Multi-View Independent Component Analysis with Shared and Individual Sources (Supplementary Material)

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

1 IDENTIFIABILITY RESULTS

Here we cite and correct needed results from [Kagan et al., 1973, Lemma 10.2.3, Theorem 10.3.1]:

Theorem 1.1 (Identifiability for independent non-constant sources [Kagan et al., 1973, Lemma 10.2.3, Theorem 10.3.1])**.** *Let $x \in \mathbb{R}^p$ be a p-dimensional random vector with two representations:*

$$A^{(1)}y^{(1)} + \mu^{(1)} = x = A^{(2)}y^{(2)} + \mu^{(2)}, \tag{1}$$

with the following properties for $i = 1, 2$:

1. *$A^{(i)} \in \mathbb{R}^{p \times k^{(i)}}$ is a (non-random) matrix with non-zero columns and for which no two columns are proportional to each other,*
2. *$\mu^{(i)} \in \mathbb{R}^p$ a (non-random) column vector,*
3. *$y^{(i)} \in \mathbb{R}^{k^{(i)}}$ is a random vector such that:*
   (a) *its $k^{(i)}$ components $\{y_1^{(i)}, \dots, y_{k^{(i)}}^{(i)}\}$ are mutually independent,*
   (b) *each of its components $y_j^{(i)}$ is a non-constant random variable (a.s.), i.e. does not have a delta-peak distribution, $j = 1, \dots, k^{(i)}$.*

Then we have the following:

$$\mu^{(2)} - \mu^{(1)} \in A^{(1)}\mathbb{R}^{k^{(1)}} = A^{(2)}\mathbb{R}^{k^{(2)}}, \qquad \operatorname{rank}(A^{(1)}) = \operatorname{rank}(A^{(2)}). \tag{2}$$

In particular, there exist $c^{(1)} \in \mathbb{R}^{k^{(1)}}$, $c^{(2)} \in \mathbb{R}^{k^{(2)}}$ such that: $\mu^{(2)} - \mu^{(1)} = A^{(1)}c^{(1)} = A^{(2)}c^{(2)}$.

Furthermore, the following statements hold:

1. *If the l-th column of $A^{(2)}$ is not proportional to any column of $A^{(1)}$, then $y_l^{(2)}$ is a normally distributed random variable.*
2. *Assume that the l-th column of $A^{(2)}$ is proportional to the j-th column of $A^{(1)}$ with proportionality constant[1] $0 \neq \lambda \in \mathbb{R}$, i.e.: $a_l^{(2)} = \lambda \cdot a_j^{(1)}$. Then there exists a (complex) polynomial g such that we have the following equation for the characteristic functions of the components $y_l^{(2)}$ and $y_j^{(1)}$ (in a neighborhood of the origin):*

$$\phi_{y_l^{(2)}}(\lambda t) = \phi_{y_j^{(1)}}(t) \cdot \exp(g(t)). \tag{3}$$

   In particular $y_l^{(2)}$ is (non-)normal if and only if $y_j^{(1)}$ is (non-)normal.

[1]Note that this proportionality constant was forgotten to be reintroduced in [Kagan et al., 1973, Theorem 10.3.1] after it was "w.l.o.g." removed in [Kagan et al., 1973, Lemmata 10.2.4, 10.2.5.].

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

The following result is a corollary from the work of [Kagan et al., 1973] and is used for proving the main result of our paper.

Theorem 1.2 (Identifiability of the single view ICA model)**.** *Let $x \in \mathbb{R}^p$ be a random variable. Assume that we have the following two representations of x:*

$$A^{(1)}(y^{(1)} + \epsilon^{(1)}) + b^{(1)} = x = A^{(2)}(y^{(2)} + \epsilon^{(2)}) + b^{(2)}, \tag{4}$$

with the following properties for $i = 1, 2$:

1. *$A^{(i)} \in \mathbb{R}^{p \times k^{(i)}}$ is a (non-random) matrix with full column rank, i.e. $\operatorname{rank}(A^{(i)}) = k^{(i)} \leq p$,*
2. *$b^{(i)} \in \mathbb{R}^p$ a (non-random) column vector,*
3. *$\epsilon^{(i)} \in \mathbb{R}^{k^{(i)}}$ is an uncorrelated k-variate normal random variable: $\epsilon^{(i)} \sim \mathcal{N}(\mu^{(i)}, \Sigma^{(i)})$, with mean $\mu^{(i)} \in \mathbb{R}^{k^{(i)}}$ and a positive-definite diagonal covariance matrix $\Sigma^{(i)} \in \mathbb{R}^{k^{(i)} \times k^{(i)}}$,*
4. *$y^{(i)} \in \mathbb{R}^{k^{(i)}}$ is a random variable such that:*
   (a) *its $k^{(i)}$-components $\{y_1^{(i)}, \ldots, y_{k^{(i)}}^{(i)}\}$ are mutually independent,*
   (b) *each of its component $y_j^{(i)}$ is a non-constant random variable (a.s.), $j = 1, \ldots, k^{(i)}$,*
   (c) *$y^{(i)}$ has no normal components, i.e. if we can write: $y^{(i)} \sim \tilde{y}^{(i)} + \hat{y}^{(i)}$ with $\tilde{y}^{(i)} \perp\!\!\!\perp \hat{y}^{(i)}$, then $\tilde{y}^{(i)}$ and $\hat{y}^{(i)}$ are non-normal,*
5. *$\epsilon^{(i)}$ is independent from $y^{(i)}$: $\epsilon^{(i)} \perp\!\!\!\perp y^{(i)}$.*

Then $k^{(1)} = k^{(2)} =: k$ and there exist a permutation matrix $P \in \mathbb{R}^{k \times k}$, an invertible diagonal matrix $\Lambda \in \mathbb{R}^{k \times k}$ and a column vector $c \in \mathbb{R}^k$ such that:

$$A^{(2)} = A^{(1)} P \Lambda,$$

and such that the corresponding random variables have the same distributions:

$$P\Lambda y^{(2)} + c \sim y^{(1)}, \qquad P\Lambda(\epsilon^{(2)} - \mu^{(2)}) \sim \epsilon^{(1)} - \mu^{(1)}, \qquad P\Lambda \Sigma^{(2)} \Lambda^\top P^\top = \Sigma^{(1)}.$$

Proof. 1. In the first part of our proof we show that $k^{(1)} = k^{(2)} =: k$ and $A^{(2)} = A^{(1)} P \Lambda$ for some permutation matrix $P \in \mathbb{R}^{k \times k}$, an invertible diagonal matrix $\Lambda \in \mathbb{R}^{k \times k}$.

First, for $i = 1, 2$ we state an equivalent formulation of the linear representation of x given in 4. According to [Kagan et al., 1973, Lemma 10.2.3], there exist a constant column vector $c^{(2)} \in \mathbb{R}^{k^{(2)}}$ such that $b^{(2)} - b^{(1)} = A^{(2)} c^{(2)}$. It follows that $\tilde{x} = x - b^{(1)} = A^{(1)}(y^{(1)} + \epsilon^{(1)}) = A^{(2)}(y^{(2)} + \epsilon^{(2)} + c^{(2)})$.

Furthermore, note that if $y^{(i)}$ is non-normal, then the random variables $g^{(1)} = y^{(1)} + \epsilon^{(1)}$ and $g^{(2)} = y^{(2)} + \epsilon^{(2)} + c^{(2)}$ are also non-normal. This follows from the fact that if $g^{(i)}$ is normal then both $y^{(i)}$ and $\epsilon^{(i)}$ would be normal according to the Lévy-Cramér theorem.

Thus, we can apply Theorem 1.1 for the two representations of $\tilde{x}$, $\tilde{x} = A^{(1)} g^{(1)}$ and $\tilde{x} = A^{(2)} g^{(2)}$. Since every component of $g^{(i)}$ is non-normal, it follows that every column of $A^{(1)}$ is proportional to a column of $A^{(2)}$ and vice versa.

Now assume w.l.o.g that $k^{(1)} > k^{(2)}$. Then, there exist two columns of $A^{(1)}$ that are proportional to a column of $A^{(2)}$. However, this is a contradiction to assumption 1. that the matrix $A^{(1)}$ has full column rank.

Thus, it follows that $k^{(1)} = k^{(2)} =: k$ and $A^{(2)} = A^{(1)} P \Lambda$ for some permutation matrix $P \in \mathbb{R}^{k \times k}$, an invertible diagonal matrix $\Lambda \in \mathbb{R}^{k \times k}$. Moreover,

$$A^{(1)}(y^{(1)} + \epsilon^{(1)}) = A^{(1)} P \Lambda (y^{(2)} + \epsilon^{(2)} + c^{(2)}).$$

Multiplying with $(A^{(1),\top} A^{(1)})^{-1} A^{(1),\top}$, which gives:

$$y^{(1)} + \epsilon^{(1)} = P\Lambda(y^{(2)} + \epsilon^{(2)} + c^{(2)}).$$

2. In the remaining we show that there exists a column vector c such that $y^{(1)} \sim P\Lambda(y^{(2)} + c^{(2)}) + c$ and $\epsilon^{(1)} - \mu^{(1)} \sim P\Lambda(\epsilon^{(2)} - \mu^{(2)})$ (or equivalently $\Sigma^{(1)} = P\Lambda\Sigma^{(2)}\Lambda^\top P^\top$). Now, define $\tilde{y}^{(2)} = P\Lambda y^{(2)}$, $\tilde{c}^{(2)} = P\Lambda c^{(2)}$ and $\tilde{\epsilon}^{(2)} = P\Lambda\epsilon^{(2)}$ which is normally distributed with mean $\tilde{\mu^{(2)}} = P\Lambda\mu^{(2)}$ and a diagonal covariance matrix $\tilde{\Sigma}^{(2)} = P\Lambda\Sigma^{(2)}\Lambda^\top P^\top$.

Define the characteristic functions of $y^{(1)}, \tilde{y}^{(2)}, \epsilon^{(1)}, \tilde{\epsilon}^{(2)}$ as $\phi_{y^{(1)}}(\cdot), \phi_{\tilde{y}^{(2)}}(\cdot), \phi_{\epsilon^{(1)}}(\cdot), \phi_{\tilde{\epsilon}^{(2)}}(\cdot) : \mathbb{R}^k \to \mathbb{R}$, from assumption 5. it follows that

$$\phi_{\epsilon^{(1)}}(t)\phi_{y^{(1)}}(t) = e^{it^\top \tilde{c}^{(2)}}\phi_{\tilde{\epsilon}^{(2)}}(t)\phi_{\tilde{y}^{(2)}}(t)$$
$$\phi_{\epsilon^{(1)}}(t)\prod_{i=1}^{k}\phi_{y_i^{(1)}}(t_i) = e^{it^\top \tilde{c}^{(2)}}\phi_{\tilde{\epsilon}^{(2)}}\prod_{i=1}^{k}\phi_{\tilde{y}_i^{(2)}}(t_i)$$

The last equation follows from assumption 4a. Now set $t_i = 0$ for all $i \neq 1$. We get for all t_1

$$\exp(it_1\mu_1^{(1)} - \Sigma_{11}^{(1)}t_1^2)\phi_{y_1^{(1)}}(t_1) = \exp(it_1\tilde{c}_1^{(2)})\exp(it_1\tilde{\mu}_1^{(2)} - \tilde{\Sigma}_{11}^{(2)}t_1^2)\phi_{\tilde{y}_1^{(2)}}(t_1).$$

W.l.o.g. we assume $0 < \Sigma_{11}^{(1)} < \tilde{\Sigma}_{11}^{(2)}$. Thus, the characteristic function given by $\exp(-(\tilde{\Sigma}_{11}^{(2)} - \Sigma_{11}^{(1)})t_1^2)$ is a well defined characteristic function of a normally distributed random variable with mean 0 and variance $\tilde{\Sigma}_{11}^{(2)} - \Sigma_{11}^{(1)}$. Then, the characteristic function of $y_1^{(1)}$ is proportional to a product of the characteristic functions of $\tilde{y}_1^{(2)}$ and a Gaussian random variable. This is a contradiction to the assumption that $y_1^{(1)}$ does not have a normal component (assumption 4c). It follows that, $\Sigma_{11}^{(1)} = \tilde{\Sigma}_{11}^{(2)}$ and for all $t_1 \in \mathbb{R}$ $\phi_{y_1^{(1)}}(t_1) = \exp it_1(\tilde{c}_1^{(2)} + \tilde{\mu}_1^{(2)} - \mu_1^{(1)})\phi_{\tilde{y}_1^{(2)}}(t_1)$, i.e. $\tilde{y}_1^{(2)} + c_1 \sim y_1^{(1)}$ where $c_1 = \tilde{c}_1^{(2)} + \tilde{\mu}_1^{(2)} - \mu_1^{(1)}$. The remaining statements can be proven analogously.

□

1.1 PROOF OF THEOREM 3.1

Proof. First, we can directly apply Theorem 1.2 to every single view $d, d \in \{1, \ldots, D\}$ which ensures the identifiability of the mixing matrices up to permutation and scaling, i.e. there exists a permutation matrix P_d and an invertible diagonal matrix Λ_d such that $A_d^{(2)} = A_d^{(1)}P_d\Lambda_d$ and $\operatorname{rank}(A_d^{(2)}) = \operatorname{rank}(A_d^{(1)}) = k_d$.

W.l.o.g., let $c^{(1)} > c^{(2)}$. That means that the shared sources in representation (1) are more than the ones in representation (2). It follows, according to Theorem 1.1, that there exist a component of the shared sources from (1) and an individual component from (2) in every view such that they are both proportional. More precisely, for any $d \in \{1, \ldots, D\}$ there exist $k, l \in \{1, \ldots, k_d\}$ such that $s_{0k}^{(1)}$ is a component of the shared sources $s_0^{(1)}$ and $s_{dl}^{(2)}$ is a component from the individual sources $s_d^{(2)}$ such that $s_{0k}^{(1)} + \epsilon_{d0k}^{(1)} = (\Lambda_d)_{ll}(s_{dl}^{(2)} + \epsilon_{d1l}^{(2)})$. Let $r \neq d$ be another view such that there exist $m \in \{1, \ldots, k_r\}$ with $s_{mr}^{(2)}$ being an individual component and $s_{0k}^{(1)} + \epsilon_{r0k}^{(1)} = (\Lambda_d)_{mm}(s_{rm}^{(2)} + \epsilon_{r1m}^{(2)})$. This is contradiction to the assumption that $s_{rm}^{(2)} \perp\!\!\!\perp s_{dl}^{(2)}$. It follows that $c^{(1)} = c^{(2)}$.

Furthermore, $\operatorname{Var}(x_d) = \sigma_d^{(1)2}A_d^{(1)}A_d^{(1),\top} = \sigma_d^{(2)2}A_d^{(2)}A_d^{(2),\top} = \sigma_d^{(2)2}A_d^{(1)}P_d\Lambda_d^2P_d^\top A_d^{(1),\top}$. Multiplying with $A_d^{(1),\dagger} = (A_d^{(1)\top}A_d^{(1)})^{-1}A_d^{(1)\top}$ from left and $A_d^{(1),\dagger,^\top} = A_d^{(1)}(A_d^{(1)\top}A_d^{(1)})^{-1}$ from right yields $\sigma_d^{(1)2}\mathbb{I}_{k_d} = \sigma_d^{(2)2}P_d\Lambda_d^2P_d^\top$. It follows that $\frac{\sigma_d^{(2)2}}{\sigma_d^{(1)2}}\Lambda_d^2 = \mathbb{I}_{k_d}$. Computing the covariance between two different views $d, l \in \{1, \ldots, D\}$ gives

$$\operatorname{Cov}(x_d, x_l) = A_{d0}^{(1)}A_{l0}^{(1),\top} = A_{d0}^{(2)}A_{l0}^{(2),\top} = A_{d0}^{(1)}\Lambda_d[c,c]\Lambda_l[c,c]A_{l0}^{(1),\top}$$

where $\Lambda_d[c,c]$ is an invertible diagonal matrix composed by the first c columns and rows of the matrix Λ_d. By multiplying with the left-inverse of $A_{d0}^{(1)}$ from the left and right-inverse of $A_d^{(1),\top}$ from the right, we get for any d and l $\Lambda_d[c,c]\Lambda_l[c,c] = \mathbb{I}_c$. It follows that all entries of Λ_d equal 1 or -1 and therefore $\frac{\sigma_d^{(2)2}}{\sigma_d^{(1)2}} = 1$ for every d.

In the remaining, we will show that the distribution of the sources is identifiable even in the cases when they have normal components. Let $s_i^{(1)}$ be component from $\tilde{s}_i^{(1)}$. Furthermore, there exist $j \in \{1, \ldots, k_d\}$ such that $s_i^{(1)} + \epsilon_i^{(1)} = s_j^{(2)} + \epsilon_j^{(2)}$.

Taking the characteristic functions from both sides yields

$$\phi_{s_i^{(1)}}(t)\phi_{\epsilon_i^{(1)}}(t) = \phi_{s_j^{(2)}}(t)\phi_{\epsilon_j^{(2)}}(t)$$

Since $\sigma_d^{(1)2} = \sigma_d^{(2)2}$ and the noise and sources are with 0 mean, the above equation simplifies to $\phi_{s_i^{(1)}}(t) = \phi_{s_j^{(2)}}(t)$, i.e. $\phi_{s_i^{(1)}}(t) \sim \phi_{s_j^{(2)}}(t)$. □

1.2 ADDITIONAL RESULTS

Theorem 1.3. *Let $x_1, \ldots, x_D$ for $D \geq 3$ be random vectors which are generated according to the model defined in Equation 1. Furthermore, we assume that we have the following two representations of $x_1, \ldots, x_D$ according to Equation 1:*

$$A_{d0}^{(1)} s_0^{(1)} + A_{d1}^{(1)} s_d^{(1)} + A_d^{(1)} \epsilon_d^{(1)} = x_d = A_{d0}^{(2)} s_0^{(2)} + A_{d1}^{(2)} s_d^{(2)} + A_d^{(2)} \epsilon_d^{(2)}, \qquad d \in \{1, \ldots, D\},$$

Additionally, to the assumptions of Equation 1 it holds that

1. *each of the components $s_{dj}^{(i)}$ of $s_d^{(i)}$ for $j = 1, \ldots, k_d^{(i)} - c^{(i)}$ is non-Gaussian.*
2. *$s_0^{(i)}$ can have Gaussian components. Furthermore, if the number of Gaussian components exceeds 2, for all $k, l \in \{1, \ldots, c\}$ with $k \neq l$ it holds that $\gamma_k^{(i)} \neq \gamma_l^{(i)}$, where $\gamma_k^{(i)}$ and $\gamma_l^{(i)}$ are the variances of the components $s_{0k}^{(i)}$ and $s_{0l}^{(i)}$*

Then, for fixed number of shared sources c and for all $d = 1, \ldots, D$ $k_d^{(1)} = k_d^{(2)} = k_d$, and there exist a permutation matrix $P_d \in \mathbb{R}^{k_d \times k_d}$ and an ivertible diagonal matrix $\Lambda_d \in \mathbb{R}^{k_d \times k_d}$ such that

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

Proof. Theorem 1.1 yields that if the individual components are not normal, then for each column of $a_j^{(1)}$ of $A_{d1}^{(1)}$ there is a column $a_i^{(2)}$ of $A_{d1}^{(1)}$ such that there exist $\lambda \neq 0$ with $a_j^{(2)} = \lambda a_j^{(1)}$. Since all mixing matrices have full column rank, it follows that there is one-to-one correspondence between the columns of $A_{d1}^{(1)}$ and the columns of $A_{d1}^{(2)}$, and thus $k_d^{(1)} = k_d^{(2)}$

If at most one of the shared components is normal please refer to Comon [1994]. Now consider the case when at least two components are normal. First, the number of normal components in both representations is the same since c is fixed and the number of non-normal components is identifiable with the same arguments as above.

Computing the covariance between two different views $d, l \in \{1, \ldots, D\}$ yields

$$\mathrm{Cov}(x_d, x_l) = A_{d0}^{(1)} \Gamma^{(1)} A_{l0}^{(1),\top} = A_{d0}^{(2)} \Gamma^{(2)} A_{l0}^{(2),\top}$$

where $\Gamma^{(i)}$ is the covariance matrix of $s_0^{(i)}$ for $i = 1, 2$. We define $A_{d0}^{\gamma,(i)} = A_{d0}^{(i)} \Gamma^{(i)\frac{1}{2}}$ for any $d \in \{1, \ldots, D\}$. Let $P_d = (A_{d0}^{\gamma,(1),\top} A_{d0}^{\gamma,(1)})^{-1} A_{d0}^{\gamma,(1),\top} A_{d0}^{\gamma,(2)}$. Following the proof of Theorem 1 [Richard et al., 2021] we get that $P_d P_l^\top = \mathbb{I}_c = P_d P_k^\top = P_k P_l^\top$ for any $d, k, l \in \{1, \ldots, D\}$. Thus, $P_l = P_d = P_k = P$ and they are orthogonal. Moreover, for all $d = 1, \ldots, D$ it holds $\tilde{s}_0^{(1)} + \tilde{\epsilon}_d^{(1)} = P(\tilde{s}_0^{(2)} + \tilde{\epsilon}_d^{(2)})$ where $\tilde{\epsilon}_d^{(i)} \sim \mathcal{N}(0, \sigma_d^{(i)2} \Gamma^{(i)-1})$ and $\tilde{s}_0^{(i)} = \Gamma^{(i)-\frac{1}{2}} s_0^{(i)}$. From the last equation it follows that $\sigma_d^{(1)2} \Gamma^{(1)-1} = P(\sigma_d^{(2)2} \Gamma^{(2)-1}) P^\top$. Lemma 2 [Richard et al., 2021] implies that P is a sign and permutation matrix. □

2 JOINT DATA LOG-LIKELIHOOD

Lemma 2.1. *Let $W \in \mathbb{R}^{c\times k}$ such that $WW^\top = \mathbb{I}_c$ and $x^1, \dots, x^N \in \mathbb{R}^k$ such that for every $j = 1, \dots, k$, we have $\sum_{i=1}^N (x_j^i)^2 = 1$ and for every $j \neq k$, we have $\sum_{i=1}^N x_j^i x_k^i = 0$. Then for every $j = 1, \dots, c$, it also holds that $\sum_{i=1}^N ((Wx^i)_j)^2 = 1$.*

Proof. Let W_j be the $j-$th row of W. Then

$$\begin{aligned}\sum_{i=1}^N ((Wx^i)_j)^2 &= \sum_{i=1}^N (\sum_{l=1}^k W_{jl}x_l^i)^2 = \sum_{i=1}^N \sum_{l=1}^k \sum_{r=1}^k W_{jl}x_l^i W_{jr}x_r^i \\ &= \sum_{l=1}^k \sum_{r=1}^k W_{jl}W_{jr} \sum_{i=1}^N x_l^i x_r^i = \sum_{l=1}^k \sum_{r=1}^k W_{jl}W_{jr}\delta_{lr} = \sum_{r=1}^k W_{jr}^2 = 1\end{aligned}$$

where $\delta_{lr} = 1$ if $l = r$ and 0 otherwise. For the fourth equation we used that $\sum_{i=1}^N (x_j^i)^2 = 1$ and $\sum_{i=1}^N x_j^i x_k^i = 0$ for all $j \neq k$; and for the last one we used $WW^\top = \mathbb{I}_c$. □

2.1 DERIVATION

Under the generative model assumptions and optimization constraints stated in Section 4, it holds

$$\mathcal{L}(W_1, \dots, W_D) = \sum_{i=1}^N \log f(\bar{z}_0^i) + \sum_{i=1}^N \sum_{d=1}^D \log p_{Z_{d,1}}(z_{d,1}^i) + N \sum_{d=1}^D \log |W_d| \tag{5}$$

$$- \frac{1}{2\sigma^2}\Big(\sum_{d=1}^D \operatorname{trace}(Z_{d,0} Z_d^{(1)\top}) - \frac{1}{D} \sum_{d=1}^D \sum_{l=1}^D \operatorname{trace}(Z_{d,0} Z_{l,0}^\top)\Big) \tag{6}$$

Proof. Let $\mathbf{x} = (x_1^\top, x_2^\top, \dots, x_D^\top)^\top \in \mathbb{R}^{K_D}$, $\tilde{\mathbf{s}} = (\tilde{s}_1^\top, \tilde{s}_2^\top, \dots, \tilde{s}_D^\top)^\top \in \mathbb{R}^{K_D}$, $\epsilon = (\epsilon_1^\top, \epsilon_2^\top, \dots, \epsilon_D^\top)^\top \in \mathbb{R}^{K_D}$, where $K_D = \sum_{d=1}^D k_d$ and for $W_d = A_d^{-1}$ define

$$\mathbf{W} = \begin{pmatrix} W_1 & 0 & \dots & 0 & 0 \\ 0 & W_2 & \dots & 0 & 0 \\ & & \ddots & & \\ 0 & 0 & \dots & W_{D-1} & 0 \\ 0 & 0 & \dots & 0 & W_D \end{pmatrix}, \ \mathbf{A} = \begin{pmatrix} A_1 & 0 & \dots & 0 & 0 \\ 0 & A_2 & \dots & 0 & 0 \\ & & \ddots & & \\ 0 & 0 & \dots & A_{D-1} & 0 \\ 0 & 0 & \dots & 0 & A_D \end{pmatrix}.$$

Furthermore, 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. $z_d = (z_{d,0}, z_{d,1})^\top$. Let $p_{\mathbf{X}}$ be the joint distribution of $x_1, \dots, x_D$, $p_{\mathbf{Z}}$ the joint distribution of $z_1, \dots, z_D$, $p_{\mathbf{Z}_0}$ the joint distribution of $z_{1,0}, \dots, z_{D,0}$, $p_{\mathbf{Z}_1}$ the joint distribution of $z_{1,1}, \dots, z_{D,1}$ and $p_{Z_{d,1}}$ the probability distribution of $z_{d,1}$.

Note that the model in Equation 1 is equivalent to $\mathbf{x} = \mathbf{A}\mathbf{z}$. By multiplying with the inverse of $\mathbf{A}$ (i.e. $\mathbf{W}$) from the left we get $\mathbf{W}\mathbf{x} = \mathbf{z}$. Then for the joint likelihood of $x_1, \dots, x_D$ we get

$$\begin{aligned} p_{\mathbf{X}}(\mathbf{x}) &= p_{\mathbf{Z}}(\mathbf{z})|\mathbf{W}| \\ &= p_{\mathbf{Z}}(\mathbf{z}) \prod_{d=1}^D |W_d| \\ &= p_{\mathbf{Z}_0}(z_{1,0}, \dots, z_{D,0}) p_{\mathbf{Z}_1}(z_{1,1}, \dots, z_{D,1}) \prod_{d=1}^D |W_d| \\ &= p_{\mathbf{Z}_0}(z_{1,0}, \dots, z_{D,0}) \prod_{d=1}^D p_{Z_{d,1}}(z_{d,1}) \prod_{d=1}^D |W_d|. \end{aligned}$$

1. Second equation: $\mathbf{W}$ is a block diagonal matrix and for all $d = 1, \dots, D$, and $W_d \in \mathbb{R}^{k_d \times k_d}$.
2. Third equation: $z_{1,0}, \dots, z_{D,0} \perp\!\!\!\perp z_{1,1}, \dots, z_{D,1}$.
3. Fourth equation follows from the fact that $z_{1,1}, \dots, z_{D,1}$ are mutually independent since $\{s_{1i}\}_{i=1}^{k_1-c}, \dots \{s_{Di}\}_{i=1}^{k_D-c}, \{\epsilon_{1i}\}_{i=1}^{k_1}, \dots, \{\epsilon_{Di}\}_{i=1}^{k_D}$ are mutually independent.

It follows that

$$\begin{aligned}
p_{\mathbf{Z}_0}(z_{1,0}, \dots, z_{D,0}) &= \int p_{\mathbf{Z}_0|S_0}(z_{1,0}, \dots, z_{D,0}|s_0) p_{S_0}(s_0) ds_0 \\
&= \int \Big(\prod_{d=1}^{D} \mathcal{N}(z_{d,0}; s_0, \sigma^2 \mathbb{I}_c) \Big) p_{S_0}(s_0) ds_0 \\
&\propto \int \exp\Big(- \sum_{d=1}^{D} \frac{\|z_{d,0} - s_0\|^2}{2\sigma^2} \Big) p_{S_0}(s_0) ds_0 \\
&= \int \exp\Big(- \frac{D\|s_0 - \bar{z}_0\|^2 + \sum_{d=1}^{D} \|z_{d,0} - \bar{z}_0\|^2}{2\sigma^2} \Big) p_{S_0}(s_0) ds_0 \\
&= \exp\Big(- \frac{\sum_{d=1}^{D} \|z_{d,0} - \bar{z}_0\|^2}{2\sigma^2} \Big) \int \exp\Big(- \frac{D\|s_0 - \bar{z}_0\|^2}{2\sigma^2} \Big) p_{S_0}(s_0) ds_0
\end{aligned}$$

where $\bar{z}_0 = \frac{1}{D} \sum_{d=1}^{D} z_{d,0}$.

- For the second and third equation recall that $z_{d,0} = s_0 + \epsilon_{d0} \in \mathbb{R}^c$, where $\epsilon_{d0} \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_c)$ and $s_0 \perp\!\!\!\perp \epsilon_{d0}$. This means that $z_{d,0}|s_0 \sim \mathcal{N}(s_0, \sigma^2 \mathbb{I}_c)$. From the following equations follow

$$\begin{aligned}
p_{\mathbf{Z}_0|S_0}(z_{1,0}, \dots, z_{D,0}|s_0) &= \prod_{d=1}^{D} p_{Z_{d,0}|s_0}(z_{d,0}|S_0) \\
&= \prod_{d=1}^{D} \mathcal{N}(z_{d,0}; s_0, \sigma^2 \mathbb{I}_c)
\end{aligned}$$

- The fourth equation results from

$$\begin{aligned}
\sum_{d=1}^{D} \|z_{d,0} - s_0\|^2 &= \sum_{d=1}^{D} \|z_{d,0} - \bar{z}_0 + \bar{z}_0 - s_0\|^2 = \sum_{d=1}^{D} \Big(\|z_{d,0} - \bar{z}_0\|^2 + 2\langle z_{d,0} - \bar{z}_0, \bar{z}_0 - s_0 \rangle + \|\bar{z}_0 - s_0\|^2 \Big) \\
&= \sum_{d=1}^{D} \|z_{d,0} - \bar{z}_0\|^2 + 2 \sum_{d=1}^{D} \langle z_{d,0} - \bar{z}_0, \bar{z}_0 - s_0 \rangle + D\|\bar{z}_0 - s_0\|^2 \\
&= \sum_{d=1}^{D} \|z_{d,0} - \bar{z}_0\|^2 + 2 \Big\langle \sum_{d=1}^{D} z_{d,0} - D \cdot \frac{1}{D} \sum_{d=1}^{D} z_{d,0}, \bar{z}_0 - s_0 \Big\rangle + D\|\bar{z}_0 - s_0\|^2 \\
&= \sum_{d=1}^{D} \|z_{d,0} - \bar{z}_0\|^2 + D\|\bar{z}_0 - s_0\|^2.
\end{aligned}$$

We define $f(\bar{z}_0) = \int \exp\Big(- \frac{D\|s_0 - \bar{z}_0\|^2}{2\sigma^2} \Big) p_{S_0}(s_0) ds_0$ similarly to [Richard et al., 2020].

Note that

$$\|z_{d,0} - \bar{z}_0\|^2 = \|z_{d,0}\|^2 - \frac{2}{D} \sum_{l=1}^{D} \langle z_{d,0}, z_{l,0} \rangle + \frac{1}{D^2} \sum_{l=1}^{D} \sum_{r=1}^{D} \langle z_{r,0}, z_{l,0} \rangle.$$

Thus, it follows that

$$\begin{aligned}\sum_{d=1}^{D}\|z_{d,0}-\bar{z}_0\|^2 &= \sum_{d=1}^{D}\Big(\|z_{d,0}\|^2-\frac{2}{D}\sum_{l=1}^{D}\langle z_{d,0},z_{l,0}\rangle+\frac{1}{D^2}\sum_{l=1}^{D}\sum_{r=1}^{D}\langle z_{r,0},z_{l,0}\rangle\Big)\\ &= \sum_{d=1}^{D}\|z_{d,0}\|^2-\frac{2}{D}\sum_{d=1}^{D}\sum_{l=1}^{D}\langle z_{d,0},z_{l,0}\rangle+D\frac{1}{D^2}\sum_{l=1}^{D}\sum_{r=1}^{D}\langle z_{r,0},z_{l,0}\rangle\\ &= \sum_{d=1}^{D}\|z_{d,0}\|^2-\frac{1}{D}\sum_{d=1}^{D}\sum_{l=1}^{D}\langle z_{d,0},z_{l,0}\rangle\end{aligned}$$

Collecting all terms together we get

$$p_{\mathbf{X}}(\mathbf{x})=\exp\Big(-\frac{\sum_{d=1}^{D}\|z_{d,0}\|^2-\frac{1}{D}\sum_{d=1}^{D}\sum_{l=1}^{D}\langle z_{d,0},z_{l,0}\rangle}{2\sigma^2}\Big)f(\bar{z}_0)\prod_{d=1}^{D}p_{Z_{d,1}}(z_{d,1})\prod_{d=1}^{D}|W_d|$$

The data log-likelihood can be expressed as

$$\begin{aligned}\sum_{i=1}^{N}\log p_{\mathbf{X}}(x_1^i,\dots,x_D^i) &= \sum_{i=1}^{N}\Big(-\frac{\sum_{d=1}^{D}\|z_{d,0}^i\|^2-\frac{1}{D}\sum_{d=1}^{D}\sum_{l=1}^{D}\langle z_{d,0}^i,z_{l,0}^i\rangle}{2\sigma^2}\\ &\quad+\log f(\bar{z}_0^i)+\sum_{d=1}^{D}\log p_{Z_{d,1}}(z_{d,1}^i)+\sum_{d=1}^{D}\log|W_d|\Big)\\ &= \sum_{i=1}^{N}\log f(\bar{z}_0^i)+\sum_{i=1}^{N}\sum_{d=1}^{D}\log p_{Z_{d,1}}(z_{d,1}^i)+N\sum_{d=1}^{D}\log|W_d|\\ &\quad-\frac{1}{2\sigma^2}\Big(\sum_{i=1}^{N}\sum_{d=1}^{D}\|z_{d,0}^i\|^2-\frac{1}{D}\sum_{i=1}^{N}\sum_{d=1}^{D}\sum_{l=1}^{D}\langle z_{d,0}^i,z_{l,0}^i\rangle\Big)\\ &= \sum_{i=1}^{N}\log f(\bar{z}_0^i)+\sum_{i=1}^{N}\sum_{d=1}^{D}\log p_{Z_{d,1}}(z_{d,1}^i)+N\sum_{d=1}^{D}\log|W_d|\\ &\quad-\frac{1}{2\sigma^2}\Big(\sum_{d=1}^{D}\operatorname{trace}(Z_{d,0}Z_{d,0}^\top)-\frac{1}{D}\sum_{d=1}^{D}\sum_{l=1}^{D}\operatorname{trace}(Z_{d,0}Z_{l,0}^\top)\Big)\end{aligned}$$

In the case when the data is pre-whitened, it holds that the unknown unmixing matrices are orthogonal, i.e. $W_dW_d^\top = W_d^\top W_d = \mathbb{I}_{k_d}$ and $|\det W_d| = 1$, and x_d and z_d are uncorrelated. Note that in the main paper, we used a different notation for the mixing matrices and sources to stress the difference before and after whitening. This notation is here omitted for simplicity.

Making similar observations as before, we get for the joint probability of the multiple views:

$$p_{\mathbf{X}}(\mathbf{x})=p_{\mathbf{Z}_0}(z_{1,0},\dots,z_{D,0})\prod_{d=1}^{D}p_{Z_{d,1}}(z_{d,1})$$

Note that after whitening $z_{d,0}=\alpha(\sigma)(s_0+\epsilon_{d0})$ with $\alpha(\sigma)=(1+\sigma^2)^{-\frac{1}{2}}$. With similar observations as above we get

$$\begin{aligned}p_{\mathbf{Z}_0|s_0}(z_{1,0},\dots,z_{D,0}|s_0) &= p_{\mathbf{Z}_0|s_0}(\alpha(\sigma)(s_0+\epsilon_{10}),\dots,\alpha(\sigma)(s_0+\epsilon_{D0})|s_0)=\prod_{d=1}^{D}p_{Z_{d,0}|S_0}(\alpha(\sigma)(s_0+\epsilon_{d0})|s_0)\\ &= \prod_{d=1}^{D}\mathcal{N}(\alpha(\sigma)(s_0+\epsilon_{d0});s_0,\sigma^2\mathbb{I}_c)=\prod_{d=1}^{D}\mathcal{N}(z_{d,0};\alpha(\sigma)s_0,\alpha(\sigma)^2\sigma^2\mathbb{I}_c)\end{aligned}$$

It follows that

$$
\begin{aligned}
p_{\mathbf{Z}_0}(z_{1,0},\dots,z_{D,0}) &= \int p_{\mathbf{Z}_0|s_0}(z_{1,0},\dots,z_{D,0}|s_0)p_{S_0}(s_0)ds_0 \\
&= \int \Big(\prod_{d=1}^{D}\mathcal{N}(z_{d,0};\alpha(\sigma)s_0,\alpha(\sigma)^2\sigma^2\mathbb{I}_c)\Big)p_{S_0}(s_0)ds_0 \\
&\propto \int \exp\Big(-\sum_{d=1}^{D}\frac{\|z_{d,0}-\alpha(\sigma)s_0\|^2}{2\alpha(\sigma)^2\sigma^2}\Big)p_{S_0}(s_0)ds_0 \\
&= \int \exp\Big(-\frac{D\|\alpha(\sigma)s_0-\bar{z}_0\|^2+\sum_{d=1}^{D}\|z_{d,0}-\bar{z}_0\|^2}{2\alpha(\sigma)^2\sigma^2}\Big)p_{S_0}(s_0)ds_0 \\
&= \exp\Big(-\frac{\sum_{d=1}^{D}\|z_{d,0}-\bar{z}_0\|^2}{2\alpha(\sigma)^2\sigma^2}\Big)\int \exp\Big(-\frac{D\|\alpha(\sigma)s_0-\bar{z}_0\|^2}{2\alpha(\sigma)^2\sigma^2}\Big)p_{S_0}(s_0)ds_0
\end{aligned}
$$

where $\bar{z}_0 = \frac{1}{D}\sum_{d=1}^{D} z_{d,0}$. We define $f_\sigma(\bar{z}_0) = \int \exp\Big(-\frac{D\|\alpha(\sigma)s_0-\bar{z}_0\|^2}{2\alpha(\sigma)^2\sigma^2}\Big)p_{S_0}(s_0)ds_0 = \int \exp\Big(-\frac{D\|s_0-(1+\sigma^2)^{\frac{1}{2}}\bar{z}_0\|^2}{2\sigma^2}\Big)p_{S_0}(s_0)ds_0$. For the data log-likelihood we get

$$
\begin{aligned}
\sum_{i=1}^{N}\log p_{\mathbf{x}}(x_1^i,\dots,x_D^i) = \sum_{i=1}^{N}\log f_\sigma(\bar{z}_0^i) + \sum_{i=1}^{N}\sum_{d=1}^{D}\log p_{Z_{d,1}}(z_{d,1}^i) - N\cdot D\cdot 1 \\
- \frac{D\cdot c}{2\alpha(\sigma)\sigma^2} + \frac{1}{2D\alpha(\sigma)^2\sigma^2}\sum_{d=1}^{D}\sum_{l=1}^{D}\mathrm{trace}(Z_{d,0}Z_{l,0}^\top)
\end{aligned}
$$

It be easily derived from 5 by making the following observations resulting from whitening

- $N\sum_{d=1}^{D}\log|W_d| = ND$ since $\forall d\ W_d$ is orthogonal
- $\mathrm{trace}(Z_{d,0}Z_{d,0}^\top) = c$ due to Lemma 2.1

□

3 REAL DATA EXPERIMENT

3.1 DATA ACQUISITION AND PREPROCESSING

A transcriptome dataset resembles a random data matrix (see Figure 1). Each column represents an experimental condition (such as knock-out or stress conditions) that cells were subjected to, and each row represents a gene. So each entry of this matrix is an expression value indicating a gene activity under a given condition (typically measured using RNA sequencing or microarrays).

Our analysis is primarily based on three large gene expression data sets, denoted by (in our code) Dataset1[2] [Arrieta-Ortiz et al., 2015] with 265 transcriptome datasets obtained from 38 unique experimental designs and Dataset2 [Nicolas et al., 2012][3] containing 262 samples from 104 different experimental conditions and Dataset3[4] collected and preprocessed RNA-seq data by [Sastry et al., 2021] with 265 samples of 93 unique conditions. We removed genes with missing values from Dataset 1 and we selected 3990 genes that are present in all three datasets.

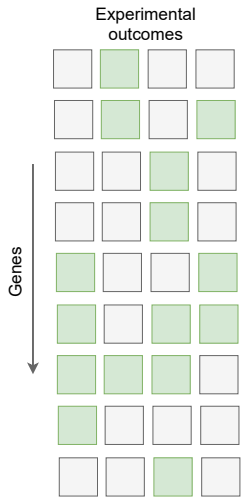


Figure 1: Example of Transcriptome Data

3.2 GENE-GENE INTERACTION PIPELINE

The main steps of our methodology are presented in Algorithm 1. First, we select the number of total and shared components. We infer latent components from the data as described in the main paper, Section 7.2. Afterward, we learn a sparse undirected graph from the estimated independent components (see Section 3.2.2).

3.2.1 Selection of Total Number of ICA Components

To select the total number of components for each single dataset, we utilize the following heuristic:

1. Estimate the sources S and the mixing matrix A from the observed data X via FastICA or another related method.
2. For each component S_k, estimate its relevance by computing $r_k = \sum_i (A_{ki})^2$.
3. Order the components' relevance from the highest to the lowest value and scale them to sum to 1, i.e. $r = \text{orderDescending}(r_1, \ldots, r_k)/(\sum_k r_k)$.
4. For $perm = 1, \ldots, P$, repeat:
   (a) Permute the features of each sample of X to form a permuted dataset $X^{perm} = \text{permuteFeaturesPerSample}(X)$.
   (b) Estimate the sources S^{perm} and the mixing matrix A^{perm} from the permuted data X^{perm} via FastICA or another related method.
   (c) For each permuted component S_k^{perm}, we estimate its relevance by computing $r_k^{perm} = \sum_i (A_{ki}^{perm})^2$.
   (d) Order the permuted components' relevance r_k^{perm} from the highest to the lowest value and scale them to sum to 1, i.e. $r^{perm} = \text{orderDescending}(r_1^{perm}, \ldots, r_k^{perm})/(\sum_k r_k^{perm})$.
5. Apply permutation testing for each value of r with respect to the values of r^{perm} and compute the corresponding p-values, i.e. $p_k = |\{r_k^{perm} | r_k \geq r_k^{perm}\}|/P$.
6. The number of components is the number of p_k's for which $p_k < 0.05$. The p-values indicate how many components have higher relevance than the components from the permuted data.

In our application, we first select the number of total components k_d for each dataset via the proposed procedure. Then we fit a ShIndICA model, for which we select the first k_d components according to their relevance. Thus, the performed dimensionality reduction step happens after training.

[2]The dataset is available at `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE67023`

[3]The dataset can be found at `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE27219`

[4]Dataset link `https://imodulondb.org/dataset.html?organism=b_subtilis&dataset=modulome`

3.2.2 Graphical Lasso

Graphical lasso (glasso) is a maximum likelihood estimator for inferring graph structure in a high-dimensional setting [Friedman et al., 2007]. This method uses l_1 regularization to estimate the precision matrix (or inverse covariance) of a set of random variables from which a graph structure can be determined. The optimization problem that glasso solves can be formalized as follows

$$\min_{\Theta \succ 0} - \log \det(\Theta) + \operatorname{tr}(\hat{\Sigma}\Theta) + \lambda \|\Theta\|_1, \tag{7}$$

where $\hat{\Sigma}$ is the empirical covariance or correlation matrix and $\Theta := \Sigma^{-1}$ denotes the precision matrix. In our setting, the input for the glasso is the Pearson's correlation matrix of the gene representations retrieved with ICA at the preceding step. We can read graph structure from the estimated matrix $\hat{\Theta}$ as follows: if the ij entry of $\hat{\Theta}$ is not 0 (i.e. $\hat{\Theta}_{ij} \neq 0$) there is an edge between the genes i and j, i.e. the genes might be co-regulated. We used the `huge`[5] R package for the implementation of the graphical lasso.

3.2.3 Extended EBIC

There are various criteria for model selection and hyperparameter tuning of glasso models. Chen and Chen [2008] propose an information criterion for Gaussian graphical models called extended BIC (EBIC) that takes the form

$$- \log \det(\Theta(E)) + \operatorname{tr}(\hat{\Sigma}\Theta(E)) + |E| \log n + 4|E|\gamma \log p, \tag{8}$$

where E is the edge set of a candidate graph and $\gamma \in [0, 1]$. Models that yield low EBIC scores are preferred. Note that positive values for γ lead to sparser graphs. Foygel and Drton [2010] suggest that $\gamma = 0.5$ is a good choice when no prior knowledge is available. In our experiments, we select the λ that minimizes the EBIC score with $\gamma = 0.5$.

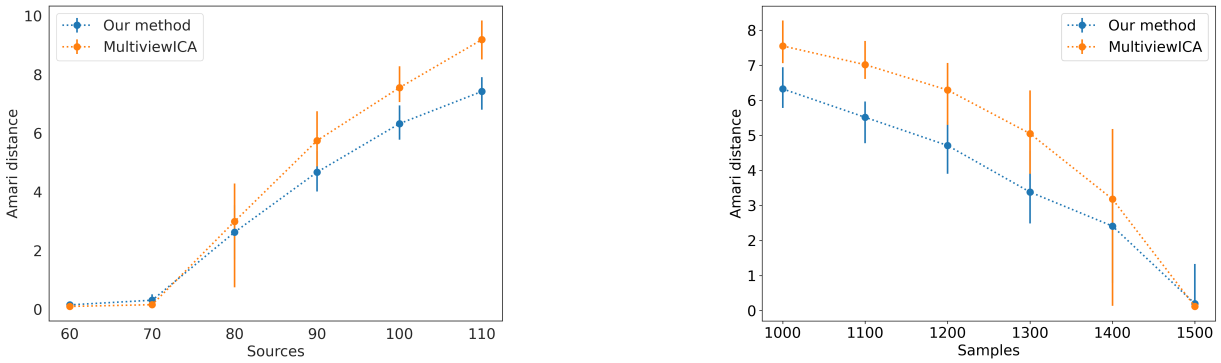
3.2.4 Method

All steps described above are summarized in the following pseudo-code.

Algorithm 1 Algorithmic description of the downstream task for $D = 2$.

1: **Input:**
   $X_1, \in \mathbb{R}^{n_1 \times p}, X_2 \in \mathbb{R}^{n_2 \times p}$ is a data matrix with n_1 and n_2 samples and p genes
   Λ is a set of regularization parameters
   γ EBIC selection parameter (8)
2: Perform a data integration method to obtain $S_1, \in \mathbb{R}^{k_1 \times p}, S_2 \in \mathbb{R}^{k_2 \times p}$
3: Concatenate $S = (S_1, S_2)^\top \in \mathbb{R}^{k_1 + k_2 \times p}$
4: Compute the Pearson correlation matrix $\hat{\Sigma} \in \mathbb{R}^{p \times p}$ of S.
5: Estimate the precision matrices $\{\hat{\Theta}^\lambda\}_{\lambda \in \Lambda}$ which solves 7 for each λ from the set Λ
6: Select the final $\hat{\Theta}^{out} \in \{\hat{\Theta}^\lambda\}_{\lambda \in \Lambda}$ according to EBIC(γ) (see 8)
7: **Output:**
   the selected $\hat{\Theta}^{out}$

[5] See `https://CRAN.R-project.org/package=huge`.



(a) (b)

Figure 2: Comparison of MultiViewICA and ShIndICA on a two-view shared response model setting. In Figure 2a, we fix the sample size and measure the Amari distance for sources $60, 70, \ldots 110$. In Figure 2b the number of sources is set to 100 and we conduct the experiments for different sample sizes (x-axis). It seems that ShIndICA outperforms MultiViewICA in both scenarios.

4 SYNTHETIC EXPERIMENTS

4.1 AMARI DISTANCE

The Amari distance [Amari et al., 1995] between two invertible matrices $A, B \in \mathbb{R}^{n \times n}$ is defined by

$$\text{amari}(A, B) := \sum_{i=1}^{n} \Big(\sum_{j=1}^{n} \frac{|c_{ij}|}{\max_k |c_{ik}|} - 1 \Big) + \sum_{j=1}^{n} \Big(\sum_{i=1}^{n} \frac{|c_{ij}|}{\max_k |c_{kj}|} - 1 \Big), \qquad C := A^{-1}B.$$

4.2 ADDITIONAL EXPERIMENTS ON SYNTHETIC DATA

Objective function motivation. In the following experiment, we compare MultiViewICA and ShIndICA when the observed data is high-dimensional on a two-view shared response model application, i.e., no individual sources. The experimental setup allows comparing standard MLE (MultiViewICA) and MLE after whitening (ShIndICA). Figure 2a reaches the two methods for fixed sample size 1000. In Figure 2b, we set the number of sources to 100 and vary the sample size. For all experiments, the noise standard deviation is 0.01. It seems that ShIndICA performs better in the case of insufficient data. This could be empirical evidence that the trace has stronger regularization properties than the MMSE term in the MultiViewICA objective.

Noisy high-dimensional views. First, we investigate the effect of noise on the Amari distance in the two-view experiment. We consider three cases when the noise's standard variation is $\sigma = 0.1, 0.5, 1$. The results are depicted in Figure 3. In the first two cases, the results are close to the ones discussed in the main paper. As expected, by adding noise with high variance ($\sigma = 1$) ShIndICA does not converge and affects the quality of the estimated mixing matrices measured

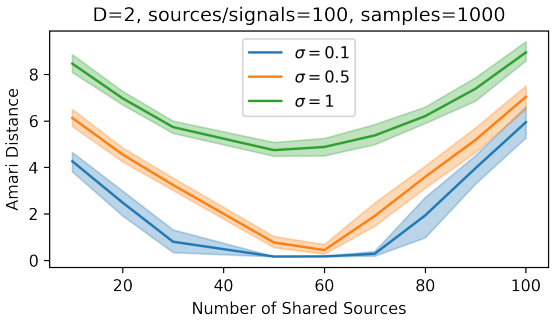


Figure 3: We have the two view case with a number of total sources and observed signals 100 and a number of samples 1000. We consider three cases of noise standard deviation: $\sigma = 0.1, 0.5, 1$. As soon as enough shared sources are present (around 60) ShIndICA reaches its lowest Amari distance value (the lower, the better) in all cases. In the first two cases ($\sigma = 0.1$ or 0.5) the Amari distance gets closer to 0 when the shared sources are 60. The error bars correspond to 95% confidence intervals based on 50 independent runs of the experiment.

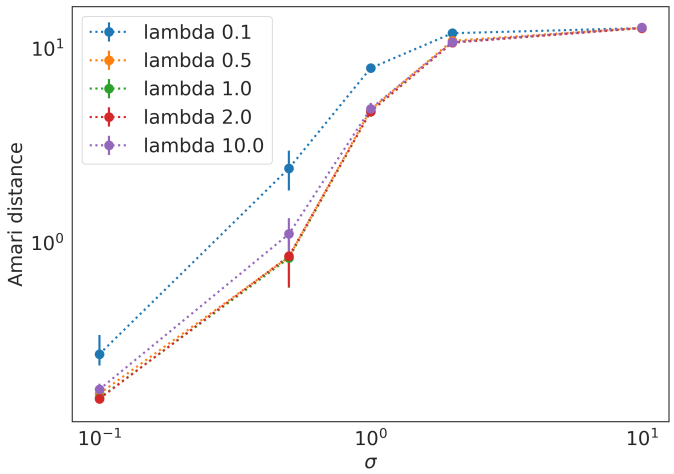


Figure 4: Choice of Hyperparameter λ. The data comes from a two-view model with 50 shared and 50 individual sources per view. The x-axis represents the noise standard deviation and the y-axis the Amari distance.

Study	Application	Observed Signals	Latent Sources	Views
Salman et al. [2019]	Identifying biomarkers	fMRI data	brain functional networks	multiple subjects
[Durieux and Wilderjans, 2019]	Mental disorders detection	fMRI data	brain functional networks	multiple subjects
[Long et al., 2020]	subgroup detection	fMRI data	brain functional networks	multiple subjects
[Huster et al., 2015]	Denoising	EEG data	brain activity patterns	multiple subjects
[Congedo et al., 2010]	Diagnosis and assessment of abnormal brain functioning	EEG data	eyes-closed resting EEG patterns	multiple subjects
[Sompairac et al., 2019]	extensive overview	tumoral omics data	gene/protein profiles	heterogeneous omics data
[Avila Cobos et al., 2018]	cell type decomposition	tissue/tumor samples	cell type-specific expressions	tissue/tumor samples
[Fraunhoffer et al., 2022]	prognostic prediction	transcriptomic profiles from PDAC epithelial and microenvironment cells	gene profile	three types of transcriptome data

Table 1: List of recent studies that use ICA as a common data analysis tool. We also provide the application, used data modalities latent sources and views interpretation.

with the Amari distance. The procedure is repeated 50 times, and the error bars are the 95% confidence intervals based on the independent runs.

Choice of λ This experiment used data from 2 views with 50 individual and 50 shared sources with varying noise standard deviation $\sigma \in \{0.1, 0.5, 1, 2, 10\}$ (x-axis). Each of the lines in Figure 4 correspond to a fixed hyperparameter $\lambda \in \{0.1, 0.5, 1, 2, 10\}$. It can be deduced that for this particular experiment for $\lambda \geq 0.5$ there is no significant difference in the model performance.

5 MODEL JUSTIFICATION

Multi-view ICA importance in the scientific community. As mentioned in our introduction, we would like to point out that ICA has proven to be a successful approach for analyzing biomedical data over the years since it solves blind source separation problems common in neuroscience and biomedicine, as stated in the main paper. Furthermore, many biomedical applications can be addressed as multi-view problems due to multiple subjects in a study (e.g., fMRI, EEG data) or data coming from different modalities (e.g., omics data). This led to the development of multi-view methods. Most of those approaches focus on shared response model setting (only shared sources), e.g., Group ICA, ShICA, MultiviewICA, IVA methods, and their corresponding variations. We list some recent scientific applications where multi-view ICA models were used in Table 1. We also interpreted the used views and latent and observed signals.

The shared response models are restrictive. There is a growing interest in examining individual variability rather than shared signals in the areas mentioned above of applications [Dubois and Adolphs, 2016] , such as [Seghier and Price, 2018, Bartolomeo et al., 2017, Long et al., 2020]. For instance, one can be interested in the effect of individual brain patterns on brain activity to develop more robust biomarkers. Another application where shared response models (GroupICA, MultiviewICA, IVA, etc.) would not be a sensible choice is data integration of omics data. This is an important research direction in computational biology, where we are interested in preserving the shared biological signal between datasets

(views) and individual ones, as illustrated in our example. Existing approaches for the tasks mentioned above consist of two steps: applying ICA/IVA on the data followed by statistical analysis (as in [Long et al., 2020]) to separate the individual from the shared sources (or vice versa). Thus, we believe ShIndICA is a valuable addition to this set of tools.

Linearity assumption in the biomedical domain. The nature of the data in the targeted domains can explain the linear assumption. More precisely, if we consider the examples from above: the linear mixing of the components in the fMRI data context has been justified by various studies, e.g. McKeown and Sejnowski [1998], and in the other applications, the linear assumption can be achieved after data transformation, e.g. log-transforming the transcriptome data. Moreover, the linearity assumption is valid in many real-life applications in the biomedical domain, where we often have a high-dimensional setting (gene activity, experimental measurements, etc.) with a low number of observed samples (participants, experiments). Moreover, in the low-data regime, if we know too little about the underlying problem, the linear approach is often a better option than eventually overparametrization it with a deep learning model. Even though a non-linear multiview version will be a valuable addition to the current active research on non-linear ICA, e.g. [Hyvärinen and Morioka, 2016, 2017, Monti et al., 2020], the identifiability justification of the proposed methods has assumptions that are hard to satisfy in real-life data scenarios (e.g. the assumption of Variability [Hyvärinen et al., 2019]. In our linear version, we assure identifiability without any requirements on how distinct the views should be.

6 MODEL ASSUMPTIONS

To prove the identifiability of the stated model, we require that four assumptions should be satisfied:

1. The mixing matrices have full-column rank. This implies that we require that the sources have a minimal representation, i.e. the number of latent sources is minimal, which is a realistic assumption.
2. The second assumption is additive noise on the sources. It can be interpreted as a measurement error on the device with variance $\sigma^2 A_d A_d^\top$. We choose this setting compared to the $A_d s_d + \epsilon_d$ because, in our case, we get a likelihood in a closed form which is not available in the latter representation. Richard et al. [2020, 2021] make a similar assumption for the shared response model setting.
3. The sources are mutually independent and non-Gaussian. This is a standard ICA assumption [Comon, 1994]. Gaussian random variables, called "white" noise, represent noise variables, which besides location and scale, do not carry real information. Thus, if all sources are Gaussian, either they cannot be identified (see, for example, Proposition 3 [Richard et al., 2020]) or additional assumptions on the variance structure need to be made to assure identifiability [Richard et al., 2021]. The non-Gaussian random variables carry meaning and are identifiable. This is not a restrictive assumption since the sources in real-life scenarios are often non-Gaussian: fMRI, EEG, and omics data. The fixed mean, and variance are also assumptions often adopted in ICA (e.g. [Richard et al., 2021, Hyvärinen and Oja, 2000]).
4. The measurement error is independent of the latent signal. This is a common assumption in measurement error models known as classical errors. It is a realistic assumption since we usually do not expect the measurement error to influence the true signal and vice versa Richard et al. [2020, 2021], Gresele et al. [2020].

References

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

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.

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

Jiahua Chen and Zehua Chen. Extended Bayesian Information Criteria for Model Selection with Large Model Spaces. *Biometrika*, 95(3):759–771, 2008.

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.

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

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.

Rina Foygel and Mathias Drton. Extended Bayesian Information Criteria for Gaussian Graphical Models. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

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.

Luigi Gresele, Paul K Rubenstein, Arash Mehrjou, Francesco Locatello, and Bernhard Schölkopf. The Incomplete Rosetta Stone Problem: Identifiability Results for Multi-View Nonlinear ICA. In *Uncertainty in Artificial Intelligence*, pages 217–227. PMLR, 2020.

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 Hiroshi Morioka. Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA. *Advances in Neural Information Processing Systems*, 29, 2016.

Aapo Hyvärinen and Hiroshi Morioka. Nonlinear ICA of Temporally Dependent Stationary Sources. In *Artificial Intelligence and Statistics*, pages 460–469. PMLR, 2017.

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.

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.

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.

Ricardo Pio Monti, Kun Zhang, and Aapo Hyvärinen. Causal Discovery with General Non-Linear Relationships using Non-Linear ICA. In *Uncertainty in Artificial Intelligence*, pages 186–195. PMLR, 2020.

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.

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.

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.

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*, 2021.

Mohamed L Seghier and Cathy J Price. Interpreting and utilising intersubject variability in brain function. *Trends in Cognitive Sciences*, 22(6):517–530, 2018.

Nicolas Sompairac, Petr V Nazarov, Urszula Czerwinska, Laura Cantini, Anne Biton, Askhat Molkenov, Zhaxybay Zhumadilov, Emmanuel Barillot, Francois Radvanyi, Alexander Gorban, et al. Independent Component Analysis for Unraveling the Complexity of Cancer Omics Datasets. *International Journal of molecular sciences*, 20(18):4414, 2019.