}$ is a mixing function. Also, we denote $\mathbf{s}_i \in \mathbb{R}^{N_m}$ as a set of latent variables linked across modalities with a joint pdf $p_{\boldsymbol{\theta}_{\mathbf{s},i}}\left(\mathbf{s}_i|\mathbf{u}\right)$.

The individual modalities can be characterized by integrating $p_{\boldsymbol{\theta}_\mathbf{s}}(\mathbf{s}|\mathbf{u})$ over $s_i^{[m']}$, $\forall i$, $\forall m', m' \neq m$, which implies the following (marginal) conditionally independent unimodal latent model:

$$p_{\boldsymbol{\theta}_\mathbf{s}^{[m]}}\left(s_1^{[m]}, \ldots, s_{N_s}^{[m]} | \mathbf{u}\right) = \prod_{i=1}^{N_s} p_{\boldsymbol{\theta}_{s,i}^{[m]}}\left(s_i^{[m]} | \mathbf{u}\right). \tag{A4}$$

Definition A.1 (Exponential family)**.** A univariate exponential family is a set of probability distributions whose pdf (or probability mass function for a discrete random variable) can be expressed as

$$p(x) = Q(x)Z(\boldsymbol{\theta})e^{\langle \mathbf{T}(x), \boldsymbol{\theta} \rangle}, \tag{A5}$$

where $\mathbf{T} : \mathbb{R} \to \mathbb{R}^k$ is the sufficient statistic, $\boldsymbol{\theta}$ is the natural parameter, $Q : \mathbb{R} \to \mathbb{R}$ is the base measure, and $Z(\boldsymbol{\theta})$ is the normalizing constant.

Assumption A.2 (Unimodal exponential family)**.** The unimodal marginals $s_i^{[m]}|\mathbf{u}$ can be accurately modeled with a univariate exponential family distribution.

According to Assumption A.2, we can write the following conditionally independent unimodal generative model:

$$\mathbf{x}^{[m]} = \mathbf{f}^{[m]}\left(\mathbf{s}^{[m]}\right) + \boldsymbol{\epsilon}^{[m]}, \quad m = 1, \ldots, N_m, \tag{A6}$$

$$\begin{aligned} p_{\boldsymbol{\theta}^{[m]}}\left(\mathbf{x}^{[m]}, \mathbf{s}^{[m]} | \mathbf{u}\right) &= p_{\mathbf{f}^{[m]}}\left(\mathbf{x}^{[m]}|\mathbf{s}^{[m]}\right) p_{\boldsymbol{\theta}_\mathbf{s}^{[m]}}\left(\mathbf{s}^{[m]}|\mathbf{u}\right), \\ &= p_{\boldsymbol{\epsilon}^{[m]}}\left(\mathbf{x}^{[m]} - \mathbf{f}^{[m]}\left(\mathbf{s}^{[m]}\right)\right) p_{\boldsymbol{\theta}_\mathbf{s}^{[m]}}\left(\mathbf{s}^{[m]}|\mathbf{u}\right), \end{aligned} \tag{A7}$$

$$
\begin{aligned}
p_{\boldsymbol{\theta}_{\mathbf{s}}^{[m]}}\left(\mathbf{s}^{[m]}|\mathbf{u}\right) &= p_{\mathbf{T}^{[m]},\boldsymbol{\lambda}^{[m]}}\left(\mathbf{s}^{[m]}|\mathbf{u}\right) \\
&= \prod_{i}^{N_s} \frac{Q_i^{[m]}\left(s_i^{[m]}\right)}{Z_i^{[m]}(\mathbf{u})} \exp\left[\sum_{j=1}^{k} T_{i,j}^{[m]}\left(s_i^{[m]}\right)\lambda_{i,j}^{[m]}(\mathbf{u})\right],
\end{aligned} \tag{A8}
$$

where $\boldsymbol{\theta}^{[m]} = \left(\mathbf{f}^{[m]}, \mathbf{T}^{[m]}, \boldsymbol{\lambda}^{[m]}\right)$ is a set of parameters of the unimodal generative model, $Q_i^{[m]}$ is the base measure, $Z_i^{[m]}(\mathbf{u})$ is the normalizing constant, $\mathbf{T}_i^{[m]} = \left(T_{i,1}^{[m]}, \ldots, T_{i,k}^{[m]}\right)$ are the sufficient statistics, $\boldsymbol{\lambda}_i^{[m]}(\mathbf{u}) = \left(\lambda_{i,1}^{[m]}(\mathbf{u}), \ldots, \lambda_{i,k}^{[m]}(\mathbf{u})\right)$ are the parameters depending on $\mathbf{u}$, and k is the dimension of each sufficient statistic.

A.2 Identifiability theory

Definition A.3 (Identifiability)**.** Let $\mathcal{P} = \left\{p_{\boldsymbol{\theta}^{[m]}} : \boldsymbol{\theta}^{[m]} \in \Theta\right\}$ be a probability set of a statistical model with parameter space Θ. We say $\mathcal{P}$ is identifiable up to an equivalence relation $\sim$ if:

$$
\forall \boldsymbol{\theta}^{[m]}, \hat{\boldsymbol{\theta}}^{[m]} \in \Theta,\ p_{\boldsymbol{\theta}^{[m]}} = p_{\hat{\boldsymbol{\theta}}^{[m]}} \Rightarrow \boldsymbol{\theta}^{[m]} \sim \hat{\boldsymbol{\theta}}^{[m]}. \tag{A9}
$$

Definition A.4 ($\sim_{\mathbf{A}^{[m]}}$-identifiability)**.** Let $\sim_{\mathbf{A}^{[m]}}$ be the equivalence relation on Θ defined as follows:

$$
\begin{aligned}
&\left(\mathbf{f}^{[m]}, \mathbf{T}^{[m]}, \boldsymbol{\lambda}^{[m]}\right) \sim \left(\hat{\mathbf{f}}^{[m]}, \hat{\mathbf{T}}^{[m]}, \hat{\boldsymbol{\lambda}}^{[m]}\right) \Leftrightarrow \\
&\exists \mathbf{A}^{[m]}, \mathbf{c}^{[m]} \mid \mathbf{T}^{[m]}\left(\mathbf{f}^{[m]^{-1}}\left(\mathbf{x}^{[m]}\right)\right) = \mathbf{A}^{[m]}\hat{\mathbf{T}}^{[m]}\left(\hat{\mathbf{f}}^{[m]^{-1}}\left(\mathbf{x}^{[m]}\right)\right) + \mathbf{c}^{[m]},\ \forall \mathbf{x}^{[m]} \in \mathcal{X},
\end{aligned} \tag{A10}
$$

where $\mathbf{A}^{[m]} \in \mathbb{R}^{kN_s \times kN_s}$ is a modality-specific invertible linear matrix, $\mathbf{c}^{[m]} \in \mathbb{R}^{kN_s}$ is a modality-specific vector, and $\mathcal{X} \subset \mathbb{R}^{N_x}$ is the image of $\mathbf{f}^{[m]}$.

Definition A.5 ($\sim_{\mathbf{P}^{[m]}}$-identifiability)**.** Let $\sim_{\mathbf{P}^{[m]}}$ be the equivalence relation on Θ defined as follows:

$$
\begin{aligned}
&\left(\mathbf{f}^{[m]}, \mathbf{T}^{[m]}, \boldsymbol{\lambda}^{[m]}\right) \sim \left(\hat{\mathbf{f}}^{[m]}, \hat{\mathbf{T}}^{[m]}, \hat{\boldsymbol{\lambda}}^{[m]}\right) \Leftrightarrow \\
&\exists \mathbf{P}^{[m]}, \mathbf{c}^{[m]} \mid \mathbf{T}^{[m]}\left(\mathbf{f}^{[m]^{-1}}\left(\mathbf{x}^{[m]}\right)\right) = \mathbf{P}^{[m]}\hat{\mathbf{T}}^{[m]}\left(\hat{\mathbf{f}}^{[m]^{-1}}\left(\mathbf{x}^{[m]}\right)\right) + \mathbf{c}^{[m]},\ \forall \mathbf{x}^{[m]} \in \mathcal{X},
\end{aligned} \tag{A11}
$$

where $\mathbf{P}^{[m]} \in \mathbb{R}^{kN_s \times kN_s}$ is a modality-specific permutation matrix and $\mathbf{c}^{[m]} \in \mathbb{R}^{kN_s}$ is a modality-specific vector.

Although unimodal sources may be $\mathbf{P}^{[m]}$-identifiable, cross-modal alignment is necessary to ensure that sources are aligned across modalities.

Definition A.6 (Cross-modal alignment)**.** Sources are *aligned* across N_m modalities if the permutation matrix $\mathbf{P}^{[m]}$ is identical for all N_m modalities:

$$
\mathbf{P}^{[m]} = \mathbf{P},\ \ \forall m \in [\![1, N_m]\!]. \tag{A12}
$$

If a unimodal generative model defined by Eqs. A6 A7 A8 follows additional assumptions, the learned model parameters and sources are identifiable up to trivial indeterminacies, as proved in Khemakhem et al.. We restate the key assumptions and theorems in Khemakhem et al. as follows.

Theorem A.7 ($\sim_{\mathbf{A}^{[m]}}$-identifiability of iVAE)**.** *Assume that the observed data are sampled from a generative model defined by Eqs. A6 A7 A8. The parameters $\boldsymbol{\theta}^{[m]} = \left(\mathbf{f}^{[m]}, \mathbf{T}^{[m]}, \boldsymbol{\lambda}^{[m]}\right)$ are $\sim_{\mathbf{A}^{[m]}}$-identifiable if the following assumptions hold:*

- *(i) The mixing function $\mathbf{f}^{[m]}$ in Eq. A6 is injective.*
- *(ii) The set $\left\{\mathbf{x}^{[m]} \in \mathcal{X} | \varphi_{\boldsymbol{\epsilon}^{[m]}}\left(\mathbf{x}^{[m]}\right) = 0\right\}$ has measure zero, where $\varphi_{\boldsymbol{\epsilon}^{[m]}}$ is the characteristic function of the density $p_{\boldsymbol{\epsilon}^{[m]}}$ defined in Eq. A7.*
- *(iii) The sufficient statistics $T_{i,j}^{[m]}$ in Eq. A8 are differentiable almost everywhere, and $\left(T_{i,j}^{[m]}\right)_{1 \le j \le k}$ are linearly independent on any subset of $\mathcal{X}$ of measure greater than zero.*

(iv) There exist $kN_s + 1$ distinct points $u_0, \ldots, u_{kN_s}$ such that the matrix $\mathbf{L} \in \mathbb{R}^{kN_s \times kN_s}$ is invertible, $\mathbf{L} = \left[\boldsymbol{\lambda}^{[m]}(u_1) - \boldsymbol{\lambda}^{[m]}(u_0), \ldots, \boldsymbol{\lambda}^{[m]}(u_{kN_s}) - \boldsymbol{\lambda}^{[m]}(u_0)\right]$.

Let $\hat{\theta}^{[m]} = \left(\hat{\mathbf{f}}^{[m]}, \hat{\mathbf{T}}^{[m]}, \hat{\boldsymbol{\lambda}}^{[m]}\right)$ be the learned parameters from some algorithm that approximates the marginal distribution of the observations from the m-th dataset. Thm. A.7 states necessarily $\hat{\theta}^{[m]} \sim_{\mathbf{A}} \theta^{[m]}$, where $\mathbf{A}^{[m]}$ is an invertible linear transformation. The recovered $\hat{\mathbf{H}}^{[m]} = \hat{\mathbf{T}}^{[m]} \circ \hat{\mathbf{f}}^{[m]^{-1}}\left(\mathbf{x}^{[m]}\right)$ are equal to the true $\mathbf{H}^{[m]} = \mathbf{T}^{[m]} \circ \mathbf{f}^{[m]^{-1}}\left(\mathbf{x}^{[m]}\right)$ up to a linear transformation $\mathbf{A}^{[m]}$ of a pointwise nonlinear transformation, where $\circ$ is a function composition defined as $\mathbf{T}^{[m]} \circ \mathbf{f}^{[m]^{-1}}(\cdot) = \mathbf{T}^{[m]}\left(\mathbf{f}^{[m]^{-1}}(\cdot)\right)$. We can further assume sufficient conditions to obtain identifiability up to permutation and pointwise nonlinear transformation ($\sim_{\mathbf{P}^{[m]}}$), as stated in the following Thm. A.8.

Theorem A.8 ($\sim_{\mathbf{P}^{[m]}}$-identifiability of iVAE)**.** *Assume the hypotheses of Thm. A.7 hold. The parameters $\boldsymbol{\theta}^{[m]}$ are $\sim_{\mathbf{P}^{[m]}}$-identifiable if the following assumptions hold:*

(i) The sufficient statistics $T_{i,j}^{[m]}$ ($k \geq 2$) are twice differentiable.

(ii) The mixing function $\mathbf{f}^{[m]}$ has all second order cross derivatives.

Theorem A.9 (Sklar's Theorem (Sklar, 1959))**.** *A multivariate joint pdf can be expressed as the product of its univariate marginal pdfs and a copula density. The latter characterizes all of the dependence between variables.*

By Thm. A.9, Eq. A3 can be written as:

$$\prod_{i=1}^{N_s} p_{\boldsymbol{\theta}_{\mathbf{s},i}}\left(s_i^{[1]}, \ldots, s_i^{[N_m]} | \mathbf{u}\right) = \prod_{i=1}^{N_s}\left(c\left(F_i^{[1]}\left(s_i^{[1]}\right), \ldots, F_i^{[N_m]}\left(s_i^{[N_m]}\right) | \mathbf{u}\right) \cdot \prod_{m=1}^{N_m} p_{\boldsymbol{\theta}_{s,i}^{[m]}}\left(s_i^{[m]} | \mathbf{u}\right)\right), \tag{A13}$$

where $c(\cdot)$ is the copula density (joint pdf) and $F_i^{[m]}\left(s_i^{[m]}\right)$ is the marginal cumulative distribution function.

Remark A.10. The copula density $c\left(F_i^{[1]}\left(s_i^{[1]}\right), \ldots, F_i^{[N_m]}\left(s_i^{[N_m]}\right) | \mathbf{u}\right)$ captures the dependence among all variables $s_i^{[1]}, \ldots, s_i^{[N_m]}$ in the i-th latent subspace. Also, under mild continuity conditions, a nonlinear univariate transformation $\tilde{s}_i^{[m]} = h_i^{[m]}\left(s_i^{[m]}\right)$ can be compensated by a corresponding $\tilde{F}_i^{[m]}(\cdot)$ such that $\tilde{F}_i^{[m]}\left(\tilde{s}_i^{[m]}\right) = F_i^{[m]}\left(s_i^{[m]}\right)$, and no changes are made to the copula. Thus, the pointwise nonlinearity $h_{i,k}^{[m]} = T_{i,k}^{[m]}\left(\mathbf{f}^{-1^{[m]}}(\cdot)\right)$ from iVAE does not alter the dependence structure between linked sources.

Remark A.11. The conditional copula $c\left(F_i^{[1]}\left(s_i^{[1]}\right), \cdots, F_i^{[N_m]}\left(s_i^{[N_m]}\right) | \mathbf{u}\right)$ implies that the dependence structure may change with segment and, thus, the MISA alignment estimation should be performed segment-wise.

Theorem A.12 ($\sim_{\mathbf{P}}$-identifiability of DeepIVA)**.** *Assume that the observed data are sampled from a multimodal generative model defined by Eqs. A1 A2 A3 A4 A7 and A8, and the assumptions in Thms. A.7 and A.8 hold. The learned model parameters and linked sources from DeepIVA are $\sim_{\mathbf{P}}$-identifiable and aligned if the SCVs exhibit dependency diversity (non-zero, distinct cross-modal covariance matrices across SCVs).*

Proof. In DeepIVA, the assumptions of Thms. A.7 and A.8 suffice for identifiability per data modality, and lead to $\sim_{\mathbf{P}^{[m]}}$-identifiable iVAE model parameters and sources for each data modality. Although these assumptions suffice for unimodal identifiability, they provide no guarantee of cross-modal alignment. In addition, following iVAE estimation, arbitrary pointwise nonlinear transformations (with respect to $\mathbf{T}^{[m]} \circ \mathbf{f}^{[m]^{-1}}$ and $\hat{\mathbf{T}}^{[m]} \circ \hat{\mathbf{f}}^{[m]^{-1}}$) remain, but they do not alter the underlying dependence structure between linked sources, based on Thm. A.9. Thus, pointwise nonlinear transformations $\hat{\mathbf{H}}^{[m]} = \hat{\mathbf{T}}^{[m]} \circ \hat{\mathbf{f}}^{[m]^{-1}}\left(\mathbf{x}^{[m]}\right)$ with distribution matching a Kotz marginal exist, and the maximum likelihood procedure in MISA will induce them (see Section 2.1.3). Consequently,

the MISA objective function can be utilized to align sources across data modalities, effectively selecting for pointwise nonlinearities that yield Kotz marginal distributions. If only the linear indeterminacies $\mathbf{A}_m$ remain and the dependency diversity condition in Thm. A.12 holds, alignment can be attained (Lahat & Jutten, 2016; Anderson et al., 2014), rendering $\mathbf{P}^{[m]} = \mathbf{P}$ and, therefore, $\hat{\boldsymbol{\theta}}$ is $\sim_{\mathbf{P}}$-identifiable up to modality-specific pointwise nonlinear transformations and a shared (aligned) arbitraty permutation. This concludes the identifiability proof for DeepIVA. □

Remark A.13. When the pointwise nonlinearities are Kotz-like (following per-modality estimation with iVAE), only the cross-modal misalignment needs to be resolved. Thus, combinatorial optimization of the MISA objective function suffices to search for modality-specific source permutation matrices $\mathbf{P}^{[m]}$. However, this form of two-step approach is highly inefficient given the poor scalability of combinatorial search when the number of modalities N_m and the number of latent sources N_s grow. Thus, we opt to pursue numerical optimization of the MISA objective function to promote learning of a pointwise nonlinear function that yields Kotz-like marginals. This also has an additional benefit of promoting source alignment in tandem with identification via unimodal iVAE training.

B SYNTHETIC DATA EXPERIMENT

B.1 DATA GENERATING PROCESS

We generate synthetic datasets with multivariate Gaussian sources. Specifically, we simulate a dataset $\mathbf{X} \in \mathbb{R}^{N \times N_s \times N_m}$ where $N = N_o \times N_u$ is the number of total observations, N_o is the number of observations per segment, N_u is the number of segments, N_s is the number of sources, and N_m is the number of modalities. Here, we set $N_m = 2$, $N_s = 10$, $N_o = 130$ and $N_u = 22$ to simulate real data. The auxiliary variable $\mathbf{u} \in \mathbb{R}^N$ is the segment label with a uniform distribution on the integer set $[\![1, N_u]\!]$. Latent sources within each modality are conditionally independent given $\mathbf{u}$.

For each segment u, we generate a covariance matrix $\mathbf{\Sigma}_u \in \mathbb{R}^{2N_s \times 2N_s}$ over both modalities. The within-modality block $\mathbf{\Sigma}_u^{[m,m]} \in \mathbb{R}^{N_s \times N_s}$ along the main block diagonal is a diagonal matrix with entries sampled from a uniform distribution $[0.2, 4]$. For source $i = \text{mod}(u, N_s)$, we set $\mathbf{\Sigma}_u^{[m,m]}[i, i] = 8$ identically in both modalities ($m = 1, 2$), assigning source i a segment-specific variance signature that we use to evaluate source identifiability across segments.

The between-modality block $\mathbf{\Sigma}_u^{[m,m']} \in \mathbb{R}^{N_s \times N_s}$ ($m \neq m'$) is a diagonal matrix defined as $\mathbf{\Sigma}_u^{[m,m']} = \mathbf{R}_u \sqrt{\mathbf{\Sigma}_u^{[1,1]} \mathbf{\Sigma}_u^{[2,2]}}$, where $\mathbf{R}_u$ is a diagonal matrix of correlation values sampled from a uniform distribution $[0.6, 0.8]$. Since $\mathbf{\Sigma}_u^{[1,1]}$, $\mathbf{\Sigma}_u^{[2,2]}$, and $\mathbf{R}_u$ are diagonal, the product and square root above are taken elementwise on the diagonal entries.

For each segment u, we draw a mean $\mu_u \in \mathbb{R}^{2N_s}$ from a uniform distribution $[-4, 4]$, and we draw N_o independent and identically distributed (i.i.d.) samples $\mathbf{s}^{[u]} \sim \mathcal{N}(\mu_u, \mathbf{\Sigma}_u)$, where $\mathbf{\Sigma}_u = \begin{bmatrix} \mathbf{\Sigma}_u^{[1,1]} & \mathbf{\Sigma}_u^{[1,2]} \\ \mathbf{\Sigma}_u^{[2,1]} & \mathbf{\Sigma}_u^{[2,2]} \end{bmatrix}$.

A multilayer perceptron with $L = 2$ layers acts as a nonlinear mixing function $\mathbf{f}^{[m]}$ for each modality m. At each layer l, we apply a Leaky ReLU (Maas et al., 2013) with a negative slope of 0.2, followed by a random orthogonal matrix $\mathbf{A}^{[m,l]}$. Composing across the L layers for each modality yields the final observed dataset $\mathbf{X}$.

For each segment in the synthetic dataset, we split the observations into training (78), validation (26), and test (26) sets using a $60\% : 20\% : 20\%$ ratio. The training set is used to fit the model, the validation set to select the number of epochs, and performance is evaluated on both the training and test sets.

B.2 RESULTS

Fig. A1 shows the aggregated RDC matrices across segments on the test set. Consistent with the training set result (Fig. 3a), DeepIVA outperforms existing methods in both unimodal source identification (M_1 MCC: 0.99; M_2 MCC: 0.99) and cross-modal source alignment (MCC: 0.94).

Fig. A2 reports the average MCC and MD across eight synthetic test sets, each generated from a different random seed. Consistent with the training set result (Fig. 3b), DeepIVA shows improved performance with an aggregated average MCC of 0.99 ($4.21\% - 11.24\%$ higher than baselines) and an aggregated average MD of 3.60 ($1.37\% - 3.23\%$ lower than baselines).

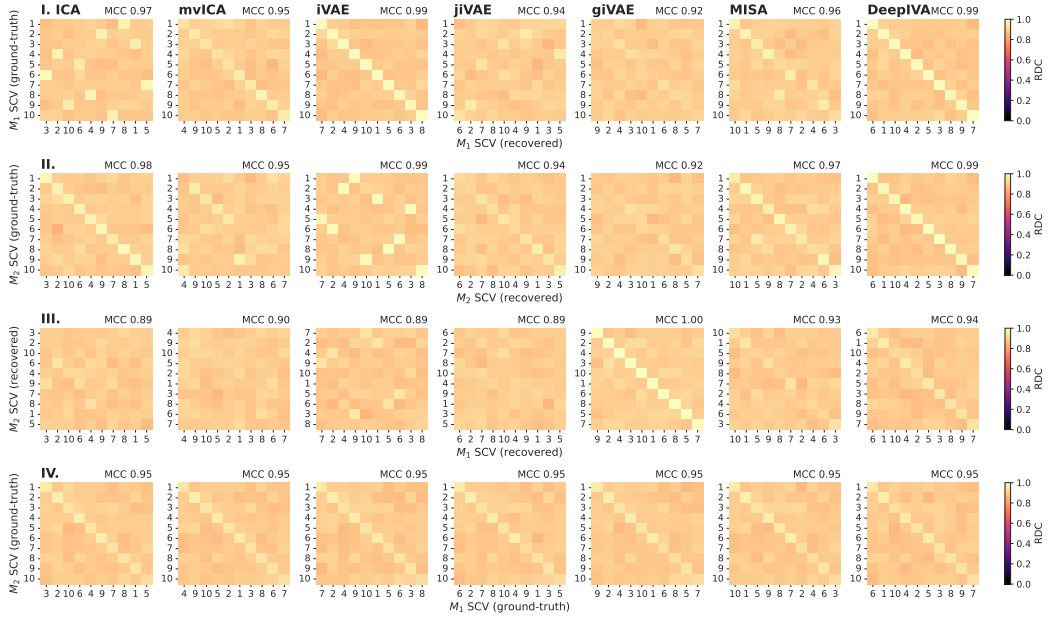


Figure A1: **Aggregated RDC matrices across segments from a synthetic dataset** (10 **sources,** 22 **segments,** 26 **samples per segment, test set).** Consistent with the training set result (Fig. 3a), DeepIVA outperforms existing methods in both unimodal source identification and cross-modal source alignment.

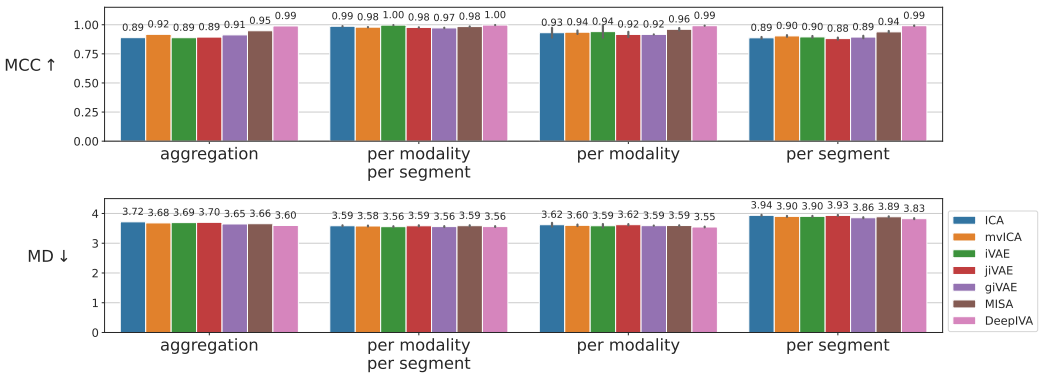


Figure A2: **MCC (↑) and MD (↓) between recovered sources and ground-truth sources for seven approaches (ICA, mvICA, iVAE, jiVAE, giVAE, MISA, DeepIVA), shown for one simulation configuration** (10 **sources,** 22 **segments,** 26 **samples per segment, test set).** The bar plot shows the mean and 95% confidence interval across 8 random seeds. Consistent with the training set result (Fig. 3b), DeepIVA shows superior performance for the aggregated metric which accounts for unimodal identifiability, cross-modal linkage, and cross-segment consistency.

C HYPERPARAMETER OPTIMIZATION

C.1 EXPERIMENT DESIGN

Empirical observations suggest that dynamically adjusting the learning rate allows DeepIVA to converge to the optimal solution more quickly and accurately. We therefore propose a hyperparameter optimization strategy for efficiently tuning the learning rate and optimizer parameters. The iVAE learning rate is fixed at 10^{-3}, while the MISA learning rate is tuned via three parameters: the initial learning rate lr_0, the target learning rate lr_t, and the decay slope $\mathrm{T}_\mathrm{decay}$ (the number of epochs between successive 10^{-5} reductions in learning rate). We additionally optimize the Adam optimizer parameters β_1 and β_2, which control the running averages of the gradient and its squared value, respectively.

We estimate the optimal hyperparameters by fitting the following quadratic function:

$$y = \mathbf{x}^\top \mathbf{A}\mathbf{x} + \mathbf{b}^\top \mathbf{x} + c, \tag{A14}$$

where $y \in \mathbb{R}$ denotes the response scalar derived from unimodal or cross-modal MCC, $\mathbf{x} = [\mathrm{lr}_0 \quad \mathrm{lr}_\mathrm{t} \quad \mathrm{T}_\mathrm{decay} \quad \beta_1 \quad \beta_2]^\top \in \mathbb{R}^5$ denotes the corresponding predictor vector, and $\mathbf{A} \in \mathbb{R}^{5\times5}$, $\mathbf{b} \in \mathbb{R}^5$, $c \in \mathbb{R}$ are the coefficients that characterize the quadratic, linear, and constant relationships between the predictors and the response, respectively.

We evaluate three values for each of five hyperparameters:

- lr_0: $0.8 \times 10^{-4}, 1.15 \times 10^{-4}, 1.5 \times 10^{-4}$
- lr_t: $0.1 \times 10^{-5}, 4 \times 10^{-5}, 7.9 \times 10^{-5}$
- $\mathrm{T}_\mathrm{decay}$: 10, 25, 40
- Adam β_1: 0.9625, 0.9812, 0.9999
- Adam β_2: 0.6, 0.7, 0.8

We generate 8 synthetic datasets with the same configuration ($N_m = 2, N_s = 10, N_o = 130, N_u = 22$) but different random seeds ($0 - 7$). For each of the five hyperparameters (lr_0, lr_t, $\mathrm{T}_\mathrm{decay}$, β_1, and β_2), we let -1, 0, or 1 denote its three candidate values in ascending order. We additionally encode the 8 random seeds using a 3-bit binary representation; for example, $(0, 0, 1)$ denotes seed 1. To recover actual hyperparameter values from continuous coded values produced by the fitted quadratic function, we define, for each hyperparameter j, $v_j(-1)$, $v_j(0)$, $v_j(1)$ as its three candidate values in ascending order. A continuous coded value $x_j \in [-1, 1]$ is then mapped to its actual value by piecewise linear interpolation between adjacent levels:

$$v_j(x_j) = \begin{cases} v_j(-1) + (x_j + 1)\left(v_j(0) - v_j(-1)\right), & x_j \in [-1, 0] \\ v_j(0) + x_j\left(v_j(1) - v_j(0)\right), & x_j \in [0, 1] \end{cases} \tag{A15}$$

Using definitive screening design (DSD) (Jones & Nachtsheim, 2011), we construct 30 experimental configurations evaluated on 8 synthetic datasets, including 2 zero-parameter configurations ($\mathbf{x} = [0 \quad 0 \quad 0 \quad 0 \quad 0]^\top$) that cover 2 of the 8 random seeds. To cover the remaining 6 seeds, we add 6 additional zero-parameter configurations, yielding 36 configurations in total. Each configuration is trained for 1000 epochs, with unimodal MCC (per modality) and cross-modal MCC recorded on the training, validation, and test sets every 5 epochs.

Prior to fitting the quadratic function, we perform an initial screening step using a linear regression model:

$$y = \mathbf{b}_0^\top \tilde{\mathbf{x}}, \tag{A16}$$

where $\tilde{\mathbf{x}} = \left[\mathrm{lr}_0 \quad \mathrm{lr}_\mathrm{t} \quad \mathrm{T}_\mathrm{decay} \quad \beta_1 \quad \beta_2 \quad \mathrm{lr}_0^2 \quad \mathrm{lr}_\mathrm{t}^2 \quad (\mathrm{T}_\mathrm{decay})^2 \quad \beta_1^2 \quad \beta_2^2\right]^\top$ is the expanded predictor vector including all linear and quadratic main effects. We compute R^2 and remove the main effect with the largest p-value if its linear and quadratic terms are both statistically insignificant ($p > 0.1$). We then perform stepwise regression on the remaining terms, including four linear terms, four quadratic terms, and all pairwise interaction terms. We iteratively remove the interaction or quadratic term with the largest p-value until all remaining terms are significant ($p < 0.1$).

Finally, we fit the quadratic function using all significant terms and derive optimal hyperparameters for each of six candidate response metrics: the unimodal, cross-modal, and combined MCCs, together with their normalized counterparts (denoted with a prime), where normalization is performed relative to the maximum value attained across the 36 experimental configurations. Let $\text{MCC}^{[1]}$, $\text{MCC}^{[2]}$, and $\text{MCC}^{[1,2]}$ denote the unimodal and cross-modal MCC for configuration $i \in \{1, \ldots, 36\}$:

$$y_{\text{um}} = \min\left(\text{MCC}^{[1]}, \text{MCC}^{[2]}\right), \tag{A17}$$

$$y'_{\text{um}} = \frac{\min\left(\text{MCC}^{[1]}, \text{MCC}^{[2]}\right)}{\max\limits_{i\in\{1,\ldots,36\}}\left(\min\left(\text{MCC}^{[1]}, \text{MCC}^{[2]}\right)\right)}, \tag{A18}$$

$$y_{\text{cm}} = \text{MCC}^{[1,2]}, \tag{A19}$$

$$y'_{\text{cm}} = \frac{\text{MCC}^{[1,2]}}{\max\limits_{i\in\{1,\ldots,36\}}\left(\text{MCC}^{[1,2]}\right)}, \tag{A20}$$

$$y_{\text{comb}} = \log_{10}\left(1 - \min\left(\text{MCC}^{[1]}, \text{MCC}^{[2]}\right)\right) + \log_{10}\left(1 - \text{MCC}^{[1,2]}\right). \tag{A21}$$

$$y'_{\text{comb}} = \log_{10}\left(1 - \frac{\min\left(\text{MCC}^{[1]}, \text{MCC}^{[2]}\right)}{\max\limits_{i\in\{1,\ldots,36\}}\left(\min\left(\text{MCC}^{[1]}, \text{MCC}^{[2]}\right)\right)}\right) + \log_{10}\left(1 - \frac{\text{MCC}^{[1,2]}}{\max\limits_{i\in\{1,\ldots,36\}}\left(\text{MCC}^{[1,2]}\right)}\right). \tag{A22}$$

C.2 Results

During initial screening, β_1 exhibits the largest p-value among the five hyperparameters and is therefore excluded from subsequent analysis; it is instead fixed at the median value of the tested range ($\beta_1 = 0.9812$). Table A1 reports the combined, unimodal (minimum across the two modalities), and cross-modal metrics for each hyperparameter configuration. The optimal solution estimated from the fitted quadratic function is $\mathbf{x}^* = [-1, -0.2, 1, 1]$, corresponding to lr_0, lr_t, T_{decay}, and β_2, respectively. Mapping these coded values to their actual values via Eq. A15 yields $\text{lr}_0 = 0.8 \times 10^{-4}$, $\text{lr}_t = 3.22 \times 10^{-5}$, $\text{T}_{\text{decay}} = 40$, and $\beta_2 = 0.8$ (contour maps shown in Fig. A3); together with $\beta_1 = 0.9812$, these constitute the final optimal hyperparameters used to train DeepIVA in both the synthetic and neuroimaging data experiments.

Table A1: **Hyperparameter optimization configurations and metrics.** Each row reports a DSD configuration's hyperparameter values and the resulting unimodal, cross-modal, and combined MCC metrics; normalized values are shown with raw MCC values in parentheses.

$\mathbf{lr}_0$ $(\times 10^{-4})$	$\mathbf{lr}_t$ $(\times 10^{-5})$	$\mathbf{T}_{\mathbf{decay}}$	β_1	β_2	$y'_{\mathbf{um}}$ ($y_{\mathbf{um}}$) ↑	$y'_{\mathbf{cm}}$ ($y_{\mathbf{cm}}$) ↑	$y'_{\mathbf{comb}}$ ($y_{\mathbf{comb}}$) ↓
1.15	4	25	0.9812	0.7	0.9997 (0.9726)	0.9908 (0.7330)	-5.5001 (-2.1355)
1.15	4	25	0.9812	0.7	0.9814 (0.9549)	0.9997 (0.7395)	-5.1928 (-1.9296)
0.8	0.1	40	0.9999	0.6	0.9921 (0.9652)	0.9972 (0.7377)	-4.6615 (-2.0400)
0.8	4	40	0.9625	0.8	0.9975 (0.9705)	0.9875 (0.7305)	-4.4980 (-2.0989)
1.15	4	25	0.9812	0.7	0.9920 (0.9651)	0.9947 (0.7358)	-4.3734 (-2.0359)
1.15	4	25	0.9812	0.7	0.9971 (0.9701)	0.9719 (0.7190)	-4.0855 (-2.0753)
1.5	0.1	10	0.9999	0.6	0.9932 (0.9663)	0.9728 (0.7196)	-3.7303 (-2.0242)
1.5	4	10	0.9999	0.6	0.9904 (0.9635)	0.9578 (0.7085)	-3.3902 (-1.9736)
0.8	0.1	40	0.9812	0.8	0.9909 (0.9641)	0.9528 (0.7048)	-3.3669 (-1.9745)
1.5	0.1	40	0.9999	0.8	0.9864 (0.9597)	0.9551 (0.7066)	-3.2147 (-1.9272)
0.8	7.9	10	0.9999	0.8	0.9538 (0.9280)	0.9827 (0.7270)	-3.0975 (-1.7064)
0.8	7.9	40	0.9625	0.6	0.9447 (0.9191)	0.9855 (0.7290)	-3.0963 (-1.6592)
1.5	0.1	25	0.9625	0.8	0.9910 (0.9642)	0.8991 (0.6651)	-3.0435 (-1.9213)
1.15	4	25	0.9812	0.7	0.9729 (0.9466)	0.9614 (0.7112)	-2.9806 (-1.8117)
1.15	4	25	0.9812	0.7	0.9224 (0.8975)	0.9861 (0.7295)	-2.9677 (-1.5569)
1.15	4	25	0.9812	0.7	0.9669 (0.9408)	0.9660 (0.7146)	-2.9494 (-1.7720)
1.5	7.9	10	0.9625	0.8	0.9471 (0.9214)	0.9779 (0.7234)	-2.9310 (-1.6629)
1.5	0.1	40	0.9625	0.7	0.9840 (0.9574)	0.9226 (0.6825)	-2.9072 (-1.8685)
1.5	0.1	10	0.9999	0.8	0.9892 (0.9624)	0.8812 (0.6519)	-2.8928 (-1.8836)
0.8	0.1	10	0.9625	0.8	0.9905 (0.9637)	0.8576 (0.6345)	-2.8678 (-1.8767)
0.8	0.1	10	0.9999	0.8	0.9924 (0.9656)	0.8061 (0.5964)	-2.8337 (-1.8570)
0.8	0.1	10	0.9625	0.6	0.9918 (0.9650)	0.8112 (0.6001)	-2.8124 (-1.8539)
0.8	0.1	40	0.9999	0.6	0.9799 (0.9534)	0.9063 (0.6705)	-2.7257 (-1.8137)
1.15	0.1	10	0.9625	0.6	0.9827 (0.9561)	0.8668 (0.6413)	-2.6371 (-1.8026)
1.15	4	25	0.9812	0.7	0.9587 (0.9327)	0.9423 (0.6971)	-2.6224 (-1.6908)
1.5	0.1	40	0.9625	0.6	0.9803 (0.9537)	0.8743 (0.6468)	-2.6056 (-1.7867)
0.8	7.9	10	0.9625	0.6	0.8740 (0.8504)	0.9253 (0.6845)	-2.0264 (-1.3261)
0.8	7.9	40	0.9625	0.8	0.8069 (0.7851)	0.9487 (0.7018)	-2.0037 (-1.1932)
0.8	7.9	10	0.9999	0.7	0.8316 (0.8091)	0.9403 (0.6956)	-1.9976 (-1.2358)
0.8	7.9	25	0.9999	0.6	0.8371 (0.8145)	0.9381 (0.6940)	-1.9964 (-1.2459)
1.5	7.9	10	0.9625	0.8	0.7602 (0.7397)	0.9386 (0.6943)	-1.8319 (-1.0993)
1.15	7.9	40	0.9999	0.8	0.7616 (0.7410)	0.9247 (0.6841)	-1.7460 (-1.0872)
1.5	7.9	40	0.9625	0.6	0.8064 (0.7846)	0.8938 (0.6612)	-1.6869 (-1.1368)
1.5	7.9	10	0.9812	0.6	0.7817 (0.7606)	0.8889 (0.6576)	-1.6153 (-1.0863)
1.5	7.9	40	0.9999	0.6	0.7581 (0.7376)	0.8963 (0.6630)	-1.6003 (-1.0535)
1.5	7.9	40	0.9999	0.8	0.7473 (0.7272)	0.8739 (0.6465)	-1.4966 (-1.0157)

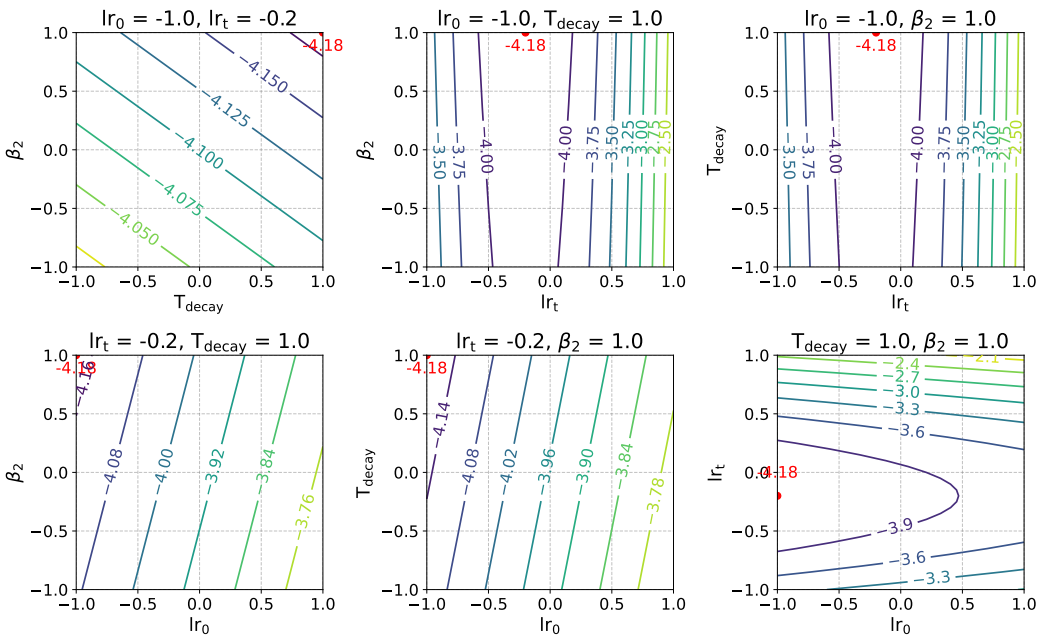


Figure A3: **Contour maps showing response surface of optimal hyperparameters.** The optimal solution estimated from the fitted quadratic function is $\mathbf{x}^* = [-1, -0.2, 1, 1]$, corresponding to lr_0, lr_t, T_{decay}, and β_2, respectively. Mapping these coded values to their actual values via Eq. A15 yields $\text{lr}_0 = 0.8 \times 10^{-4}$, $\text{lr}_t = 3.22 \times 10^{-5}$, $\text{T}_{\text{decay}} = 40$, and $\beta_2 = 0.8$; together with $\beta_1 = 0.9812$, these constitute the final optimal hyperparameters used to train DeepIVA in both the synthetic and neuroimaging data experiments.

D NEUROIMAGING DATA EXPERIMENT

D.1 DATA PREPROCESSING

D.1.1 STRUCTURAL MRI

T1-weighted structural MRI (sMRI) scans are acquired using a 3D MPRAGE sequence with the following parameters: repetition time (TR) $=$ 2000ms, inversion time (TI) $=$ 880ms, in-plane acceleration factor $= 2$, voxel size $= 1 \times 1 \times 1\text{mm}^3$, and acquisition matrix $= 208 \times 256 \times 256$.

The sMRI preprocessing steps include tissue segmentation and normalization to the Montreal Neurological Institute (MNI) template using SPM12 (`http://www.fil.ion.ucl.ac.uk/spm/`) (Ashburner et al., 2014), yielding gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) tissue probability maps. The normalized GM tissue probability maps are then spatially smoothed using a 10mm full width at half maximum (FWHM) Gaussian kernel and resampled to $3 \times 3 \times 3\text{mm}^3$ resolution. We define a group-level GM mask to restrict the analysis to GM voxels. Specifically, the normalized GM tissue probability maps at $1 \times 1 \times 1\text{mm}^3$ resolution are averaged across all subjects to produce a group-average GM map. This map is binarized using a threshold of 0.2 and resampled to $3 \times 3 \times 3\text{mm}^3$ resolution, resulting in a binary mask of 44318 voxels.

D.1.2 FUNCTIONAL MRI

Resting-state functional MRI (fMRI) scans are acquired with the following parameters: TR $=$ 735ms, echo time (TE) $=$ 39ms, multiband factor $= 8$, in-plane acceleration factor $= 1$, flip angle $= 52°$, voxel size $= 2.4 \times 2.4 \times 2.4\text{mm}^3$, and acquisition matrix $= 88 \times 88 \times 64$.

We utilize distortion-corrected, FIX-denoised (Griffanti et al., 2014) and normalized fMRI data from the UK Biobank data resource. For each subject, the resulting voxel-wise time series is bandpass filtered ($0.01 - 0.08$Hz) to remove low-frequency drift and physiological noise, then transformed to the frequency domain via fast Fourier transform (FFT). The amplitude of low-frequency fluctuations (ALFF) is computed as the mean square root of the power spectrum across $0.01 - 0.08$Hz (Zang et al., 2007). To improve test-retest reliability (Zhao et al., 2018), a subject-specific mean-scaled ALFF (mALFF) map is obtained by dividing each subject's ALFF map by its global mean ALFF value. The mALFF maps are then smoothed using a 6mm FWHM Gaussian filter and resampled to $3 \times 3 \times 3\text{mm}^3$ resolution. Finally, the same group-average GM mask is applied to the mALFF maps, resulting in 44318 voxels.

D.2 SAMPLE DISTRIBUTIONS

Fig. A4 shows sex and age distributions across the training, validation, and test sets. Fig. A5 shows sample counts across segments for each of the training, validation, and test sets. Table A2 lists each segment index alongside its designated sex, lower and upper age boundaries, and sample counts across the training, validation, and test sets. Distributions are approximately consistent across all three sets.

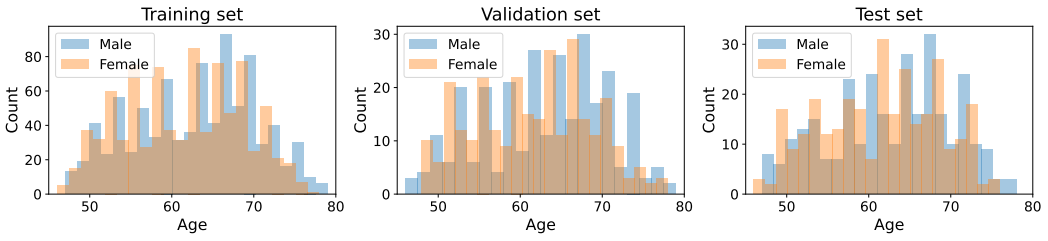


Figure A4: **Age and sex distributions across the training, validation, and test sets.**

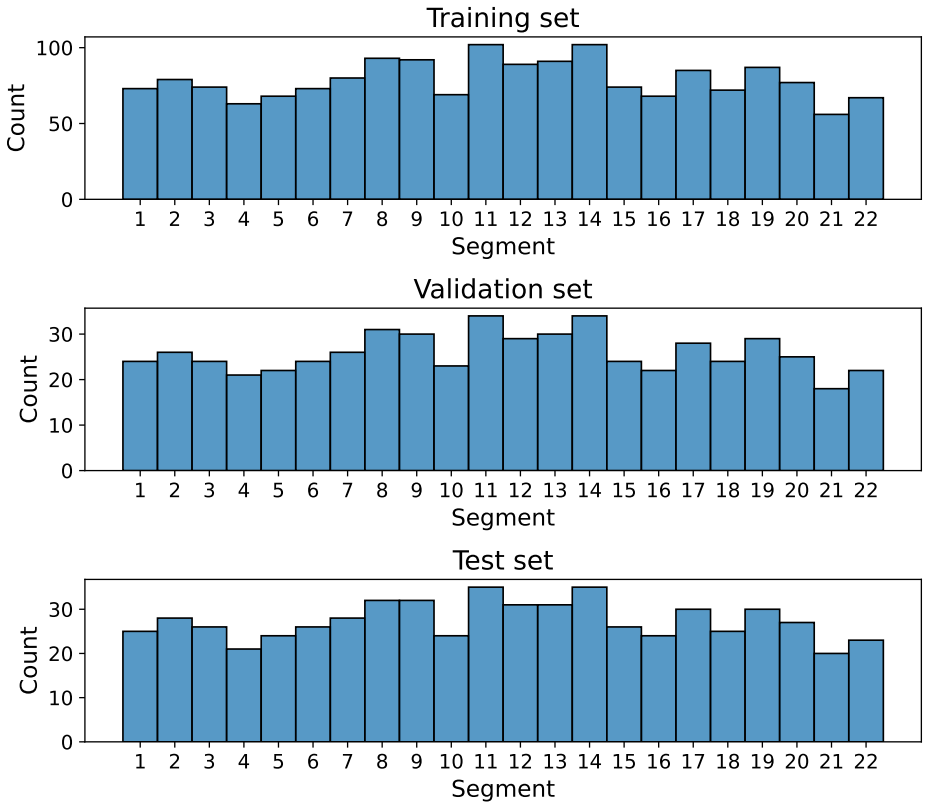


Figure A5: **Sample counts across segments for each of the training, validation, and test sets.**

Table A2: **Segment information for age- and sex-stratified subgroups.** Abbreviations: M, male; F, female; N_{train}, N_{valid}, N_{test}, number of samples in the training, validation, and test sets, respectively.

Segment index (u)	**Sex**	**Age range (yr)**	N_{train}	N_{valid}	N_{test}
1	M	46-51	73	24	25
2	M	51-54	79	26	28
3	M	54-57	74	24	26
4	M	57-59	63	21	21
5	M	59-61	68	22	24
6	M	61-63	73	24	26
7	M	63-65	80	26	28
8	M	65-67	93	31	32
9	M	67-69	92	30	32
10	M	69-71	69	23	24
11	M	71-79	102	34	35
12	F	46-51	89	29	31
13	F	51-54	91	30	31
14	F	54-57	102	34	35
15	F	57-59	74	24	26
16	F	59-61	68	22	24
17	F	61-63	85	28	30
18	F	63-65	72	24	25
19	F	65-67	87	29	30
20	F	67-69	77	25	27
21	F	69-71	56	18	20
22	F	71-79	67	22	23

D.3 ORDER SELECTION

To select the model order (the number of latent sources), we evaluate three information-theoretic criteria (Li et al., 2007) based on the singular values of the multimodal group principal component analysis (MGPCA) covariance matrix derived from the multimodal neuroimaging dataset: the Akaike information criterion (AIC) (Akaike, 1998), the Kullback-Leibler information criterion (KIC) (Cavanaugh, 1999), and the minimum description length (MDL) (Rissanen, 1978).

According to Fig. A6, AIC, KIC, and MDL select the orders of 25, 18, and 10 for the full dataset, and 17, 12, and 7 for the training set. The curves flatten at 10 or more sources, indicating that 10 sources already capture the majority of the information. Therefore, 10 sources are recovered for each neuroimaging modality.

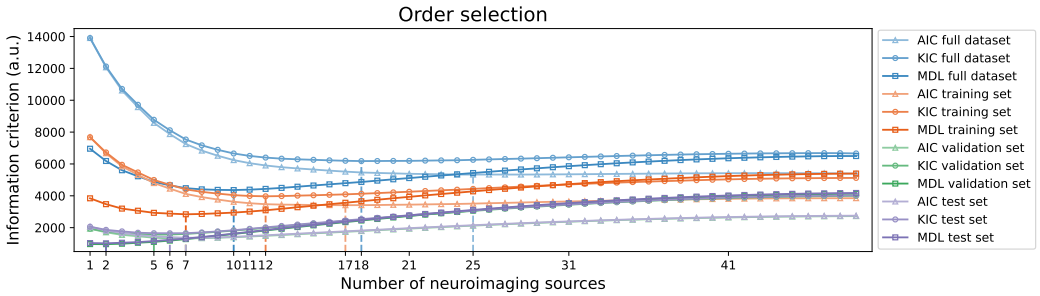


Figure A6: **Order selection for multimodal neuroimaging data.** AIC, KIC, and MDL select the orders of 25, 18, and 10 for the full dataset. The curves flatten at 10 or more sources, suggesting that 10 sources capture the majority of the information; thus, 10 sources are recovered for each modality.

D.4 BRAIN SPATIAL MAP RECONSTRUCTION

Prior to spatial map reconstruction, each source $\mathbf{s}_i^{[m]} \in \mathbb{R}^N$ (where N is the total number of subjects) is normalized by dividing its standard deviation across the subject dimension. The normalized source is then partitioned into segments, indexed by u, each containing a subset of N_o^u observations (subjects), such that $N = \sum_u N_o^u$.

For each segment u, the spatial map is obtained via least-squares regression of the original (unnormalized) m-th modality onto the corresponding segment of the normalized source:

$$\mathbf{a}_i^{[m,u]} = \left({\mathbf{s}_i}^{[m,u]\top} \mathbf{s}_i^{[m,u]} \right)^{-1} {\mathbf{s}_i}^{[m,u]\top} \mathbf{X}^{[m,u]}, \tag{A23}$$

where $\mathbf{a}_i^{[m,u]} \in \mathbb{R}^{N_x}$ is the resulting spatial weight vector, $\mathbf{s}_i^{[m,u]} \in \mathbb{R}^{N_o^u}$ is the segment-u portion of the normalized source, $\mathbf{X}^{[m,u]} \in \mathbb{R}^{N_o^u \times N_x}$ is the original modality-m data matrix for segment u (rows: observations, columns: feature values), and N_x is the number of features (voxels). The vector $\mathbf{a}_i^{[m,u]}$ is then reshaped into a 3D tensor using the gray matter mask's voxel coordinates and visualized in 3D space.

The spatial maps of SCVs 1 and 7 for the test set (Fig. A7) show several patterns consistent with those observed for the training set (Fig. 4b). For sex-related SCV 1, females relative to males show stronger positive intensity in the cerebellum and occipital pole, and similar negative intensity in the inferior temporal gyrus in sMRI; they also show similar negative activation in the subcallosal cortex and stronger negative activation in the occipital pole in fMRI. For age-related SCV 7, younger subjects show slightly stronger negative intensity in the insular cortex and precuneus cortex in sMRI. Additionally, younger subjects show increased negative activation in the lateral occipital cortex in fMRI.

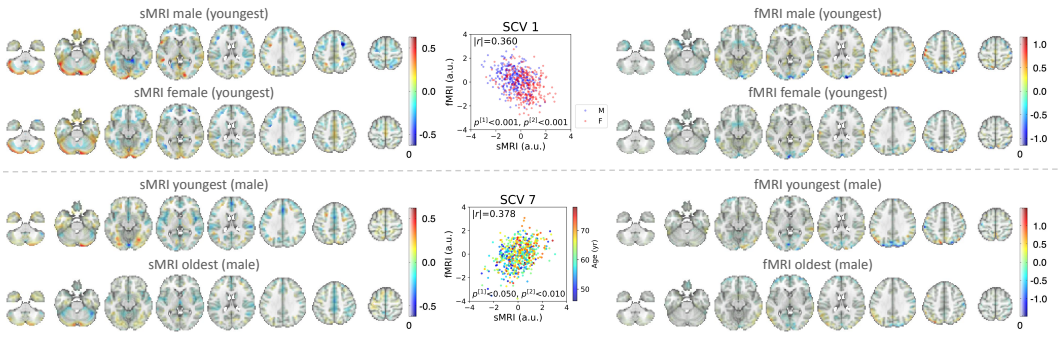


Figure A7: **DeepIVA SCVs associated with sex (top) and age (bottom) for the test set.** The Pearson correlation coefficient (r) indicates the cross-modal source correlation. The p-value indicates group differences for sMRI ($p^{[1]}$) and fMRI ($p^{[2]}$) sources separately (two-sample t-test; Bonferroni-correction for 20 comparisons). The axial slices show the average reconstructed segment-specific data across subjects for each modality, with voxel intensity mapped to both color hue and opacity.

E EVALUATION METRICS

Here we describe how MCC and MD are computed in detail. We first compute the RDC matrix $\mathbf{R}$ between two sets of sources (e.g., recovered sources and ground-truth sources; estimated sMRI sources and estimated fMRI sources) for each modality and each segment. Note that we compute an RDC matrix for *each* segment separately, instead of computing it across all segments by convention. Our segment-specific RDC can more precisely characterize the data within each segment and effectively mitigate the noise introduced when all segments are taken simultaneously. Next, we aggregate the RDC matrices over segments by taking the mean to obtain an RDC matrix $\mathbf{R}^{[m]}$ *per modality* (mean aggregation). We also obtain an aggregated RDC matrix $\mathbf{R}^{[u]}$ *per segment* by taking the minimum across modalities for the entries corresponding to the sorted indices (i.e., the entries along the main diagonal after sorting) from a linear sum assignment problem (LSAP) solver (Crouse, 2016), and then taking the maximum for the remaining entries across all modalities (min-max aggregation). This min-max aggregation penalizes approaches that fail to detect cross-modal linkage, even when unimodal identifiability is high. To compute the final aggregated RDC matrix, we use min-max aggregation of $\mathbf{R}^{[m]}$ across modalities. We use the permuted indices from the modality-specific RDC matrix $\mathbf{R}^{[m]}$ which yields the lowest MD value as the *global* sorting indices to sort the other RDC matrices. For each sorted RDC matrix $\mathbf{R}_s$, we compute the MCC between the 25th percentile and the 75th percentile, as well as the MD, slightly adjusted from Eq. 4 in Nordhausen et al. (2011):

$$\mathrm{MD}\left(\mathbf{R}\right) = \frac{1}{2} \cdot \left(1 + \frac{1}{d} \cdot \mathrm{trace}\left(\mathbf{R}\mathbf{R}^{\top}\right) - \frac{2}{d} \cdot \mathrm{trace}\left(\mathbf{R}_s\right)\right), \tag{A24}$$

where $\mathbf{R}$ is the unsorted matrix, $\mathbf{R}_s$ is the sorted matrix, and d is the dimension of $\mathbf{R}$.