Multi-View Independent Component Analysis with Shared and Individual Sources

Teodora Pandeva[1,2] **Patrick Forré**[1]

[1]AI4Science, AMLab, University of Amsterdam, The Netherlands
[2]Swammerdam Institute for Life Sciences, University of Amsterdam, The Netherlands

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, referred to as ShIndICA, 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. Furthermore, we propose a model selection procedure for recovering the number of shared sources. Finally, we empirically demonstrate the advantages of our model over baselines. We apply ShIndICA in a challenging real-life task, using three transcriptome datasets provided by three different labs (three different views). The recovered sources were used for a downstream graph inference task, facilitating the discovery of a plausible representation of the data's underlying graph structure.

1 INTRODUCTION

Independent Component Analysis (ICA) is a method used to solve Blind Source Separation (BSS) problems [Comon, 1994]. The aim is to separate independent latent sources from mixed observed signals, thus revealing essential structures in various types of data. Historically, linear ICA has proven to be a successful approach in recovering spatially independent sources, such as regions of brain activity from magnetoencephalography (MEG) data [Vigário et al., 1997], or functional MRI (fMRI) data [McKeown and Sejnowski, 1998]. The utility of ICA extends beyond neuroscience, with applications in omics data analysis, for example, [Zheng et al., 2008, Zhou and Altman, 2018, Nazarov et al., 2019, Dubois et al., 2019, Tan et al., 2020, Rusan et al., 2020, Cary et al., 2020, Aynaud et al., 2020, Urzúa-Traslaviña et al., 2021]. In these works, the interpretation of the latent sources hinges on the assumption that each experimental outcome is a linear mixture of different cell types, disease states, or other independent biological processes. For example, the gene expression data from a tumor biopsy might include signals from cancer cells, immune cells, and other cell types within the tumor microenvironment. Each of these cell types has distinct gene expression profiles that get mixed together in the observed data [Avila Cobos et al., 2018].

The rapid 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 and Wilderjans, 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. A scenario highlighting this limitation is when one is interested in investigating the individual-specific brain functions (view-specific) and shared phenotypes patterns in individuals' brain activity in a natural stimuli experiment [Dubois and Adolphs, 2016, Bartolomeo et al., 2017].

Another application, where the estimation of both shared and view-specific sources is essential, is omics data integration (e.g. see [Smilde et al., 2017, Li et al., 2018]). For instance, we often have datasets from different sources (such as gene expression, proteomics, etc.) collected under various conditions and perhaps from different cohorts of individuals. The shared signals across these datasets represent stable patterns, such as consistently expressed genes across different types of cells or universally present metabolic pathways. These shared patterns can help us understand the fundamen-

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

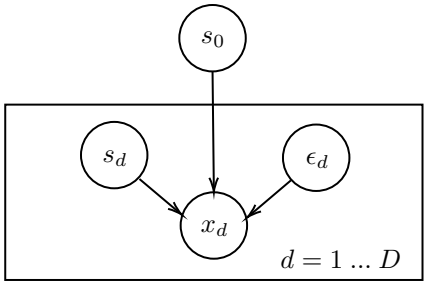

Figure 1: A graphical representation of Equation 1 where x_d is the observed variable, s_0 denotes the shared sources, s_d the view-specific ones and ϵ_d is the Gaussian noise.

tal biological processes common to all cells. On the other hand, view-specific signals reflect unique patterns under particular conditions or within specific cohorts. For instance, some genes may be expressed only under certain stress conditions, or particular biological patterns may be unique to a specific cohort of individuals with a disease. These view-specific patterns can provide insights into the specific factors that differentiate one condition or cohort from another.

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 a 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 that 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 to a data integration problem of two large transcriptome datasets. We show empirically that our method works well when the estimated components are used for a graph inference task.

Contributions. We summarize our contributions:

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, as well as their dimensions.
3. We derive the closed-form joint likelihood of ShIndICA, which is used for estimating the mixing matrices.
4. By leveraging the generative model assumptions, we propose a selection criterion for inferring the correct number of shared sources.

2 PROBLEM FORMALIZATION

Consider the following D-view multivariate linear BSS model where for $d \in \{1, \ldots, D\}$ with a graphical representation in Figure 1:

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

and it holds that

1. $x_d \in \mathbb{R}^{k_d}$ is a random vector,
2. $\tilde{s}_d = (s_0^\top, s_d^\top)^\top$ are latent non-Gaussian sources with $\mathbb{E}[\tilde{s}_d] = 0$ and $\mathrm{Var}[\tilde{s}_d] = \mathbb{I}_{k_d}$, and $s_0 \in \mathbb{R}^c$ and $s_d \in \mathbb{R}^{k_d - c}$ the shared and individual sources,
3. $A_d \in \mathbb{R}^{k_d \times k_d}$ is an invertible mixing matrix, 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 random variables are mutually independent.

We name the proposed generative model ShIndICA. This model builds upon the MultiViewICA model by Richard et al. [2020], which assumes the presence of shared sources exclusively. The Gaussian noise in Equation 1 mirrors a device measurement error with variance $\sigma^2 A_d A_d^\top$, akin to [Richard et al., 2020, 2021]. We choose this configuration as it yields a closed-form joint data likelihood (see Section 4), which is not available for alternative representations. Assumption 5 maintains that noise does not interact with the true signal, a concept typical for measurement error models. Moreover, when $D = 1$, ShIndICA reverts to a standard linear ICA model, solved by Comon [1994], Bell and Sejnowski [1995], Hyvärinen and Oja [2000] for independent non-Gaussian latent sources $z := \tilde{s}_1 + \epsilon_1$.

3 IDENTIFIABILITY RESULTS

Due to the absence of labels in unsupervised learning, the algorithm's reliability cannot be directly confirmed outside of simulations. Therefore, we rely on theoretical guarantees to trust the algorithm's estimation of quantities of interest. In the case of BSS problems, it is necessary for sources and mixing matrices to be uniquely determined (or *identifiable*) by the data, particularly when dealing with large samples.

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 defined in Equation 1, we extend the results by Kagan et al. [1973], Comon [1994], Davies [2004], Richard et al. [2020]. Compared to previous work, we provide identifiability guarantees not only for the mixture matrices up to sign and permutation, but also for the *sources and noise distributions* (up to the same sign and permutation), and the latent (both shared and individual) sources dimensions[1]. Moreover, our identifiability statement holds for a more general case than Equation 1 where 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 Section 1 of the Supplementary Material:

Theorem 3.1 *Let $x_1, \ldots, x_D$ 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 $d \in \{1, \ldots, D\}$, 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 $d = 1, \ldots, D$ we get that $k_d^{(1)} = k_d^{(2)} =: k_d$, and there exists 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. Conversely, in a single-view scenario, Kagan et al. [1973] demonstrates that identifying any arbitrary non-Gaussian source distribution is unfeasible unless we introduce an additional constraint mandating the latent sources to have non-normal components (refer to Theorem A.2).[2].

Furthermore, to identify the linear structure, it is necessary to assume the non-normality of the latent sources—a standard presumption in ICA literature [Comon, 1994], as previously mentioned. In a multi-view shared ICA scenario, Richard et al. [2021] posits that the sources can be Gaussian if we impose additional conditions regarding the diversity of noise distributions. However, this is not relevant in our case as we do not incorporate these assumptions in 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 data log-likelihood of Equation 1 for N observed samples per view (proved in Section 2 of the Supplementary Material), given by

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

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

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

where $Z_{d,0} \in \mathbb{R}^{c \times N}$ for $d = 1, \ldots, D$ is the data matrix that stores N observations of $z_{d,0}$.

The first term in Equation 2 refers to data log-likelihood of the estimated shared sources $\bar{z}_0^i = \sum_{d=1}^{D} z_{d,0}^i / D$ with density $f(\bar{z}_0) = \int \exp\Big(- \frac{D \|s_0 - \bar{z}_0\|^2}{2\sigma^2} \Big) p_{S_0}(s_0) ds_0$; the second term is the data log-likelihood of the view-specific sources. The second line in the above equation

[1] Note that the identifiability of the source distributions is a weaker notion of identifiability than the almost sure 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.

can be further simplified by assuming that the data matrices $X_1 \in \mathbb{R}^{k_1 \times N}, \ldots, X_D \in \mathbb{R}^{k_D \times N}$ are whitened.

Joint data log-likelihood after whitening. Whitening is a data pre-processing procedure 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 $\tilde{A}_d$.

After whitening, the joint data log-likelihood has the form

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

where analogously to Equation 2 we abbreviate: $\tilde{z}_d = (\tilde{z}_{d,0}^\top, \tilde{z}_{d,1}^\top)^\top = \tilde{W}_d \tilde{x}_d$ and $\tilde{\bar{z}}_0^i = \sum_{d=1}^{D} \tilde{z}_{d,0}^i / D$. Note that after whitening $\operatorname{trace}(\tilde{Z}_{d,0} \tilde{Z}_{d,1}^\top) = c$ and $|\tilde{W}_d| = 1$ and thus vanish from Equation 3 (see Section 2 of the Supplementary Material for detailed derivations).

Understanding the loss. In our work, we use Equation 3 for parameter estimation. The first line illustrates the log-likelihoods of the sources, while the second line functions as a regularization term, essential for uncovering the shared information among views. This regularization term maximizes the alignment between the estimated noisy shared signals across each pair of views.

Optimization. Both the density of shared and individual sources, denoted by $f_\sigma(\cdot)$ and $p_{\tilde{Z}_{d,1}}$, are unknown and, thus, approximated by a *prior* nonlinear function $g(s)$. For instance, we use $g(s) = -\log\cosh(s)$ in our experiments. Importantly, we do not directly optimize for the noise variance σ^2; instead, we establish it as a Lagrange multiplier through the relationship $\lambda = \frac{1+\sigma^2}{\sigma^2}$ [3].

5 MODEL SELECTION

The selection of the number of shared sources c can be accomplished in an unsupervised manner, leveraging the assumptions of the data generation model. To be precise, we consider $k < k_d$ for all $d = 1, \ldots, D$ as a potential candidate for the unknown shared number of sources c. Recall that $z_d = W_d x_d$ where $(z_{d,0}^\top, z_{d,1}^\top)^\top$ are the estimated sources for the d-th view, and $z_{d,0} \in \mathbb{R}^k$. To signify the dependency of $z_{d,0}$ on k, we write it as $z_{d,0}^{(k)}$. We introduce a goodness-of-fit measure called the Normalized Reconstruction Error (NRE), designed to evaluate the quality of $z_{d,0}^{(k)}$ for varying values of k, defined as:

$$\mathrm{NRE}(k) := \sum_{d=1}^{D} \frac{\|\hat{z}_d^{(k)}\|^2}{k},$$

where we make use of the variable $\hat{z}_d^{(k)} = z_{d,0}^{(k)} - \bar{z}_0^{(k)} = z_{d,0}^{(k)} - \frac{1}{D}\sum_{l=1}^{D} z_{l,0}^{(k)}$. Now, if we assume k^* to be the correct guess (i.e. $k^* = c$) and that the true unmixing matrices W_d's have been estimated, the $z_{d,0}^{(k^*)}$ for $d = 1, \ldots, D$ become well-aligned. Consequently, $\hat{z}_d^{(k^*)}$ follows a normal distribution with a mean $\mathbb{E}[\hat{z}_d^{(k^*)}] = 0$ and variance $\mathrm{Var}(\hat{z}_d^{(k^*)}) = \frac{D-1}{D}\sigma^2 \mathbb{I}_{k^*}$ for each $d = 1, \ldots, D$. We can then relate this to the log-likelihood of $\hat{z}_d^{(k^*)}$:

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

The two quantities differ by translation and multiplication with constants (indicated with $\overset{+}{\propto}$). Thus, $\mathrm{NRE}(k^*)$ approaches $(D-1)\sigma^2$ in the large sample size limit, i.e. $\mathrm{NRE}(k^*) \approx D \cdot \frac{D-1}{D}\frac{\sigma^2 \cdot k^*}{k^*}$. Overestimation of shared sources by choosing $k > c$ breaches the generative model assumptions, indicating that the estimated $z_{d,0}^{(k)}$ are misaligned, leading to a $\hat{z}_d^{(k)}$ with $\mathrm{NRE}(k) > (D-1)\sigma^2 \approx \mathrm{NRE}(k^*)$. For $k < c$, our method outputs well-aligned $\hat{z}_d^{(k)}$ leading to $\mathrm{NRE}(k) \approx \mathrm{NRE}(k^*)$. Hence, we can apply an elbow or the following method to select the optimal parameter.

Calculating NRE. We repeat the procedure for various k

1. We divide the data (applicable to each view) into two separate sets, with sizes N_0 and N_1, which do not necessarily have the same sample sizes.
2. We then estimate the unmixing matrices for a fixed k on the training set and estimate the shared sources on the test data.
3. We calculate the mean $\mathrm{NRE}(k)$ on the recovered test shared sources (not on the training set due to potential overfitting).

We choose the maximum of all k's that minimize the NRE:

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

where

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

[3] 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.

is the average NRE score over all observed test samples. This score serves as a measure of the model's goodness of fit and indicates how well the true shared sources can be reconstructed from the unseen data. Even with $k << c$, we can still achieve high-quality shared sources due to model fitting, as we will demonstrate empirically. Consequently, we prefer to select the highest possible k that yields the minimum average shared source reconstruction error.

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 higher order moments across view [Lee et al., 2008, Anderson et al., 2011, Vía et al., 2011, Anderson et al., 2014, Engberg et al., 2016]. 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 applying 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 and Jordan, 2005], its kernelized version [Bach and Jordan, 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 [Salzmann et al., 2010, Kang and Choi, 2011, 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 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 extension of `pytorch` [Lezcano-Casado, 2019]. The gradient-based method applied for training is L-BFGS. Before running any of the ICA-based methods 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 ensure that the algorithm consistently outputs the shared sources in the same order across all views, we initialized the unmixing matrices using Canonical Correlation Analysis (CCA). We made this decision based on the property of CCA weights to constitute orthogonal matrices and the fact that they facilitate pairing and ordering of the transformed components across all views. For all conducted experiments, we fixed the parameter $\lambda = 1$ where $\lambda := \frac{1+\sigma^2}{\sigma^2}$. The code for our method is publicly available at `https://github.com/tpandeva/shindica/`.

Baselines Implementation. We benchmark our ShIndICA against the standard single-view ICA method, Infomax [Ablin et al., 2018]. To adapt Infomax to the multi-view context, we apply it independently to each view. To match the shared components across different views, we use the Hungarian algorithm [Kuhn and Yaw, 1955] on their cross-correlation. For settings that involve a shared response model, we draw comparisons between ShIndICA and 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]. GroupICA incorporates a two-step preprocessing procedure that initially whitens the data in the single views, followed by a dimensionality reduction step on the joint views.

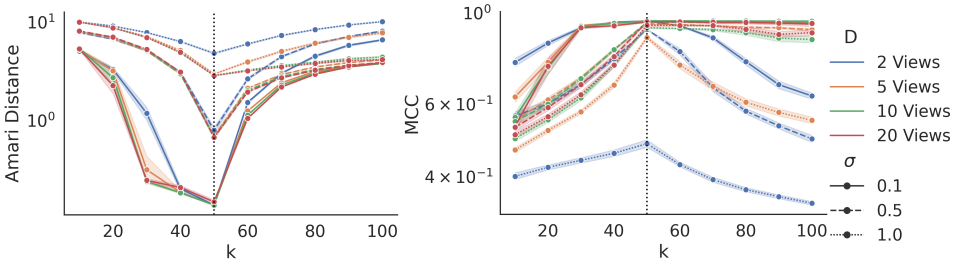

Figure 2: We generate data with 50 individual and 50 shared sources (annotated by a dashed line) for $D = 2, 5, 10, 20$ and noise standard deviation $\sigma = 0.1, 0.5, 1$. We train the model with varying k (x-axis), and compute the average Amari distance (left plot) and the MCC (right plot) of the estimated shared sources and the ground truth ones. 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.

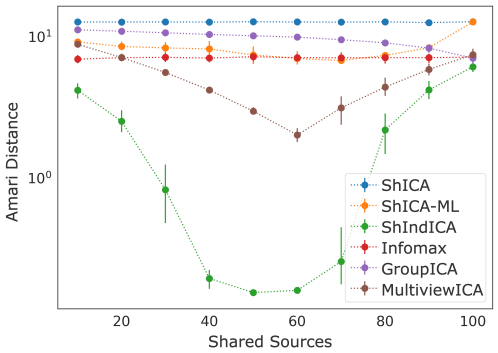

Figure 3: ShIndICA's (this paper) performance is compared with ShICA, Infomax, GroupICA, MultiViewICA, and ShICA-ML. Datasets from two views, each with 100 sources and 1000 samples, are used. The known number of shared sources varies from 10 to 100 (x-axis). The Amari distance (y-axis) measures accuracy with lower values indicating better performance. ShIndICA consistently outperforms the other methods.

7.1 SYNTHETIC EXPERIMENTS

Data Simulation. We simulated the data using the Laplace distribution $\exp(-\frac{1}{2}|x|)$. The mixing matrices are sampled with entries following a normal distribution with a mean of 1 and a standard deviation of 0.1. 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 Section 4 of the Supplementary Material.

Motivational Example: Noiseless Views. This example illustrates the advantage of our method compared to the other multi-view ICA methods for modeling view-specific and individual sources. In Figure 3, 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 well as MultiViewICA which is a state-of-the-art model designed for this task. More experiments on the noisy views are provided in Section 4 of the Supplementary Material.

Shared Sources Estimation. This experiment illustrates ShIndICA's performance when the number of sources is unknown beforehand and user-defined. We generate a dataset comprising 50 shared and 50 individual sources from $D = 2, 5, 10, 20$ views with noise standard deviations of $\sigma = 0.1, 0.5, 1$. The number of shared sources given for training is varied from 10 to 100 and a model is trained on each dataset for every choice of this hyperparameter. The results are summarized in Figure 2, 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 used for the data generation. We first evaluate ShIndICA's overall performance using the Amari distance between the estimated and actual mixing matrices. The left plot of Figure 2 shows the Amari distance reaching its lowest when the correct value of c is guessed. Next, we evaluate the quality of the recovered shared sources (the average shared sources across all views, $\bar{z}_0$) by calculating the

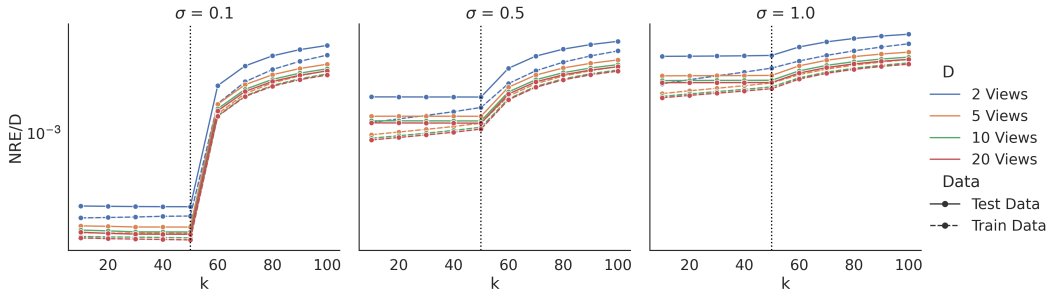

Figure 4: We generate data with 100 sources, of which 50 are shared (dashed line) for different views $D = 2, 5, 10, 20$ and noise variances $\sigma = 0.1, 0.5, 1$. We compute the NRE on the test and train data for different candidates of $c = 10, \ldots, 100$. If we "overestimate" the number of shared sources we see that the NRE score increases.

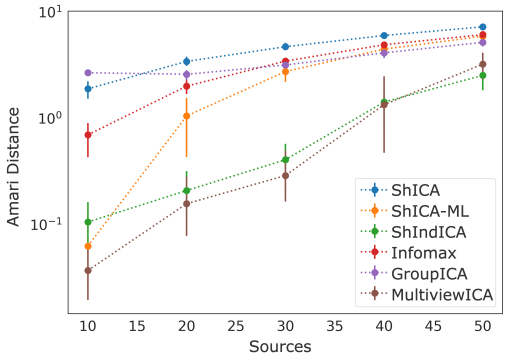

Figure 5: The data is generated according to a model where no individual sources are present and the noise per view is uniformly sampled from the interval $[1, 2]$. The number of views is set to 10, and the sample size is 1000. We vary the number of sources from 10 to 50 (y-axis). ShIndICA and MultiviewICA have the best Amari distance compared to the other methods.

mean cross-correlation (MCC) between the estimates and the actual sources. This involves aligning the ground truth components with the estimated ones using the Hungarian algorithm and computing the mean correlations between these matched pairs. The right plot in Figure 2 implies that even with high noise variance, high-quality estimates of the shared sources (high MCC scores) are attainable given sufficient views. This holds even when c is overestimated.

Model Selection. The preceding experiment indicates the critical role of hyper-parameter c in ShIndICA's training and performance. The NRE score, as introduced in Section 5, aids in selecting the correct number of sources by serving as a goodness-of-fit measure. The same data generation models as before are considered. Each model is trained with various shared sources k. Figure 4 summarizes the results: the x-axis refers to the hyper-parameter k used for training, and the y-axis denotes the corresponding NRE score on both the training and test data (each with a sample size $N = 1000$), distinguished by the line style. Notably, in all scenarios, the NRE remains low if the number of sources is underestimated and it rises upon overestimating c, particularly when the noise variance is low (left plot). Furthermore, due to overfitting, the NRE score on the training data is increasing with increasing k. Therefore, the NRE on the test data is more suitable for model selection than the NRE on the training data. In contrast, the NRE score on the test data remains stable and attains its minimum at the correct number of sources.

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 views is 10 and the number of sources varies from 10 to 50. Figure 5 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 Assumption. Transcriptome datasets are relevant to the field of genomics. After preprocessing they have the form of random data matrices, where each row corresponds 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.

Consider the cocktail party problem, a classical application of Independent Component Analysis (ICA). Imagine a room filled with speakers and several microphones placed strate-

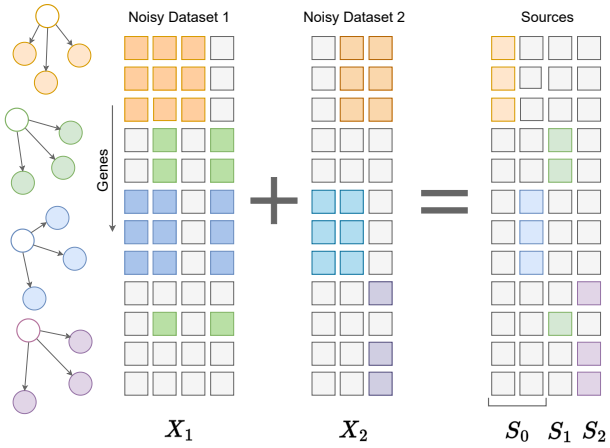

Figure 6: ICA for Data Fusion. We are given two views X_1 and X_2 that share information in terms of gene activity. The common information here is the activation of the orange and blue networks. There are view-specific signals: the green (X_1) and the purple (X_2) networks. The goal of the data integration task is to extract the shared signal S_0 and the view-specific sources S_1 and S_2.

gically to record everyone. Each microphone picks up a different combination of all voices. The objective of ICA, in this context, is to isolate the original speakers' speeches from the mixed microphone signals. Analogously, we can interpret a transcriptome dataset as the microphones in this scenario. Each microphone represents an experimental condition, and similar to how individuals act independently at a party, we posit that the regulators in the cell operate independently. Therefore, each column of the transcriptome data represents a mixed measure of transcriptional regulator effects. This hypothesis has found support in several studies, such as [Sastry et al., 2019, 2021a, Fraunhoffer et al., 2022]. Upon applying ICA, we aspire to recover the sources symbolizing these independent regulatory pathways.

To formalize, let $X \in \mathbb{R}^{n \times p}$ denote a transcriptome data matrix with n samples (or experimental outcomes) and p genes. We assume that the transcriptome matrix follows a linear latent model, meaning there exist matrices $A \in \mathbb{R}^{n \times k}$ and $S \in \mathbb{R}^{k \times p}$ such that $X = AS$. The k components are representations of gene expression levels. If a group of genes appears over or under-expressed in a specific component, they are typically assumed to share a functional property in the genome. Furthermore, if these components operate independently, it implies that they correspond to distinct genetic pathways. In other words, sets of genes that demonstrate overexpression or underexpression in these components are assumed to act independently from each other under the observed experimental conditions.

ICA for Data Fusion. We combine noisy transcriptome datasets by considering that the different datasets (or views) share information but also have view-specific ones. The shared information can be the activity of regulatory pathways invariant across conditions, such as housekeeping genes. In the example shown in Figure 6, we have two noisy datasets X_1 and X_2, presumably provided by different labs. The shared information is the orange and blue networks. After successfully combining the two views, we would like to retrieve these shared sources denoted by S_0 and the dataset-specific ones S_1 and S_2 efficiently. These are colored green and purple in our diagram.

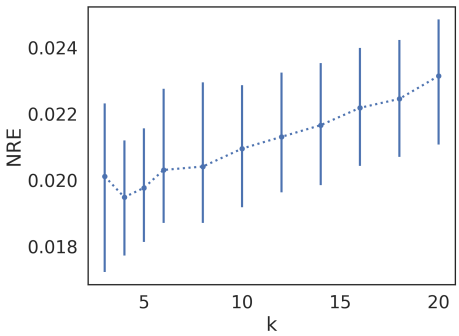

Figure 7: The NRE score computed on test data for the transcriptome data for various k. This procedure was repeated 50 times and the error bars represent the estimated 95% confidence interval.

Datasets. In this example, we consider the bacterium *Bacillus subtilis*, the best studied Gram-positive bacteria which can be found in the soil and human intestines. Our goal is to combine three publicly available datasets (views) provided by different labs [Nicolas et al., 2012, Arrieta-Ortiz et al., 2015, Sastry et al., 2021b]. Each of the datasets contains gene expression levels of about 4000 genes measured across more than 250 experimental outcomes. For a detailed description of the datasets refer to Section 3 of the Supplementary Material.

Model Selection. For the data fusion task, we do not have any prior knowledge about the shared information between the three datasets (three 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 sets 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 7. The NRE score reaches its minimum for $k = 4$, indicating the number of shared sources.

Data Integration for Co-Regulation Inference. Often, the primary objective of transcriptome data fusion studies is

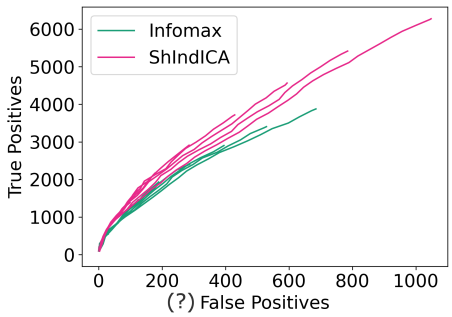

Figure 8: We compare the top ten models with ShIndICA and Infomax. We order the edges from the networks according to their strength. We count the true positives (y-axis) and possibly (?) false positive edges (x-axis) in the first $100, 200, \ldots$ edges. ShIndICA combined with glasso identifies more true positive edges than Infomax with glasso.

to predict novel gene-gene interactions from the data, such as co-regulation patterns. Here, we consider an application for estimating an undirected graph where the nodes represent the genes, and the edges indicate that these genes have a common regulator. Given the high-dimension-low-sample-size nature of transcriptome datasets (i.e., the number of genes exceeds the number of samples), it is often challenging to discern meaningful relationships. A tool like the Graphical Lasso (glasso) [Friedman et al., 2007] becomes particularly valuable here, as it is well-suited for uncovering the underlying graphical structure of the observed data. The output of glasso is a precision matrix from which the graphical structure can be directly determined: each non-zero entry from the precision matrix indicates an edge between the corresponding variables.

In this scenario, instead of feeding the "raw" data samples directly into the glasso, we utilize the components generated by the data integration algorithm. The rationale behind this is that the combined data, resulting from the integration of multiple datasets, can provide a more comprehensive and accurate picture, which, in turn, can enhance the performance of the glasso in identifying the true co-regulation relationships among genes.

Use Case: Co-Regulation in *Bacillus subtilis*. Our analysis focuses on the comparative evaluation of ShIndICA and the naive ICA approach (Infomax as utilized in the previous example) within the defined downstream data integration task. We specify the number of shared sources for ShIndICA to be 5, and for the naive Infomax approach, we set it to 0. After performing PCA whitening on the data, the number of sources per view is brought down to 72, 30, and 59, respectively. A comprehensive description of the exact component selection process is provided in Section 3 of the Supplementary Material.

After applying each method (ShIndICA or Infomax), we fit 30 glasso models, each with varying penalization parameters, on the estimated components. Using a statistical goodness-of-fit measure, we select the top 10 models (for a detailed understanding, refer to Algorithm 1 in Section 3 of the Supplementary Material). To evaluate the efficacy of our results, we gather a ground truth network from the online *Subti*Wiki database [Pedreira et al., 2021], which comprises 5,952 pairs of regulator and regulated genes. As our method predicts pairs of co-regulated genes, we modify the ground truth network into an undirected graph, establishing connections between genes with a shared regulator.

Figure 8 compares the ten output graphs generated by the glasso combined with ShIndICA or Infomax. For every estimated graph, we rank the edges according to their strength and count the true positive (y-axis) and false positive (x-axis) edges within the top $100, 200, \ldots$ edges. ShIndICA outperforms Infomax, as demonstrated by all ten selected glasso models with curves above the ones from the glasso models combined with Infomax. In other words, when paired with glasso, ShIndICA identifies more true positive edges than when glasso is combined with Infomax.

8 DISCUSSION

We introduced ShIndICA, a novel noisy linear Independent Component Analysis (ICA) approach, designed to leverage shared information across different views to extract both shared and view-specific sources effectively. Our method is backed by theoretical guarantees affirming the identifiability of the model's linear structure, latent source, noise distributions, and the number of shared and individual sources. The unmixing matrices are estimated through the maximization of the joint log-likelihood of the observed views, while a novel goodness-of-fit measure guides the selection of the number of shared sources. Our empirical studies demonstrate the robust performance of ShIndICA on simulated data, even in the case of model misspecification. We introduced a novel strategy for merging transcriptome data, showing that our model can align estimated sources with biologically meaningful signals. ShIndICA not only combines transcriptome data effectively but also boosts the effectiveness of the chosen graphical inference models in extracting biologically relevant insights.

Acknowledgements

We sincerely value the contributions made by various individuals to this work. Our profound thanks to Joris Mooij for his insightful discussions and manuscript reviews. We are also grateful to Leendert Hamoen and Martijs Jonker for their valuable perspectives on transcriptome data fusion. Finally, we thank Sara Magliacane for the constructive feedback that greatly improved our manuscript.

References

Pierre Ablin, Jean-François Cardoso, and Alexandre Gramfort. Faster ICA under Orthogonal Constraint. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4464–4468. IEEE, 2018.

Shun-ichi Amari, Andrzej Cichocki, and Howard Yang. A New Learning Algorithm for Blind Signal Separation. *Advances in Neural Information Processing Systems*, 8, 1995.

Matthew Anderson, Tuelay Adali, and Xi-Lin Li. Joint Blind Source Separation with Multivariate Gaussian Model: Algorithms and Performance Analysis. *IEEE Transactions on Signal Processing*, 60(4):1672–1683, 2011.

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.

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep Canonical Correlation Analysis. In *International Conference on Machine Learning*, pages 1247–1255. PMLR, 2013.

Mario L Arrieta-Ortiz, Christoph Hafemeister, Ashley Rose Bate, Timothy Chu, Alex Greenfield, Bentley Shuster, Samantha N Barry, Matthew Gallitto, Brian Liu, Thadeous Kacmarczyk, et al. An experimentally supported model of the Bacillus subtilis global transcriptional regulatory network. *Molecular Systems Biology*, 11(11): 839, 2015.

Francisco Avila Cobos, Jo Vandesompele, Pieter Mestdagh, and Katleen De Preter. Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics*, 34(11):1969–1979, 2018.

Marie-Ming Aynaud, Olivier Mirabeau, Nadege Gruel, Sandrine Grossetête, Valentina Boeva, Simon Durand, Didier Surdez, Olivier Saulnier, Sakina Zaïdi, Svetlana Gribkova, et al. Transcriptional programs define intratumoral heterogeneity of ewing sarcoma at single-cell resolution. *Cell reports*, 30(6):1767–1779, 2020.

Francis R Bach and Michael I Jordan. Kernel Independent Component Analysis. *Journal of Machine Learning Research*, 3(Jul):1–48, 2002.

Francis R Bach and Michael I Jordan. A Probabilistic Interpretation of Canonical Correlation Analysis. *Technical Report 688, Department of Statistics, University of California*, 2005.

Paolo Bartolomeo, Tal Seidel Malkinson, and Stefania De Vito. Botallo's error, or the quandaries of the universality assumption. *Cortex*, 86:176–185, 2017.

Anthony J Bell and Terrence J Sejnowski. An Information-Maximization Approach to Blind Separation and Blind Deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

Vince D Calhoun, Tulay Adali, Godfrey D Pearlson, and James J Pekar. A Method for Making Group Inferences from functional MRI Data Using Independent Component Analysis. *Human Brain Mapping*, 14(3):140–151, 2001.

Michael Cary, Katie Podshivalova, and Cynthia Kenyon. Application of transcriptional gene modules to analysis of caenorhabditis elegans' gene expression data. *G3: Genes, Genomes, Genetics*, 10(10):3623–3638, 2020.

Pierre Comon. Independent Component Analysis, a New Concept? *Signal Processing*, 36(3):287–314, 1994.

Marco Congedo, Roy E John, Dirk De Ridder, and Leslie Prichep. Group Independent Component Analysis of Resting State EEG in Large Normative Samples. *International Journal of Psychophysiology*, 78(2):89–99, 2010.

Mike Davies. Identifiability Issues in Noisy ICA. *IEEE Signal processing letters*, 11(5):470–473, 2004.

Julien Dubois and Ralph Adolphs. Building a science of individual differences from fMRI. *Trends in Cognitive Sciences*, 20(6):425–443, 2016.

Sydney Dubois, Bruno Tesson, Sylvain Mareschal, Pierre-Julien Viailly, Elodie Bohers, Philippe Ruminy, Pascaline Etancelin, Pauline Peyrouze, Christiane Copie-Bergman, Bettina Fabiani, et al. Refining diffuse large b-cell lymphoma subgroups using integrated analysis of molecular profiles. *EBioMedicine*, 48:58–69, 2019.

Jeffrey Durieux and Tom F Wilderjans. 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.

Astrid ME Engberg, Kasper W Andersen, Morten Mørup, and Kristoffer H Madsen. 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.

Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning Robust Representations via Multi-View Information Bottleneck. *ICLR*, 2020.

Nicolas A Fraunhoffer, Analía Meilerman Abuelafia, Martin Bigonnet, Odile Gayet, Julie Roques, Remy Nicolle, Gwen Lomberk, Raul Urrutia, Nelson Dusetti, and Juan Iovanna. Multi-omics data integration and modeling unravels new mechanisms for pancreatic cancer and improves prognostic prediction. *NPJ Precision Oncology*, 6 (1):1–16, 2022.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 12 2007.

Ying Guo and Giuseppe Pagnoni. A Unified Framework for Group Independent Component Analysis for Multi-Subject fMRI Data. *NeuroImage*, 42(3):1078–1093, 2008.

Harold Hotelling. Relations between Two Sets of Variates. In *Breakthroughs in Statistics*, pages 162–190. Springer, 1936.

Rene J Huster, Sergey M Plis, and Vince D Calhoun. Group-level component analyses of EEG: validation and evaluation. *Frontiers in Neuroscience*, 9:254, 2015.

Aapo Hyvärinen and Erkki Oja. Independent Component Analysis: Algorithms and Applications. *Neural Networks*, 13(4-5):411–430, 2000.

Aapo Hyvärinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ICA using Auxiliary Variables and Generalized Contrastive Learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, 2019.

Abram Meerovich Kagan, Yurii Vladimirovich Linnik, Calyampudi Radhakrishna Rao, et al. *Characterization Problems in Mathematical Statistics*. Wiley-Interscience, 1973.

Yoonseop Kang and Seungjin Choi. Restricted Deep Belief Networks for Multi-View Learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part II 22*, pages 130–145. Springer, 2011.

Arto Klami, Seppo Virtanen, and Samuel Kaski. 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.

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.

Mario Lezcano-Casado. Trivializations for Gradient-Based Optimization on Manifolds. In *Advances in Neural Information Processing Systems, NeurIPS*, pages 9154–9164, 2019.

Yifeng Li, Fang-Xiang Wu, and Alioune Ngom. A Review on Machine Learning Principles for Multi-View Biological Data Integration. *Briefings in bioinformatics*, 19(2): 325–340, 2018.

Qunfang Long, Suchita Bhinge, Vince D Calhoun, and Tülay Adali. Independent Vector Analysis for Common Subspace Analysis: Application to Multi-Subject fMRI Data yields Meaningful Subgroups of Schizophrenia. *NeuroImage*, 216:116872, 2020.

Ana S Lukic, Miles N Wernick, Lars Kai Hansen, and Stephen C Strother. An ICA Algorithm for Analyzing Multiple Data Sets. In *Proceedings. International Conference on Image Processing*, volume 2, pages II–II. IEEE, 2002.

Mona Maneshi, Shahabeddin Vahdat, Jean Gotman, and Christophe Grova. Validation of Shared and Specific Independent Component Analysis (SSICA) for Between-Group Comparisons in fMRI. *Frontiers in Neuroscience*, 10:417, 2016.

Martin J McKeown and Terrence J Sejnowski. Independent Component Analysis of fMRI Data: Examining the Assumptions. *Human Brain Mapping*, 6(5-6):368–372, 1998.

Petr V Nazarov, Anke K Wienecke-Baldacchino, Andrei Zinovyev, Urszula Czerwińska, Arnaud Muller, Dorothée Nashan, Gunnar Dittmar, Francisco Azuaje, and Stephanie Kreis. 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.

Pierre Nicolas, Ulrike Mäder, Etienne Dervyn, Tatiana Rochat, Aurélie Leduc, Nathalie Pigeonneau, Elena Bidnenko, Elodie Marchadier, Mark Hoebeke, Stéphane Aymerich, et al. Condition-dependent transcriptome reveals high-level regulatory architecture in Bacillus subtilis. *Science*, 335(6072):1103–1106, 2012.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. *NIPS-W*, 2017.

Tiago Pedreira, Christoph Elfmann, and Jörg Stülke. The current state of SubtiWiki, the database for the model organism Bacillus subtilis. *Nucleic Acids Research*, 50 (D1):D875–D882, 10 2021.

Hugo Richard, Luigi Gresele, Aapo Hyvarinen, Bertrand Thirion, Alexandre Gramfort, and Pierre Ablin. Modeling Shared Responses in Neuroimaging Studies through Multiview ICA. *Advances in Neural Information Processing Systems*, 33:19149–19162, 2020.

Hugo Richard, Pierre Ablin, Bertrand Thirion, Alexandre Gramfort, and Aapo Hyvarinen. Shared Independent Component Analysis for Multi-Subject Neuroimaging. *Advances in Neural Information Processing Systems*, 34: 29962–29971, 2021.

Zeid M Rusan, Michael P Cary, and Roland J Bainton. Granular Transcriptomic Signatures Derived from Independent Component Analysis of Bulk Nervous Tissue for Studying Labile Brain Physiologies. *bioRxiv*, 2020.

Mustafa S Salman, Yuhui Du, Dongdong Lin, Zening Fu, Alex Fedorov, Eswar Damaraju, Jing Sui, Jiayu Chen, Andrew R Mayer, Stefan Posse, 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.

Mathieu Salzmann, Carl Henrik Ek, Raquel Urtasun, and Trevor Darrell. Factorized Orthogonal Latent Spaces. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 701–708. JMLR Workshop and Conference Proceedings, 2010.

Anand V Sastry, Ye Gao, Richard Szubin, Ying Hefner, Sibei Xu, Donghyuk Kim, Kumari Sonal Choudhary, Laurence Yang, Zachary A King, and Bernhard O Palsson. The Escherichia coli transcriptome mostly consists of independently regulated modules. *Nature Communications*, 10(1):1–14, 2019.

Anand V Sastry, Alyssa Hu, David Heckmann, Saugat Poudel, Erol Kavvas, and Bernhard O Palsson. Independent Component Analysis recovers Consistent Regulatory Signals from Disparate Datasets. *PLoS Computational Biology*, 17(2):e1008647, 2021a.

Anand V. Sastry, Saugat Poudel, Kevin Rychel, Reo Yoo, Cameron R. Lamoureux, Siddharth Chauhan, Zachary B. Haiman, Tahani Al Bulushi, Yara Seif, and Bernhard O. Palsson. Mining all publicly available expression data to compute dynamic microbial transcriptional regulatory networks. *bioRxiv*, 2021b.

Age K Smilde, Ingrid Måge, Tormod Naes, Thomas Hankemeier, Mirjam Anne Lips, Henk AL Kiers, Ervim Acar, and Rasmus Bro. Common and Distinct Components in Data Fusion. *Journal of Chemometrics*, 31(7):e2900, 2017.

Justin Tan, Anand V Sastry, Karoline S Fremming, Sara P Bjørn, Alexandra Hoffmeyer, Sangwoo Seo, Bjørn G Voldborg, and Bernhard O Palsson. Independent component analysis of E. coli's transcriptome reveals the cellular processes that respond to heterologous gene expression. *Metabolic Engineering*, 61:360–368, 2020.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive Multiview Coding. In *European Conference on Computer Vision*, pages 776–794. Springer, 2020.

Carlos G Urzúa-Traslaviña, Vincent C Leeuwenburgh, Arkajyoti Bhattacharya, Stefan Loipfinger, Marcel ATM van Vugt, Elisabeth GE de Vries, and Rudolf SN Fehrmann. Improving gene function predictions using independent transcriptional components. *Nature Communications*, 12 (1):1–14, 2021.

Gaël Varoquaux, Sepideh Sadaghiani, Jean Baptiste Poline, and Bertrand Thirion. CanICA: Model-Based Extraction of Reproducible Group-Level ICA Patterns from fMRI Time Series. *arXiv preprint arXiv:0911.4650*, 2009.

Javier Vía, Matthew Anderson, Xi-Lin Li, and Tülay Adalı. 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.

Ricardo Vigário, Veikko Jousmäki, Matti Hämäläinen, Riitta Hari, and Erkki Oja. Independent Component Analysis for Identification of Artifacts in Magnetoencephalographic Recordings. *Advances in Neural Information Processing Systems*, 10, 1997.

Seppo Virtanen. *Bayesian Exponential Family Projections*. PhD thesis, Aalto University, 2010.

Weiran Wang, Xinchen Yan, Honglak Lee, and Karen Livescu. Deep Variational Canonical Correlation Analysis. *arXiv preprint arXiv:1610.03454*, 2016.

Hejia Zhang, Po-Hsuan Chen, Janice Chen, Xia Zhu, Javier S Turek, Theodore L Willke, Uri Hasson, and Peter J Ramadge. A searchlight factor model approach for locating shared information in multi-subject fMRI analysis. *arXiv preprint arXiv:1609.09432*, 2016.

Chun-Hou Zheng, De-Shuang Huang, Xiang-Zhen Kong, and Xing-Ming Zhao. Gene Expression Data Classification Using Consensus Independent Component Analysis. *Genomics, Proteomics & Bioinformatics*, 6(2):74–82, 2008.

Weizhuang Zhou and Russ B Altman. Data-driven human transcriptomic modules determined by Independent Component Analysis. *BMC Bioinformatics*, 19(1):1–25, 2018.