Multi-View Independent Component Analysis with Shared and Individual Sources

Abstract

Independent component analysis (ICA) is a blind source separation method for linear disentanglement of independent latent sources from observed data. We investigate the special setting of noisy linear ICA where the observations are split among different views, each receiving a mixture of shared and individual sources. We prove that the corresponding linear structure is identifiable, and the sources distribution can be recovered. To computationally estimate the sources, we optimize a constrained form of the joint log-likelihood of the observed data among all views. We also show empirically that our objective recovers the sources also in the case when the measurements are corrupted by noise. Furthermore, we propose a model selection procedure for recovering the number of shared sources which we verify empirically. Finally, we apply the proposed model in a challenging real-life application, where the estimated shared sources from two large transcriptome datasets (observed data) provided by two different labs (two different views) lead to recovering (shared) sources utilized for finding a plausible representation of the underlying graph structure.

1 INTRODUCTION

Independent component analysis (ICA) is a method for solving blind source separation (BSS) problems [Comon, 1994] where the goal is to separate independent latent sources from mixed observed signals and, thus, uncover essential structures in various data types. Historically, linear ICA has proven to be a successful approach for recovering spatially independent sources representing brain activity regions from magnetoencephalography (MEG) data [Vigário et al., 1997] or functional MRI (fMRI) data [McKeown and Sejnowski, 1998]. The utility of ICA is not only limited to neuroscience, but it has a wide range of applications in omics data analysis, e.g. [Zheng et al., 2008, Nazarov et al., 2019, Zhou and Altman, 2018, Tan et al., 2020, Urzúa-Traslaviña et al., 2021, Rusan et al., 2020, Cary et al., 2020, Dubois et al., 2019, Aynaud et al., 2020]. In these works, the interpretation of the latent sources relies on the assumption that each experimental outcome is a linear mixture of independent biological processes (the sources). For example, the latent sources could represent gene profiles that are used to predict gene regulation [Sastry et al., 2021, 2019] or cell-type specific expressions from tumor samples [Avila Cobos et al., 2018] for studying cell-type decompositions in cancer research.

The fast advancement of technology in the biomedical domain has provided a unique opportunity to find valuable insights from large-scale data integration studies. Many of these applications can be transformed into multiview BSS problems. A significant body of research has been devoted to developing multiview ICA methods focused on unraveling group-level (shared) brain activity patterns in multi-subject fMRI and EEG datasets [Salman et al., 2019, Huster et al., 2015, Congedo et al., 2010, Durieux et al., 2019, Congedo et al., 2010, Calhoun et al., 2001]. However, these methods cannot be applied directly to problems where one is interested in retrieving both shared and view-specific signals, e.g. investigating the individual-specific brain functions (view-specific) and shared phenotypes patterns in individuals' brain activity in a natural stimuli experiment [Dubois et al., 2016, Bartolomeo et al., 2017]. Another application, where the estimation of both shared and view-specific sources is essential, is omics data integration. A typical example is combining heterogeneous gene expression data sets for achieving better gene regulation discovery. In this scenario, the observed samples are realizations of diverse and complex experiments. The shared information between the datasets refer to genes with stable expression across almost all conditions and the individual signals represent experiment-specific gene activities such as measurements of gene knock-outs, stress conditions, etc.

Submitted to the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023). **To be used for reviewing only**.

Summary. To address these and similar scientific applications, we formalize the described multi-view BSS problem as a linear noisy generative model for a multi-view data regime, assuming that the mixing matrix and number of individual sources are view-specific. We call the resulting model, ShIndICA. By requiring that the sources are non-Gaussian and mutually independent and the linear mixing matrices have full column rank, we provide identifiability guarantees for the mixing matrices and latent sources in distribution. We maximize the joint log-likelihood of the observed views to estimate the mixing matrices. Furthermore, we provide a model selection criterion for selecting the correct number of shared sources. Finally, we apply ShIndICA on a data integration problem of two large transcriptome datasets. We show empirically that our method works well compared to the baselines when the estimated components are used for a graph inference task.

Contributions. Our contributions can be summarized as follows:

1. We propose a new multi-view generative BSS model with shared and individual sources, called ShIndICA.
2. We provide theoretical guarantees for the identifiability of the recovered linear structure and the source and noise distributions.
3. We derive the closed form joint likelihood of ShIndICA which is used for estimating the mixing matrices.
4. We propose a selection criterion for inferring the correct number of shared sources derived from the generative model assumptions.

2 PROBLEM FORMALIZATION

Consider the following D-view multivariate linear BSS model where for $d \in \{1, \dots, D\}$

$$\boxed{x_d = A_d(\tilde{s}_d + \epsilon_d) = A_{d0}s_0 + A_{d1}s_d + A_d\epsilon_d,} \tag{1}$$

and it holds that

1. $x_d \in \mathbb{R}^{k_d}$ is a random vector with $\mathbb{E}[x_d] = 0$,
2. $\tilde{s}_d = (s_0^\top, s_d^\top)^\top$ are latent non-Gaussian random sources with $s_0 \in \mathbb{R}^c$ and $s_d \in \mathbb{R}^{k_d - c}$ being the shared and individual sources and $\mathbb{E}[\tilde{s}_d] = 0$ and $\mathrm{Var}[\tilde{s}_d] = \mathbb{I}_{k_d}$,
3. $A_d \in \mathbb{R}^{k_d \times k_d}$ is a mixing matrix with full column rank, A_{d0} and A_{d1} are the columns corresponding to the shared and individual sources,
4. $\epsilon_d \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_{k_d})$ is Gaussian noise,
5. all latent source components and noise variables are mutually independent.

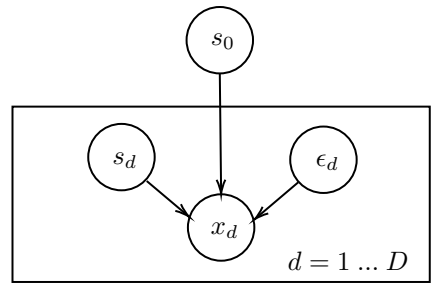

Figure 1: A graphical representation of Equation 1 where x_d is the observed variable, s_0 denotes the shared sources, s_d the view-specific ones and ϵ_d is the Gaussian noise.

Note that for $D = 1$ the model becomes a standard linear ICA model which is solved by Comon [1994], Hyvärinen and Oja [2000], Bell and Sejnowski [1995] for independent non-Gaussian latent sources $z := \tilde{s}_1 + \epsilon_1$. The Gaussian noise in Equation 1 can be interpreted as a measurement error on the device with variance $\sigma^2 A_d A_d^\top$ (similarly to [Richard et al., 2020, 2021]). We choose this setting compared to the $A_d\tilde{s}_d + \epsilon_d$ because we can derive a joint data likelihood in a closed form (see Section 4) which is not available in the latter representation. Moreover, assumption 5 implies that the noise is not expected to influence the true signal and vice versa which is a common assumption in measurement error models known as classical errors. See Figure 1 for a graphical representation of Equation 1.

3 IDENTIFIABILITY RESULTS

In unsupervised machine learning methods, the reliability of the algorithm cannot be directly verified outside of simulations due to the non-existence of labels. For this reason, theoretical guarantees are necessary to trust that the algorithm estimates the quantities of interest. For a BSS problem solution, such as ICA, we want the sources and mixing matrices to be (up to certain equivalence relations) unambiguously determined (or *identifiable*) by the data, at least in the large sample limit.

Identifiability results for noiseless single-view ICA are proved by [Comon, 1994]. It turns out that if at most one of the latent sources is normal and the mixing matrix is invertible, then both the mixing matrix and sources can be recovered *almost surely* up to permutation, sign and scaling. However, this result does not hold in the general additive noise setting. Davies [2004] shows that if the mixing matrix has a full column rank, then the structure is identifiable, but not the latent sources.

By employing the multi-view ($D \geq 2$) noisy setting inspired from our model (see Equation 1), we extend the results by Comon [1994], Davies [2004], Kagan et al. [1973], Richard et al. [2020]. Compared to previous work, we provide identifiability guarantees not only for the mixture ma-

trices up to sign and permutation, but also for the *source and noise distributions* (up to the same sign and permutation), and the latent (both shared and individual) sources dimensions[1]. Moreover, our identifiability results hold for a more general case than Equation 1 since the noise distribution can be view-specific, and the mixing matrices can be non-square. This is stated in the following Theorem 3.1, proved in Appendix A:

Theorem 3.1. *Let $x_1, \dots, x_D$ for $D \geq 2$ be random vectors with the following two representations:*

$$A_d^{(1)}\Big(\begin{bmatrix} s_0^{(1)} \\ s_d^{(1)} \end{bmatrix} + \epsilon_d^{(1)}\Big) = x_d = A_d^{(2)}\Big(\begin{bmatrix} s_0^{(2)} \\ s_d^{(2)} \end{bmatrix} + \epsilon_d^{(2)}\Big),$$

where $d \in \{1, \dots, D\}$, with the following properties for $i = 1, 2$

1. *$A_d^{(i)} \in \mathbb{R}^{p_d \times k_d^{(i)}}$ is a (non-random) matrix with full column rank, i.e. $\operatorname{rank}(A_d^{(i)}) = k_d^{(i)}$,*
2. *$\epsilon_d^{(i)} \in \mathbb{R}^{k_d^{(i)}}$ and $\epsilon_d^{(i)} \sim \mathcal{N}(0, \sigma_d^{(i)2}\mathbb{I}_{k_d^{(i)}})$ is a $k_d^{(i)}$-variate normal random variable,*
3. *$\tilde{s}_d^{(i)} = (s_0^{(i)\top}, s_d^{(i)\top})^\top$ with $s_0^{(i)} \in \mathbb{R}^{c^{(i)}}$ and $s_d^{(i)} \in \mathbb{R}^{k_d^{(i)} - c^{(i)}}$ is a random vector such that:*
   (a) *the components of $\tilde{s}_d^{(i)}$ are mutually independent and each of them is a.s. a non-constant random variable,*
   (b) *$\tilde{s}_d^{(i)}$ is non-normal with 0 mean and unit variance.*
4. *$\epsilon_d^{(i)}$ is independent from $s_0^{(i)}$ and $s_d^{(i)}$: $\epsilon_d^{(i)} \perp\!\!\!\perp s_0^{(i)}$ and $\epsilon_d^{(i)} \perp\!\!\!\perp s_d^{(i)}$.*

Then, the number of shared sources is identifiable, i.e. $c^{(1)} = c^{(2)} =: c$ and for all $d = 1, \dots, D$ we get that $k_d^{(1)} = k_d^{(2)} =: k_d$, and there exist a sign matrix Γ_d and a permutation matrix $P_d \in \mathbb{R}^{k_d \times k_d}$ such that:

$$A_d^{(2)} = A_d^{(1)} P_d \Gamma_d,$$

and furthermore the source and noise distributions are identifiable, i.e.

$$\begin{bmatrix} s_0^{(2)} \\ s_d^{(2)} \end{bmatrix} \sim \Gamma_d^{-1} P_d^{-1} \begin{bmatrix} s_0^{(1)} \\ s_d^{(1)} \end{bmatrix}, \qquad \sigma_d^{(2)} = \sigma_d^{(1)}.$$

Note that the requirement $D \geq 2$ is essential for the identifiability of the non-Gaussian latent source and noise distributions. In contrast, in the single-view case, Kagan et al. [1973] shows that we cannot identify any arbitrary non-Gaussian source distribution unless we impose an additional constraint on the latent sources to have non-normal components (e.g., see Theorem A.2)[2].

Moreover, a necessary assumption for the identifiability of the linear structure is the non-normality of the latent sources, which is a standard assumption in the ICA literature [Comon, 1994] as we stated above. In the more restrictive multi-view shared ICA case, Richard et al. [2021] shows that the sources can be Gaussian if we impose additional assumptions about the diversity of the noise distributions. However, this is not applicable in our case since we do not make these assumptions for our model.

4 JOINT DATA LOG-LIKELIHOOD

Here, we derive the joint log-likelihood of the observed views which we use for estimating the mixing matrices. Following the standard ICA approaches [Bell and Sejnowski, 1995, Hyvärinen and Oja, 2000], instead of optimizing directly for the mixing matrices A_d, we estimate their inverses $W_d = A_d^{-1}$, called unmixing matrices.

Let $z_d := W_d x_d = \tilde{s}_d + \epsilon_d$, and $z_{d,0} := s_0 + \epsilon_{d0} \in \mathbb{R}^c$ and $z_{d,1} := s_d + \epsilon_{d1} \in \mathbb{R}^{k_d - c}$, i.e. $z_d = (z_{d,0}^\top, z_{d,1}^\top)^\top$ are the estimated noisy sources of the d-th view. Furthermore, let $p_{Z_{d,1}}$ be the probability distribution of $z_{d,1}$ and $|W_d| = |\det W_d|$. Then we can derive the the data log-likelihood of Equation 1 for N observed samples per view (proved in Appendix B), which is given by

$$\mathcal{L}(W_1, \dots, W_D) = \sum_{i=1}^{N} \Big(\log f(\bar{s}_0^i) + \sum_{d=1}^{D} \log p_{Z_{d,1}}(z_{d,1}^i) \Big) \tag{2}$$

$$- \frac{1}{2\sigma^2} \sum_{d=1}^{D} \Big(\operatorname{trace}(Z_{d,0} Z_{d,0}^\top) - \frac{1}{D} \sum_{l=1}^{D} \operatorname{trace}(Z_{d,0} Z_{l,0}^\top) \Big)$$

$$+ N \sum_{d=1}^{D} \log |W_d| + C$$

where $Z_{d,0} \in \mathbb{R}^{c \times N}$ for $d = 1, \dots, D$ is the data matrix that stores N observations of $z_{d,0}$, We estimate the shared sources via $\bar{s}_0^i = \sum_{d=1}^{D} z_{d,0}^i / D$ with probability distribution $f(\bar{s}_0) = \int \exp\Big(- \frac{D\|s_0 - \bar{s}_0\|^2}{2\sigma^2} \Big) p_{S_0}(s_0) ds_0$.

We further simplify the loss function by assuming that the data matrices $X_1 \in \mathbb{R}^{k_1 \times N}, \dots, X_D \in \mathbb{R}^{k_D \times N}$ are whitened. That consists of linearly transforming the random variables' realizations x_d such that the resulting variable $\tilde{x}_d = K_d x_d$ has uncorrelated components, i.e. unit variance, $\mathbb{E}[\tilde{x}_d \tilde{x}_d^\top] = \mathbb{I}_{k_d}$, where K_d is the whitening matrix. This step transforms the mixing matrix to an orthogonal one $\tilde{A}_d$.

[1] Note that the identifiability of the source distributions is a weaker notion of identifiability than the almost surely one (i.e. recovering the exact sources) in the noiseless case [Comon, 1994].

[2] A random variable x is said to have non-normal components if for every representation $x \sim v + w$ with $v \perp\!\!\!\perp w$, then v and w are non-normal.

In the new optimization problem after whitening, we aim to find orthogonal unmixing matrices $\tilde{W}_d = \tilde{A}_d^\top$ such that they maximize the transformed data log-likelihood:

$$\mathcal{L}(\tilde{W}_1, \dots, \tilde{W}_D) \propto \sum_{i=1}^{N} \Big(\log f_\sigma(\tilde{\bar{s}}_0^i) + \sum_{d=1}^{D} \log p_{\tilde{Z}_{d,1}}(\tilde{z}_{d,1}^i) \Big) \tag{3}$$
$$+ \frac{1+\sigma^2}{2D\sigma^2} \sum_{d=1}^{D} \sum_{l=1}^{D} \mathrm{trace}(\tilde{Z}_{d,0} \tilde{Z}_{l,0}^\top),$$

where analogously to Equation 2: $\tilde{z}_d = (\tilde{z}_{d,0}^\top, \tilde{z}_{d,1}^\top)^\top = \tilde{W}_d \tilde{x}_d$ and $\tilde{\bar{s}}_0^i = \sum_{d=1}^{D} \tilde{z}_{d,0}^i / D$. Note that after whitening we have $\mathrm{trace}(\tilde{Z}_{d,0} \tilde{Z}_{d,1}^\top) = c$ and $|\tilde{W}_d| = 1$ are constants and thus vanish from Equation 3 (see Appendix B for detailed derivations). The first line of Equation 3 represents the sources log-likelihoods and the second line has the role of a regularization term for finding the shared information between the views. In our work, Equation 3 is used for the parameter estimation where both the density of the shared and individual sources $f_\sigma(\cdot)$ and $p_{\tilde{Z}_{d,1}}$ are approximated by a nonlinear function $g(s)$, e.g. $g(s) = -\log\cosh(s)$ for super-Gaussian or $g(s) = e^{-s^2/2}$ for sub-Gaussian sources. Moreover, we treat the noise variance σ^2 as a Lagrange multiplier via the relation $\lambda = \frac{1+\sigma^2}{\sigma^2}$. Finally, after training we compute the mixing matrices $\hat{A}_d$ by setting $\hat{A}_d = K_d^{-1} \tilde{W}_d$. Thus, we recover the true ones A_d up to scaling with $(1+\sigma^2)^{\frac{1}{2}}$, sign and column permutation.

5 MODEL SELECTION

By leveraging the data generation model assumptions, we can select the number of shared sources c in a completely unsupervised way. More precisely, let $k < k_d$ for all $d = 1, \dots, D$ be a candidate for c which is unknown. Under the assumption that k is a correct guess (i.e. $k = c$), our generative model yields that $\hat{z}_d = z_{d,0} - \frac{1}{D} \sum_{l=1}^{D} z_{l,0}$ is normally distributed with 0 mean and variance $\frac{D-1}{D} \sigma^2 \mathbb{I}_k$ for each $d = 1, \dots, D$, where $z_{d,0}$ is defined as in Equation 2. We propose an evaluation metric, called normalized reconstruction error (NRE), defined by the following relation to the log-likelihood of $\hat{z}_d$:

$$\mathrm{NRE}(k) := \sum_{d=1}^{D} \frac{\|\hat{z}_d\|^2}{k} \overset{+}{\propto} - \sum_{d=1}^{D} \frac{\log p(\hat{z}_d)}{k}$$
$$= \sum_{d=1}^{D} \frac{D\|\hat{z}_d\|^2}{2(D-1)\sigma^2 k} - \frac{\cancel{k} \log(2\pi\sigma^2)}{2\cancel{k}}.$$

The two quantities differ by translation and multiplication with constants that do not depend on the parameter of interest k. Thus, by minimizing $\mathrm{NRE}(k)$ we maximize the sum of the (normalized) log-likelihoods of the normal variables $\hat{z}_d$. Intuitively, due to the normalization, $\mathrm{NRE}(k)$ can be interpreted as the average reconstruction error over the k shared sources (summed over all views). This allows for a fair comparison of the NRE scores for different k.

We select an optimal parameter by employing the following procedure. First, we split the data (that applies for each view) into two disjunct sets D_0 and D_1, with not necessarily the same sample sizes. We estimate the unmixing matrices for a fixed k on D_0 (train set) and estimate the shared sources on the test data D_1. Then we compute the mean $\mathrm{NRE}(k)$ on the recovered test shared sources (not on the train set due to possible overfitting, see Section 7). We repeat this for various k and we choose the maximum of all ks that minimize NRE, i.e.

$$k^* = \max\{\arg\min_k \overline{\mathrm{NRE}}(k)\},$$

where

$$\overline{\mathrm{NRE}}(k) = \frac{1}{N_1} \sum_{i \le |D_1|} \mathrm{NRE}(k)_i = \frac{1}{|D_1|} \sum_{i \le |D_1|} \frac{\|\hat{z}_d^i\|^2}{k}$$

is the average NRE score over all observed test samples in D_1. The NRE score serves as a goodness of fit measure and indicates how well the true shared sources are reconstructed from the test data. Due to the model fitting, we can get high-quality shared sources even when $k << c$, as we will demonstrate this empirically. Thus, we prefer to select the highest possible k for which the average shared sources reconstruction error is minimal.

6 RELATED WORK

The existing body of work on linear multi-view BSS, inspired by the ICA literature, considers mostly *shared* response model applications (i.e., no individual sources), some of them adopting a maximum likelihood approach [Guo and Pagnoni, 2008, Richard et al., 2020, 2021] to model the noisy views of the proposed models. Other methods, such as independent vector analysis (IVA), relax the assumption about the shared sources by assuming that they have the same first or highest order moments across view [Lee et al., 2008, Anderson et al., 2011, 2014, Engberg et al., 2016, Vía et al., 2011]. Many of these approaches, such as Group ICA [Calhoun et al., 2001], shared response ICA (SR-ICA) [Zhang et al., 2016], MultiViewICA [Richard et al., 2020], and ShICA[Richard et al., 2021], incorporate a dimensionality reduction step for every view (CCA [Varoquaux et al., 2009, Richard et al., 2021] or PCA) to extract the mutual signal between the multiple objects before applying an ICA procedure on the reduced data. However, there are no guarantees that the pre-processing procedure will entirely remove the influence of the object-specific sources on the transformed data. In the ICA literature, there exist three

methods for extracting shared and individual sources from data. Maneshi et al. [2016] proposes a heuristic way of using FastICA for the given task without discussing the identifiability of the results; [Long et al., 2020] suggests to apply ICA on each view separately followed by statistical analysis to separate the individual from the shared sources; [Lukic et al., 2002] exploits temporal correlations rather than the non-Gaussianity of the sources and thus is not applicable in the context we are considering.

A common tool for analyzing multi-view data is canonical correlation analysis (CCA), initially proposed by Hotelling [1936]. It finds two datasets' projections that maximize the correlation between the projected variables. Gaussian-CCA [Bach et al., 2005], its kernelized version [Bach et al., 2002] and deep learning [Andrew et al., 2013] formulations of the classical CCA problem aim to recover shared latent sources of variations from the multiple views. There are extensions of CCA that model the observed variables as a linear combination of group-specific and dataset-specific latent variables: estimated with Bayesian inference methods [Klami et al., 2013] or exponential families with MCMC inference [Virtanen, 2010]. However, most of them assume that the latent sources are Gaussian or non-linearly related to the observed data [Wang et al., 2016] and thus lack identifiability results.

Existing non-linear multiview versions such as [Tian et al., 2020, Federici et al., 2020] cannot recover both shared and individual signals across multiple measurements, and do not assure the identifiability of the proposed generative models. There are identifiable deep non-linear versions of ICA (e.g. [Hyvärinen et al., 2019]) which can be employed for this task. However, their assumptions for achieving identifiability are often hard to satisfy in real-life applications, especially in the biomedical domains with low-data regimes.

7 EXPERIMENTS

Model Implementation and Training. We used the python library `pytorch` [Paszke et al., 2017] to implement our method. We model each view with a separate unmixing matrix. To impose orthogonality constraints on the unmixing matrices, we made use of the `geotorch` library, which is an the extension of `pytorch` [Lezcano-Casado, 2019]. The stochastic gradient-based method applied for training is L-BFGS. Before running any of the ICA-based methods (our or the baselines), we whiten every single view by performing PCA to speed up computation. We estimate the mixing matrix up to scale (due to the whitening) and permutation (see Sections 3 and 4). To force the algorithm to output the shared sources in the same order across all views we initialize the unmixing matrices by means of CCA. This follows from the fact that the CCA weights are orthogonal matrices, and the transformed views' components are paired and ordered across views. For all conducted experiments, we fixed the parameter λ from Equation 3 to 1.

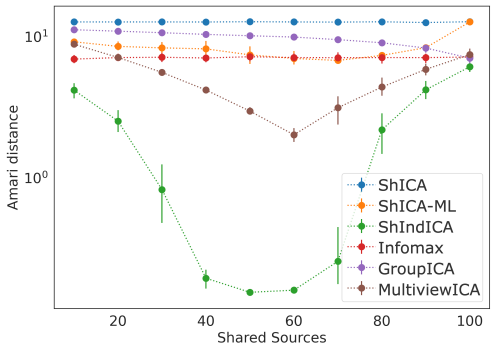

Figure 2: Comparison of ShIndICA (this paper) to ShICA, Infomax, GroupICA, MultiViewICA and ShICA-ML. The datasets come from two different views with total number of sources 100 and sample size 1000. We vary the number of the true number of shared sources from 10 to 100 (x-axis), which are considered to be known to the user before training. We compute the Amari distance (y-axis) between the estimated unmixing matrices and ground truth (the lower the better) in each case. ShIndICA consistently outperforms all baselines.

Baselines Implementation. We compare ShIndICA to the standard single-view ICA method Infomax [Ablin et al., 2018]. To adapt it to the multi-view setting, we run Infomax on each view separately, and then we apply the Hungarian algorithm [Kuhn and Yaw, 1955] to match components from different views based on their cross-correlation. For the shared response model settings, ShIndICA is compared to related methods such as MultiViewICA Richard et al. [2020], ShICA, ShICA-ML Richard et al. [2021], and GroupICA as proposed by Richard et al. [2020]. The latter involves a two-step pre-processing procedure, first whitening the data in the single views and then dimensionality reduction on the joint views. For the data integration experiment we use a method based on partial least squares estimation, closely related to CCA, that extracts between-views correlated components and view-specific ones. This method is provided by the `OmicsPLS` R package Bouhaddani et al. [2018] and is especially developed for data integration of omics data. We refer to this method as PLS.

7.1 SYNTHETIC EXPERIMENTS

Data Simulation. We simulated the data using the Laplace distribution $\exp(-\frac{1}{2}|x|)$, and the mixing matrices are sampled with normally distributed entries with mean 1 and 0.1 standard deviation. The realizations of the observed views are obtained according to the proposed model. In the different scenarios described below we vary the noise distribution. We conducted each experiment 50 times and based on that we provided error bars in all figures where applicable. Additional experiments are provided in Appendix D.2.

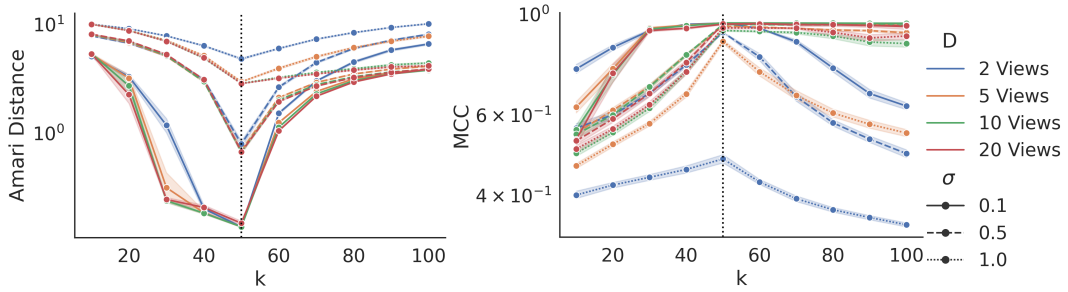

Figure 3: We generate data with 100 sources (50 shared annotated by a dashed line) for $D = 2, 5, 10, 20$ views and noise standard deviation $\sigma = 0.1, 0.5, 1$. We train the model with varying k, (x-axis) and we compute the average Amari distance (left plot) and the MCC of the assumed shared sources compared to the ground truth (right plot). While the Amari distance suggests that we get the best mixing matrix estimates when we "guessed" the right number of sources, the shared sources MCC plot shows that we can estimate the true shared sources with high quality if D is large enough also for overestimated c.

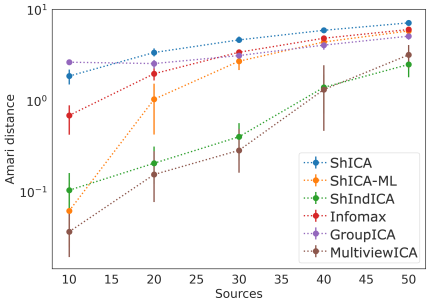

Figure 4: The data is generated according to a model, where no individual sources are present and the noise per view is uniformly sampled from the interval $[1, 2]$. The number of views is set to 10 and the sample size is 1000. We vary the number of sources from 10 to 50 (y-axis). ShIndICA and MultiviewICA have the best Amari distance compared to the other methods.

Motivational Example: Noiseless Views. This example illustrates the advantage of our method compared to the other multiview ICA methods for modelling view specific and individual sources. In Figure 2, we consider a noiseless view setting, where we fixed the dimension to be 100 and we vary the number of shared sources c from 10 to 100 in a two view setting. We fit a model for every c which is considered to be known. The quality of the mixing matrix estimation is measured with the Amari distance [Amari et al., 1995], which cancels if the estimated matrix differs from the ground truth one up to scale and permutation. We can see that as soon as the ratio of shared sources to individual sources gets around 1:1 we can recover the mixing matrices with very high accuracy (the Amari distance is almost 0) compared to the baseline methods which cannot perform well in this setting. Moreover, even in the case when all sources are shared, i.e. the baselines' model assumption is satisfied, our method performs as good as MultiViewICA which is a state of the art model designed for this task. More experiments on the noisy views are provided in Appendix D.2.

Shared Sources Estimation. This experiment exemplifies the performance of ShIndICA if the number of sources is a priori unknown and specified by the user. We sample a data set with 50 shared and 50 individual sources from $D = 2, 5, 10, 20$ views and noise standard deviation $\sigma = 0.1, 0.5, 1$. We vary the input number of shared sources from 10 to 100 and for each choice of this hyperparameter, we train a model on every dataset. The results are summarized in Figure 3, where the x-axis indicates the number of shared sources given for the training. The line colors and styles denote the number of views and noise distribution, respectively, used for the data generation. First, we assess the overall performance of ShIndICA in terms of the Amari distance of the estimated mixing matrices and ground truth ones. Figure 3 (left plot) shows that the Amari distance is the lowest when we guess correctly c. Furthermore, we assess the quality of the recovered shared sources (the average shared sources across all views $\bar{s}_0$) by computing the mean cross-correlation (MCC) between the estimates and the ground truth. That involves pairing the ground truth components with the estimated ones using the Hungarian algorithm and then computing the mean correlations between the aligned pairs. Figure 3 (right plot) suggests that even in the high noise variance case, we can get high-quality estimates of the shared sources (high MCC scores) if there are enough views present. This also holds when we overestimate c.

Model Selection. The previous experiment suggests that the hyper-parameter c is essential for the training and performance of ShIndICA. The NRE score, introduced in Section5, serves as a goodness of fit measure for selecting the correct number of sources. We consider the same data generation models as before. Again, we trained each model with various shared sources k. Figure 5 summarizes the results, where the x-axis refers to the hyper-parameter k. The y-axis is the corresponding NRE score on both train and test data (both with sample size =1000) indicated by the line style. First, in all cases, we observe that NRE remains low if the number of sources is lower than the true ones and it

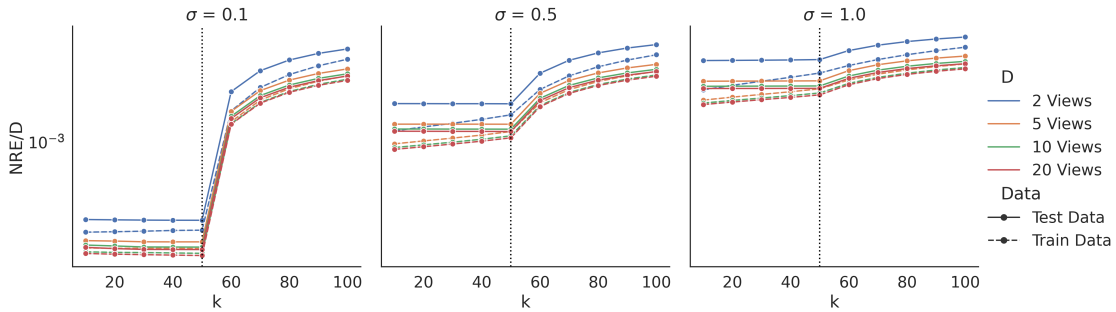

Figure 5: We generate data with 100 sources out of which 50 are shared (dashed line) for different views $D = 2, 5, 10, 20$ and noise variances $\sigma = 0.1, 0.5, 1$. We compute the NRE on the test and train data for different candidates of $c = 10, \ldots, 100$. If we "overestimate" the number of shared sources we see that the NRE score increases.

increases as soon as we overestimate c, especially when the noise variance is low (left plot). Moreover, due to overfitting, the NRE score computed on the train data takes its minimum for the lowest k. In contrast, the NRE on the test data that remains constant or for large D even reaches its minimum at the correct number of sources. Thus, it makes it more suitable for model selection than the NRE on the train data.

Robustness to model misspecification in a shared response model application. Here we want to investigate the robustness of our model when the noise has a view-specific variance. To provide a fair comparison to the baseline methods, we apply our method to a shared response setting, i.e. no individual sources are available. For this experiment the view-specific variances are uniformly sampled from $[1, 2]$, the number of view is 10 and the number of sources varies from 10 to 50. Figure 4 shows that ShIndICA and MultiViewICA show consistently the best model performances (lowest Amari distance between estimates and ground truth matrices) compared to the other methods.

7.2 DATA FUSION OF TRANSCRIPTOME DATA

Background and Data Generation Assumptiom. Transcriptome datasets are relevant for the field of genomics. After preprocessing they have the form of random data matrices, where each row correspond to a gene and each column refers to an experiment. Based on these datasets, scientists try to infer gene-gene interactions in the genome. Combining as many datasets as possible enables getting better gene regulatory predictions. This is a challenging task due to the batch effects (non-biological noise) in the data. We do a one-to-one translation of this data integration task to our proposed model by assuming that each view represents a different lab, each experiment is a noisy linear combination of independent gene pathways. We also assume that some gene pathways get activated due to the specific experimental design (individual sources) and others show stable activation level across all experiments (shared sources).

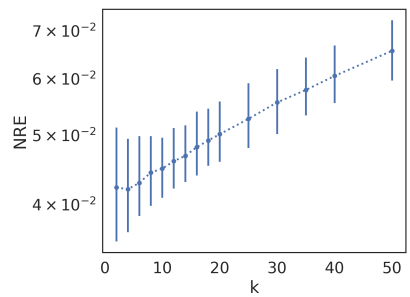

Figure 6: The NRE score computed on test data for the transcriptome data for various k. This procedure was repeated 50 times and the error bars represent the estimated 95% confidence interval.

Datasets. In this example, we consider the bacterium *B. subtilis*, for which a very rich collection of the discovered gene-gene interactions are publicly available, which we use as our ground truth model in the graph inference task. Our goal is to "denoise" and combine two publicly available datasets [Arrieta-Ortiz et al., 2015, Nicolas et al., 2012]. Each of the datasets contain gene expression levels of about 4000 genes measured across more than 250 experimental outcomes. For detailed description of the datasets see Appendix C.

Model Selection. In this real-life application we do not have any prior knowledge about the shared information between the two datasets (two views). Therefore, we utilize the model selection procedure in Section 5 to choose the number of shared sources. In this case, we randomly split the data into train and test set with proportions 3:1. We estimated the mixing matrices on the train data for different k. We reconstruct the test set shared sources and compute the corresponding NRE scores. This procedure is repeated 50 times for different splits and the results are displayed in Figure 6. The NRE score reaches its minimum for $k = 4$ which indicates the number of shared sources. Furthermore, we provided a biological interpretation of the estimated shared sources, by matching them to gene pathways, visualized in Figure 7. The x-axis represent the genes, decoded by num-

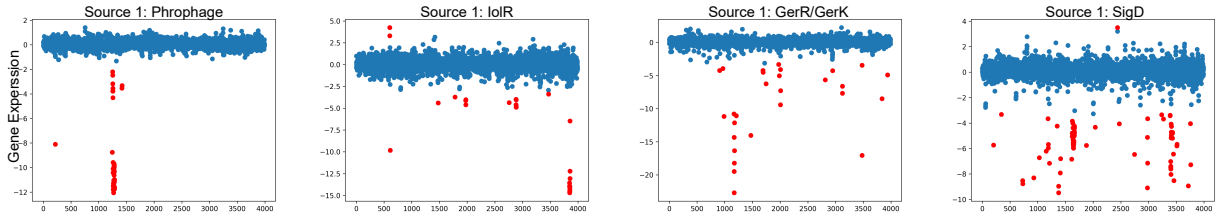

Figure 7: Gene Expressions from the shared latent sources. The red markers are outliers. To a great extent, they can be related to functional groups (source 1) and gene regulatory networks (sources (2-4)), which are given in the titles of the subfigures.

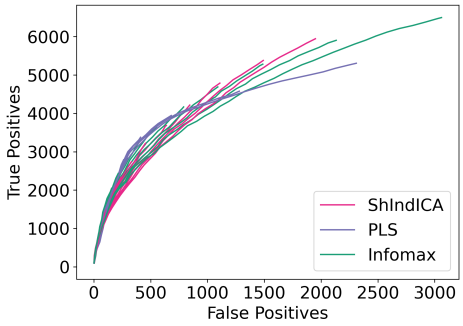

Figure 8: We compare the top ten models with ShIndICA, PLS and Infomax. We order the edges from the networks according to their strength. We count the true positives (y-axis) and possibly false positive edges (x-axis) in the first $100, 200, \ldots$ edges. ShIndICA and Infomax outperform PLS for higher total number of edges.

bers $0 - 3994$, and the y-axis the corresponding latent "gene expressions" in the latent source. Each marker represents one gene, and the red markers annotate the outliers. We compared the outliers with the available ground truth regulatory network and interestingly, we could conclude that almost all red markers from the first source belong to prophage genes, and the ones from the other three sources are regulated by the *iolR* and *gerE/gerK* and *sigD* regulators, respectively.

Data Integration for Co-regulation Inference. The combined datasets can be used for co-regulation prediction. More precisely, in this application, we want to estimate an undirected graph with nodes referring to the genes and with edges connecting genes with a common regulator. Since the transcriptome datasets are in the high-dimension-low-sample-size regime (number of genes>number of samples), usually graphical lasso [Friedman et al., 2007] is well-suited for inferring graphical structure from the observed data. In this case, instead of using the "raw" data samples as input data to the graphical lasso, we use the samples extracted from the data integration algorithms. Ideally, the combined data will boost the graphical lasso performance.

Experiment. We evaluate ShIndICA, PLS and naive ICA approach (Infomax as in the previous example) on the defined data integration downstream task. We select the number of shared sources for ShIndICA to be 5, for PLS - 10 (selected by cross-validation procedure provided by [Bouhaddani et al., 2018]) and 0 for the naive Infomax approach. The data is whitened with PCA and the number of sources per view is reduced to 180. After applying each method, we fit 30 graphical lasso models for different penalization parameters on the estimated components. We select the top 10 models by employing a statistical goodness-of-fit measure, called EBIC (see Appendix C for more details). In Figure 8 we compare the 10 output graphs from the graphical lasso for each pre-processing method in the following way. For each estimated graph, we order the edges according to their strength. Then we count the true positive (y-axis) and false positive (x-axis) edges in the first $100, 200, \ldots$ edges. From Figure 8. We can conclude that PLS shows better performance at the beginning and gets outperformed by the other two methods (especially ShIndICA) for a number of edges > 5000. ShIndICA performs slightly better than Infomax. The reason could be that both models output similar sources due to the small k specified in the ShIndICA case. We also run the graphical lasso on the pooled data without any pre-processing. Surprisingly, the EBIC evaluates the empty graph as the best model describing the data.

8 DISCUSSION

We proposed a novel noisy linear ICA approach that utilizes the prior knowledge that the different views share information to infer both shared and view-specific sources, called ShIndICA. We provided theoretical guarantees for the identifiability of the model's linear structure, latent source and noise distributions and the number of shared and individual sources. We estimate the unmixing matrices by maximizing the joint log-likelihood of the observed views. Furthermore, we proposed a goodness of fit measure for choosing the number of shared sources. Our empirical results showed that our model performs well on simulated data also when the model is misspecified. We also suggested a novel strategy for combining transcriptome data and empirically showed that the estimated sources can be matched to biologically meaningful signals. Moreover, our model improves the performance of a graphical inference model chosen for the particular task. In future work, we would like to address some possible extensions, such as allowing for dependency between the sources of different views. This resembles more real-life applications like the one considered above.

References

P. Ablin et al. Faster ica under orthogonal constraint. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4464–4468. IEEE, 2018.

S. Amari et al. A new learning algorithm for blind signal separation. *Advances in neural information processing systems*, 8, 1995.

M. Anderson et al. Joint blind source separation with multivariate gaussian model: Algorithms and performance analysis. *IEEE Transactions on Signal Processing*, 60(4): 1672–1683, 2011.

M. Anderson et al. Independent vector analysis: Identification conditions and performance bounds. *IEEE Transactions on Signal Processing*, 62(17):4399–4410, 2014.

G. Andrew et al. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255. PMLR, 2013.

M. Arrieta-Ortiz et al. An experimentally supported model of the bacillus subtilis global transcriptional regulatory network. *Molecular systems biology*, 11(11):839, 2015.

F. Avila Cobos et al. Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics*, 34(11):1969–1979, 2018.

M. Aynaud et al. Transcriptional programs define intratumoral heterogeneity of ewing sarcoma at single-cell resolution. *Cell reports*, 30(6):1767–1779, 2020.

F. Bach et al. Kernel independent component analysis. *Journal of machine learning research*, 3(Jul):1–48, 2002.

F. Bach et al. A probabilistic interpretation of canonical correlation analysis. 2005.

P. Bartolomeo et al. Botallo's error, or the quandaries of the universality assumption. *Cortex*, 86:176–185, 2017.

A. Bell and T. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

S. Bouhaddani et al. Integrating omics datasets with the omicspls package. *BMC bioinformatics*, 19(1):1–9, 2018.

V. Calhoun et al. A method for making group inferences from functional mri data using independent component analysis. *Human brain mapping*, 14(3):140–151, 2001.

M. Cary et al. Application of transcriptional gene modules to analysis of caenorhabditis elegans' gene expression data. *G3: Genes, Genomes, Genetics*, 10(10):3623–3638, 2020.

P. Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

M. Congedo et al. Group independent component analysis of resting state eeg in large normative samples. *International Journal of Psychophysiology*, 78(2):89–99, 2010.

M. Davies. Identifiability issues in noisy ica. *IEEE Signal processing letters*, 11(5):470–473, 2004.

J. Dubois et al. Building a science of individual differences from fmri. *Trends in cognitive sciences*, 20(6):425–443, 2016.

S. Dubois et al. Refining diffuse large b-cell lymphoma subgroups using integrated analysis of molecular profiles. *EBioMedicine*, 48:58–69, 2019.

J. Durieux et al. Partitioning subjects based on high-dimensional fmri data: comparison of several clustering methods and studying the influence of ica data reduction in big data. *Behaviormetrika*, 46(2):271–311, 2019.

A. Engberg et al. Independent vector analysis for capturing common components in fmri group analysis. In *2016 international workshop on pattern recognition in neuroimaging (prni)*, pages 1–4. IEEE, 2016.

M. Federici et al. Learning robust representations via multiview information bottleneck. *ICLR*, 2020.

J. Friedman et al. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 12 2007. ISSN 1465-4644. doi: 10.1093/biostatistics/kxm045. URL `https://doi.org/10.1093/biostatistics/kxm045`.

Y. Guo and G. Pagnoni. A unified framework for group independent component analysis for multi-subject fmri data. *NeuroImage*, 42(3):1078–1093, 2008.

H. Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics*, pages 162–190. Springer, 1936.

R. Huster et al. Group-level component analyses of eeg: validation and evaluation. *Frontiers in neuroscience*, 9: 254, 2015.

A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5): 411–430, 2000.

A. Hyvärinen et al. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, 2019.

A. Kagan et al. *Characterization problems in mathematical statistics*. Wiley-Interscience, 1973.

A Klami et al. Bayesian canonical correlation analysis. *Journal of Machine Learning Research*, 14(4), 2013.

H. W. Kuhn and Bryn Yaw. The hungarian method for the assignment problem. *Naval Res. Logist. Quart*, pages 83–97, 1955.

J. Lee et al. Independent vector analysis (iva): multivariate approach for fmri group study. *Neuroimage*, 40(1):86–109, 2008.

M. Lezcano-Casado. Trivializations for gradient-based optimization on manifolds. In *Advances in Neural Information Processing Systems, NeurIPS*, pages 9154–9164, 2019.

Q. Long et al. Independent vector analysis for common subspace analysis: Application to multi-subject fmri data yields meaningful subgroups of schizophrenia. *NeuroImage*, 216:116872, 2020.

A. Lukic et al. An ica algorithm for analyzing multiple data sets. In *Proceedings. International Conference on Image Processing*, volume 2, pages II–II. IEEE, 2002.

M. Maneshi et al. Validation of shared and specific independent component analysis (ssica) for between-group comparisons in fmri. *Frontiers in neuroscience*, 10:417, 2016.

M. McKeown and T. Sejnowski. Independent component analysis of fmri data: examining the assumptions. *Human brain mapping*, 6(5-6):368–372, 1998.

P. Nazarov et al. Deconvolution of transcriptomes and mirnomes by independent component analysis provides insights into biological processes and clinical outcomes of melanoma patients. *BMC medical genomics*, 12(1):1–17, 2019.

P. Nicolas et al. Condition-dependent transcriptome reveals high-level regulatory architecture in bacillus subtilis. *Science*, 335(6072):1103–1106, 2012.

A. Paszke et al. Automatic differentiation in pytorch. 2017.

H. Richard et al. Modeling shared responses in neuroimaging studies through multiview ica. *Advances in Neural Information Processing Systems*, 33:19149–19162, 2020.

H. Richard et al. Shared independent component analysis for multi-subject neuroimaging. *Advances in Neural Information Processing Systems*, 34:29962–29971, 2021.

Z. Rusan et al. Granular transcriptomic signatures derived from independent component analysis of bulk nervous tissue for studying labile brain physiologies. *bioRxiv*, 2020.

M. Salman et al. Group ica for identifying biomarkers in schizophrenia:'adaptive'networks via spatially constrained ica show more sensitivity to group differences than spatio-temporal regression. *NeuroImage: Clinical*, 22:101747, 2019.

A. Sastry et al. The escherichia coli transcriptome mostly consists of independently regulated modules. *Nature communications*, 10(1):1–14, 2019.

A. Sastry et al. Independent component analysis recovers consistent regulatory signals from disparate datasets. *PLoS computational biology*, 17(2):e1008647, 2021.

J. Tan et al. Independent component analysis of e. coli's transcriptome reveals the cellular processes that respond to heterologous gene expression. *Metabolic Engineering*, 61:360–368, 2020.

Y. Tian et al. Contrastive multiview coding. In *European conference on computer vision*, pages 776–794. Springer, 2020.

C. Urzúa-Traslaviña et al. Improving gene function predictions using independent transcriptional components. *Nature communications*, 12(1):1–14, 2021.

G. Varoquaux et al. Canica: Model-based extraction of reproducible group-level ica patterns from fmri time series. *arXiv preprint arXiv:0911.4650*, 2009.

J. Vía et al. A maximum likelihood approach for independent vector analysis of gaussian data sets. In *2011 IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6. IEEE, 2011.

R. Vigário et al. Independent component analysis for identification of artifacts in magnetoencephalographic recordings. *Advances in neural information processing systems*, 10, 1997.

S. Virtanen. *Bayesian exponential family projections*. PhD thesis, Aalto University, 2010.

W. Wang et al. Deep variational canonical correlation analysis. *arXiv preprint arXiv:1610.03454*, 2016.

H. Zhang et al. A searchlight factor model approach for locating shared information in multi-subject fmri analysis. *arXiv preprint arXiv:1609.09432*, 2016.

C. Zheng et al. Gene expression data classification using consensus independent component analysis. *Genomics, proteomics & bioinformatics*, 6(2):74–82, 2008.

W Zhou and Russ B Altman. Data-driven human transcriptomic modules determined by independent component analysis. *BMC bioinformatics*, 19(1):1–25, 2018.