MULTI-VIEW INDEPENDENT COMPONENT ANALYSIS FOR OMICS DATA INTEGRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

With the increasing number of omics data, there is a great need to incorporate these datasets together to create a better and more robust understanding of the underlying biological processes. We transform this problem into a noisy multi-view independent component analysis (ICA) task by assuming that each observed dataset (view) is a linear mixture of independent latent biological processes. Furthermore, we assume that each view contains a mixture of shared and individual sources. To computationally estimate the sources, we optimize a constrained form of the joint log-likelihood of the observed data among all views. Finally, we apply the proposed model in a challenging real-life application, where the estimated shared sources from two large transcriptome datasets (observed data) provided by two different labs (two different views) lead to a more plausible representation of the underlying graph structure than existing baselines.

1 INTRODUCTION

With the fast advancement of technology, growing medical and biological data from omics (e.g., genomics, transcriptomics, epigenomics, microbiomics) can be collected and combined to provide valuable insights for disease development, to improve the performance of downstream tasks such as gene regulation discovery, cancer prediction, etc. However, analyzing these datasets can be very challenging and inefficient without considering measurement errors and batch effects (non-biological noise in the data) in the data, and often missing ground truth knowledge, etc.

This work focuses on a specific task: unsupervised data integration from the same omics modality. Our approach is based on two key assumptions. First, we adopt the paradigm of a linear mixture of independent signals in molecular data analysis such as independent regulatory pathways in the data (Sompairac et al., 2019; Avila Cobos et al., 2018; Fraunhoffer et al., 2022). Furthermore, we assume that the different datasets contain shared and dataset (view)-specific information. For example, the shared information could represent population-specific and the individual sources can correspond to technical artefacts but also signals of medical/biological interest. In transcriptomics, the shared sources could be the housekeeping genes activity and the individual ones - experiment-specific gene activity due to knock-out experiments, stress conditions etc.

Independent component analysis (ICA) is often used for modeling omics data (Zheng et al., 2008; Nazarov et al., 2019; Zhou & Altman, 2018; Tan et al., 2020; Sastry et al., 2021; 2019; Urzúa-Traslaviña et al., 2021; Rusan et al., 2020; Cary et al., 2020; Dubois et al., 2019; Aynaud et al., 2020) and satisfies our first assumption. Its goal is to separate independent latent sources from mixed observed signals and, thus, uncover essential data structures in various data types. In the multiview scenarios, ICA based methods has been developed mainly for fMRI data analysis (MultiViewICA, ShICA-ML,(Richard et al., 2020; 2021), Group ICA (Calhoun et al., 2001), independent vector analysis (IVA) methods (Lee et al., 2008; Anderson et al., 2011; 2014; Engberg et al., 2016; Vía et al., 2011)), where all views contain only shared sources and no view-specific ones. We propose a novel multiview extension of ICA that models the different views (datasets) to have both shared and view-specific sources in the presence of noise. Our method provides an identifiable model with a known closed-form likelihood for estimating the parameters. The resulting framework is applied to a transcriptome data integration task for the bacteria *B.subtilis* with a well-studied gene regulatory network and boosts the data-driven gene regulation discovery compared to other omics data integration approaches.

2 Problem Formalization

Consider the following D-view multivariate linear model

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

where we assume that for $d = 1, \ldots, D$: 1. $x_d \in \mathbb{R}^{k_d}$ is a random vector of *signals* with $\mathbb{E}[x_d] = 0$; 2. $\tilde{s}_d = (s_0^\top, s_d^\top)^\top$ are latent random *sources* with and $s_0 \in \mathbb{R}^c$ and $s_d \in \mathbb{R}^{k_d-c}$ being the shared and individual sources and $\mathbb{E}[\tilde{s}_d] = 0$ and $\mathrm{Var}[\tilde{s}_d] = \mathbb{I}_{k_d}$; 3. $A_d \in \mathbb{R}^{k_d \times k_d}$ is a mixing matrix with full column rank, A_{d0} and A_{d1} are the columns corresponding to the shared and individual sources; 4. $\epsilon_d \sim \mathcal{N}(0, \sigma^2\mathbb{I}_{k_d})$ is Gaussian noise, or measurement error, on the sources (similar to (Richard et al., 2020; 2021)). Additionally, we assume that the variables $s_{01}, \ldots, s_{0c}, s_{11}, \ldots, s_{1(k_1-c)}, \ldots, s_{D1}, \ldots, s_{D(k_D-c)}, \epsilon_{11}, \ldots, \epsilon_{1k_1}, \ldots, \epsilon_{D1}, \ldots, \epsilon_{Dk_D}$ are mutually independent. Thus, we require that the noise variables do not influence the latent signal and vice versa. The imposed model assumptions guarantee the *identifiability* of the proposed data generation model as defined by Comon (1994), i.e. we can reconstruct the mixing matrices A_d up to scale and permutation from the observed data.

In this work, we aim to estimate the mixing matrices A_d from observed random matrices of $X_d \in \mathbb{R}^{k_d \times N}, d = 1, \ldots, D$. Translated to our data integration task k_d refers to the number of observed experimental outcomes in dataset d and N to the measured entities of interest, such as gene expressions in transcriptomics. Thus, from k_d observed samples in dataset d, we estimate k_d latent sources that represent independent biological pathways with respect to the data. In addition, the noise variable represents non-biological noise. The identifiability of the proposed model guarantees that the reconstructed (noisy) latent representations *are the true ones up to scaling and permutation under the paradigm of linear mixture of signals.*

3 Joint Data Likelihood

Here, we derive the joint log-likelihood of the observed views, which we use for estimating the mixing matrices. We treat the D data matrices $X_d \in \mathbb{R}^{k_d \times N}, d = 1, \ldots, D$ as a collection of N independent observations of the k_d experimental outcomes. Thus, we do not explicitly model the dependence relationship between the N entities of interest. However, as we see empirically in the next section, this dependency is still preserved in the constructed latent sources.

Let $z_d := W_d x_d = \tilde{s}_d + \epsilon_d$, and $z_d^{(1)} := s_0 + \epsilon_{d0} \in \mathbb{R}^c$ and $z_d^{(2)} := s_d + \epsilon_{d1} \in \mathbb{R}^{k_d-c}$, i.e. $z_d = (z_d^{(1)\top}, z_d^{(2)\top})^\top$. Furthermore, let $p_{Z_d^{(2)}}$ be the probability distribution of $z_d^{(2)}$ and $|W_d| = |\det W_d|$. Then the data log-likelihood of 1 is given by

$$\begin{aligned}\mathcal{L}(W_1, \ldots, W_D) = &\sum_{i=1}^{N} \log f(\bar{s}_0^i) + \sum_{i=1}^{N}\sum_{d=1}^{D} \log p_{Z_d^{(2)}}(z_d^{(2)i}) + N\sum_{d=1}^{D} \log|W_d| \\ &- \frac{1}{2\sigma^2}\Big(\sum_{d=1}^{D} \mathrm{trace}(Z_d^{(1)} Z_d^{(1)\top}) - \frac{1}{D}\sum_{d=1}^{D}\sum_{l=1}^{D} \mathrm{trace}(Z_d^{(1)} Z_l^{(1)\top})\Big) + C\end{aligned} \tag{2}$$

where $f(\bar{s}_0) = \int \exp\Big(-\dfrac{D\|s_0 - \bar{s}_0\|^2}{2\sigma^2}\Big)p_{S_0}(s_0)ds_0$, $Z_d^{(1)} \in \mathbb{R}^{c \times N}$ for $d = 1, \ldots, D$ is a data matrix and $\bar{s}_0^i = \sum_{d=1}^{D} z_d^{(1)i}/D$. We further simplify the loss function 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. That consists of centering and linearly transforming the random variables' realizations x_d such that the resulting variable $\tilde{x}_d = K_d x_d$ has unit variance, $\mathbb{E}[\tilde{x}_d\tilde{x}_d^\top] = \mathbb{I}_{k_d}$, where K_d is the whitening matrix. Thus, from the last equation we get that $\mathbb{I}_{k_d} = \mathbb{E}[\tilde{x}_d\tilde{x}_d^\top] = (1+\sigma^2)K_d A_d A_d^\top K_d^\top$. It follows that the matrix $(1+\sigma^2)^{\frac{1}{2}}K_d A_d$ is orthogonal, which we estimate by the matrix W_d. After training we set $\hat{A}_d = K_d^{-1}W_d$ which differs from the true one by $(1+\sigma^2)^{\frac{1}{2}}$. Due to the orthogonal constraints the objective function becomes

$$\mathcal{L}(W_1,\dots,W_D) \propto \sum_{i=1}^{N} \log f_\sigma(\bar{s}_0^i) + \sum_{i=1}^{N}\sum_{d=1}^{D} \log p_{Z_d^{(2)}}(z_d^{(2)i}) + \frac{1+\sigma^2}{2D\sigma^2}\sum_{d=1}^{D}\sum_{l=1}^{D} \operatorname{trace}(Z_d^{(1)} Z_l^{(1)\top}) \tag{3}$$

where here $f_\sigma(\bar{s}_0) = \int \exp\Big(-\frac{D\|s_0 - (1+\sigma^2)^{\frac{1}{2}}\bar{s}_0\|^2}{2\sigma^2}\Big)p_{S_0}(s_0)ds_0$. All proofs can be found in Appendix A. Note that in our optimization procedure, both $f_\sigma(\bar{s}_0)$ and $p_{Z_d^{(2)}}$, we approximate by the negative of a nonlinear function $g(s)$, e.g. $g(s) = \log\cosh(s)$ for super-Gaussian or $g(s) = -e^{-s^2/2}$ for sub-Gaussian sources. Moreover, since we do not estimate σ^2 we treat it as a Lagrange multiplier via the relation $\lambda = \frac{1+\sigma^2}{\sigma^2}$.

4 Data Fusion of Transcriptome Data

Model Implementation. Before running any ICA-based method, we whiten every single view by performing PCA to speed up computation. 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 optimization method applied for training is L-BFGS we initialize the unmixing matrices with canonical correlation analysis (CCA) (Hotelling, 1936).

Motivational Example: Simulated Data. Here we exemplify the advantages of our method compared to other group ICA models and a naive ICA method (called Infomax), where we run Infomax-ICA on each view separately. We simulated the data using the Laplace distribution $\exp(-\frac{1}{2}|x|)$, and the mixing matrices are sampled with normally distributed entries with mean 1 and 0.1 standard deviation. The realizations of the observed views are obtained according to the proposed model with $\sigma = 0.1$. 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. In Figure 1 we vary the number of shared sources from 10 to 100 for a total number of sources 100 and sample size 1000. We can see that as soon as the ratio of shared sources to individual sources gets around 1:1 we can recover almost. This is not the case for the baseline methods.

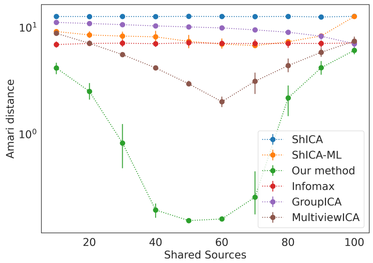


Figure 1: We compare the performance of our method, ShICA, Infomax, GroupICA, MultiViewICA and ShICA-ML Amari distance (the lower the better) for different number of shared sources (x-axis). The error bars correspond to 95% confidence intervals based on 50 independent runs of the experiment. The datasets come from two different views with total number of sources 100 and sample size 1000. We vary the number of shared sources from 10 to 100.

Data Integration Task. Based on transcriptome datasets, scientists try to infer gene-gene interactions in the genome. The goal of the data integration task is to ”denoise” the datasets, such that the transformed data can be used as samples for a graph inference algorithm. More precisely, in this application, we apply graphical lasso (glasso) (Friedman et al., 2007) on the combined datasets to estimate an undirected graph with nodes referring to the genes and with edges connecting genes with a common regulator. We compare our method to IVA-L-SOS, PLS (a CCA-based approach that extracts between-views correlated components and view-specific ones, provided by the `OmicsPLS` R package Bouhaddani et al. (2018)) and naive ICA approach (Infomax as in the previous example).

Experiment on *B. subtilis*. In this example, we consider the bacterium *B. subtilis*, for which a very rich collection of the discovered gene-gene interactions are publicly available, which we use as our ground truth model. For this data integration task we use two publicly available datasets (Arrieta-Ortiz et al., 2015; Nicolas et al., 2012). Each of the datasets contain gene expression levels of about 4000 genes measured across more than 250 experimental outcomes. [1] As in most real-life

[1] For detailed description of the datasets and procedure we refer to Appendix B.

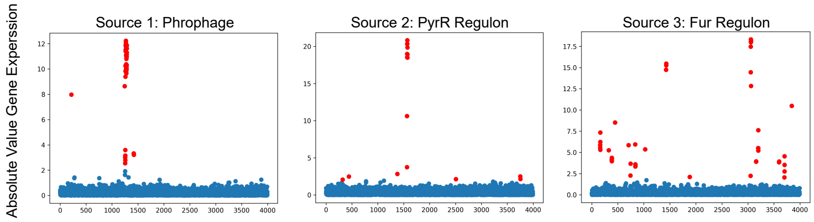


Figure 2: Absolute Value of Gene Expressions from the three ”strongest” shared latent sources. The red markers are outliers and can be related to functional groups (see titles).

applications, the number of latent sources per dataset is unknown. We treat it as a hyperparameter for each model, i.e., we perform grid search on $\{50, 60, 70, \ldots, 200\}^2$, for the total number of sources for both datasets. Note that for IVA-L-SOS the number of sources in both datasets should coincide. The number of shared sources for our method and PLS varies between $10, 20, 30$, and 40. We fit 30 graphical lasso models for different penalization parameters on the estimated components. We select the top 10 models by employing a statistical goodness-of-fit measure, called EBIC, for each combination of hyperparameters. Then, the hyperparameter setting is selected, yielding the best true positive/false positive ratio curves (as the ones shown in Figure 3). The resulting hyperparameter settings are IVA-L-SOS (130 latent sources), Our Method (50 for dataset 1, 60 for dataset 2, 40 shared sources), PLS (180 for dataset 1, 80 for dataset 2, 10 shared sources), and Infomax (200 for dataset 1, 50 for dataset 2).

True Positives vs False Positives. The below-described evaluation is used for our hyperparameter selection. The output graph from the graphical lasso for each preprocessing method is compared to the ground truth one. For each estimated graph, we order the edges according to their strength. Then we count the true positive and false positive edges in the first $100, 200, \ldots$ edges. Then for each method separately, we select the hyperparameter combination for which the graphical lasso has the best true positive/false positive ratio curves. The results are depicted in Figure 3, where the best models for each method are compared. We can conclude that our model boosts the graphical lasso's performance compared to the others. We also run the graphical lasso on the pooled data without any pre-processing. Surprisingly, the method outputs an empty graph, i.e., the goodness-of-fit measure we use evaluates the empty graph as the best model describing the data.

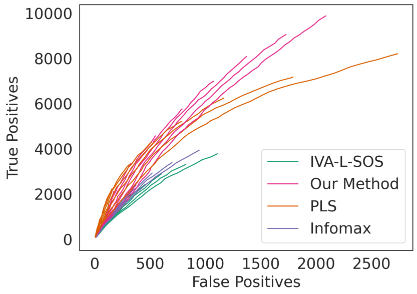


Figure 3: We compare the top ten models with our model, PLS, Infomax, and IVA-L-SOS. We order the edges from the selected networks according to their strength. We count the true positives (y-axis) and possibly false positive edges in the first 100, 200,... edges (x-axis). Our method and PLS outperform the other two methods, and for our method, the true positive/false positive rate increases faster than for PLS.

Qualitative Interpretation of the shared sources. Figure 2 visualizes the ”gene expression levels” of the shared sources with the highest correlation. Each marker represents one gene, and the red markers annotate the outliers. We can provide a meaningful interpretation of all outliers. All red markers from the first source belong to prophage genes, and the ones from the second and the third latent sources are regulated by the *pyrR* and *fur* regulators, respectively.

5 Discussion

We suggested a novel strategy for combining omics data by assuming that the observed data follows a noisy multiview linear ICA model with both shared and view-specific latent sources. We adopted a maximum likelihood strategy for estimating the unmixing matrices by maximizing the joint log-likelihood of the observed views and showed empirically that our procedure improves the performance of a graph inference algorithm. In future work, we would like to investigate how to choose the number of shared sources and apply our method to other omics data.

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.

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.

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.

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

Said el Bouhaddani, Hae-Won Uh, Geurt Jongbloed, Caroline Hayward, Lucija Klarić, Szymon M Kiełbasa, and Jeanine Houwing-Duistermaat. Integrating omics datasets with the omicspls package. *BMC bioinformatics*, 19(1):1–9, 2018.

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.

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.

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)*, pp. 1–4. IEEE, 2016.

Rina Foygel and Mathias Drton. Extended bayesian information criteria for gaussian graphical models. *arXiv preprint arXiv:1011.6640*, 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. ISSN 1465-4644. doi: 10.1093/biostatistics/kxm045. URL `https://doi.org/10.1093/biostatistics/kxm045`.

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*, pp. 217–227. PMLR, 2020.

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

Maxim (https://math.stackexchange.com/users/491644/maxim). Multivariate transformation formula correct? Mathematics Stack Exchange, 2019. URL `https://math.stackexchange.com/q/3325459`. URL:https://math.stackexchange.com/q/3325459 (version: 2019-08-17).

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*, pp. 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*, pp. 859–868. PMLR, 2019.

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*, pp. 9154–9164, 2019.

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*, pp. 186–195. PMLR, 2020.

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.

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.

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, 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.

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.

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.

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*, pp. 1–6. IEEE, 2011.

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.

A OPTIMIZATION

Lemma A.1. *Let $W \in \mathbb{R}^{c\times k}$ such that $WW^\top = \mathbb{I}_c$ and $x^1, \ldots, x^N \in \mathbb{R}^k$ such that for every $j = 1, \ldots, k$, 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 $j = 1, \ldots, c$, 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 $WW^\top = \mathbb{I}_c$. □

Under the generative model assumptions and optimization constraints stated in 2 it holds

$$\mathcal{L}(W_1, \ldots, W_D) = \sum_{i=1}^N \log f(\bar{s}_0^i) + \sum_{i=1}^N \sum_{d=1}^D \log p_{Z_d^{(2)}}(z_d^{(2)i}) + N\sum_{d=1}^D \log |W_d| \tag{4}$$

$$- \frac{1}{2\sigma^2}\Big(\sum_{d=1}^D \operatorname{trace}(Z_d^{(1)} Z_d^{(1)\top}) - \frac{1}{D}\sum_{d=1}^D \sum_{l=1}^D \operatorname{trace}(Z_d^{(1)} Z_l^{(1)\top})\Big) \tag{5}$$

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

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

Furthermore, let $z_d := W_d x_d = \tilde{s}_d + \epsilon_d$, and $z_d^{(1)} := s_0 + \epsilon_{d0} \in \mathbb{R}^c$ and $z_d^{(2)} := s_d + \epsilon_{d1} \in \mathbb{R}^{k_d - c}$, i.e. $z_d = (z_d^{(1)}, z_d^{(2)})^\top$. Let $p_{\mathbf{X}}$ be the joint distribution of $x_1, \ldots, x_D$, $p_{\mathbf{Z}}$ the joint distribution of $z_1, \ldots, z_D$, $p_{\mathbf{Z}^{(1)}}$ the joint distribution of $z_1^{(1)}, \ldots, z_D^{(1)}$, $p_{\mathbf{Z}^{(2)}}$ the joint distribution of $z_1^{(2)}, \ldots, z_D^{(2)}$ and $p_{Z_d^{(2)}}$ the probability distribution of $z_d^{(2)}$.

Note that the model in 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 $x_1, \ldots, x_D$ we get

$$\begin{aligned} p_{\mathbf{X}}(\mathbf{x}) &= p_{\mathbf{Z}}(\mathbf{z})|\mathbf{W}| \\ &= p_{\mathbf{z}}(\mathbf{z}) \prod_{d=1}^D |W_d| \\ &= p_{\mathbf{Z}^{(1)}}(z_1^{(1)}, \ldots, z_D^{(1)}) p_{\mathbf{Z}^{(2)}}(z_1^{(2)}, \ldots, z_D^{(2)}) \prod_{d=1}^D |W_d| \\ &= p_{\mathbf{Z}^{(1)}}(z_1^{(1)}, \ldots, z_D^{(1)}) \prod_{d=1}^D p_{Z_d^{(2)}}(z_d^{(2)}) \prod_{d=1}^D |W_d|. \end{aligned}$$

1. Second equation: $\mathbf{W}$ is a block diagonal matrix and for all $d = 1, \ldots, D$, and $W_d \in \mathbb{R}^{k_d \times k_d}$.

2. Third equation: $z_1^{(1)}, \dots, z_D^{(1)} \perp\!\!\!\perp z_1^{(2)}, \dots, z_D^{(2)}$.
3. Fourth equation follows from the fact that $z_1^{(2)}, \dots, z_D^{(2)}$ 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

$$\begin{aligned}
p_{\mathbf{Z}^{(1)}}(z_1^{(1)}, \dots, z_D^{(1)}) &= \int p_{\mathbf{Z}^{(1)}|S_0}(z_1^{(1)}, \dots, z_D^{(1)}|s_0) p_{S_0}(s_0) ds_0 \\
&= \int \Big(\prod_{d=1}^{D} \mathcal{N}(z_d^{(1)}; s_0, \sigma^2 \mathbb{I}_c) \Big) p_{S_0}(s_0) ds_0 \\
&\propto \int \exp\Big(- \sum_{d=1}^{D} \frac{\|z_d^{(1)} - s_0\|^2}{2\sigma^2} \Big) p_{S_0}(s_0) ds_0 \\
&= \int \exp\Big(- \frac{D\|s_0 - \bar{s}_0\|^2 + \sum_{d=1}^{D} \|z_d^{(1)} - \bar{s}_0\|^2}{2\sigma^2} \Big) p_{S_0}(s_0) ds_0 \\
&= \exp\Big(- \frac{\sum_{d=1}^{D} \|z_d^{(1)} - \bar{s}_0\|^2}{2\sigma^2} \Big) \int \exp\Big(- \frac{D\|s_0 - \bar{s}_0\|^2}{2\sigma^2} \Big) p_{S_0}(s_0) ds_0
\end{aligned}$$

where $\bar{s}_0 = \frac{1}{D} \sum_{d=1}^{D} z_d^{(1)}$.

- For the second and third equation recall that $z_d^{(1)} = s_0 + \epsilon_{d0} \in \mathbb{R}^c$, where $\epsilon_{d0} \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_c)$ and $s_0 \perp\!\!\!\perp \epsilon_{d0}$. This means that $z_d^{(1)}|s_0 \sim \mathcal{N}(s_0, \sigma^2 \mathbb{I}_c)$. From the following equations follow

$$\begin{aligned}
p_{\mathbf{Z}^{(1)}|S_0}(z_1^{(1)}, \dots, z_D^{(1)}|s_0) &= \prod_{d=1}^{D} p_{Z_d^{(1)}|s_0}(z_d^{(1)}|S_0) \\
&= \prod_{d=1}^{D} \mathcal{N}(z_d^{(1)}; s_0, \sigma^2 \mathbb{I}_c)
\end{aligned}$$

- The fourth equation results from

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

We define $f(\bar{s}_0) = \int \exp\Big(- \frac{D\|s_0 - \bar{s}_0\|^2}{2\sigma^2} \Big) p_{S_0}(s_0) ds_0$.

Note that

$$\|z_d^{(1)} - \bar{s}_0\|^2 = \|z_d^{(1)}\|^2 - \frac{2}{D} \sum_{l=1}^{D} \langle z_d^{(1)}, z_l^{(1)} \rangle + \frac{1}{D^2} \sum_{l=1}^{D} \sum_{r=1}^{D} \langle z_r^{(1)}, z_l^{(1)} \rangle.$$

Thus, it follows that

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

Collecting all terms together we get

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

The data log-likelihood can be expressed as

$$\begin{aligned}\sum_{i=1}^{N}\log p_{\mathbf{X}}(x_1^i,\dots,x_D^i) &= \sum_{i=1}^{N}\Big(-\frac{\sum_{d=1}^{D}\|z_d^{(1)i}\|^2-\frac{1}{D}\sum_{d=1}^{D}\sum_{l=1}^{D}\langle z_d^{(1)i},z_l^{(1)i}\rangle}{2\sigma^2}\\ &\quad+\log f(\bar{s}_0^i)+\sum_{d=1}^{D}\log p_{Z_d^{(2)}}(z_d^{(2)i})+\sum_{d=1}^{D}\log|W_d|\Big)\\ &= \sum_{i=1}^{N}\log f(\bar{s}_0^i)+\sum_{i=1}^{N}\sum_{d=1}^{D}\log p_{Z_d^{(2)}}(z_d^{(2)i})+N\sum_{d=1}^{D}\log|W_d|\\ &\quad-\frac{1}{2\sigma^2}\Big(\sum_{i=1}^{N}\sum_{d=1}^{D}\|z_d^{(1)i}\|^2-\frac{1}{D}\sum_{i=1}^{N}\sum_{d=1}^{D}\sum_{l=1}^{D}\langle z_d^{(1)i},z_l^{(1)i}\rangle\Big)\\ &= \sum_{i=1}^{N}\log f(\bar{s}_0^i)+\sum_{i=1}^{N}\sum_{d=1}^{D}\log p_{Z_d^{(2)}}(z_d^{(2)i})+N\sum_{d=1}^{D}\log|W_d|\\ &\quad-\frac{1}{2\sigma^2}\Big(\sum_{d=1}^{D}\mathrm{trace}(Z_d^{(1)}Z_d^{(1)\top})-\frac{1}{D}\sum_{d=1}^{D}\sum_{l=1}^{D}\mathrm{trace}(Z_d^{(1)}Z_l^{(1)\top})\Big)\end{aligned}$$

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

$$p_{\mathbf{X}}(\mathbf{x})=p_{\mathbf{Z}^{(1)}}(z_1^{(1)},\dots,z_D^{(1)})\prod_{d=1}^{D}p_{Z_d^{(2)}}(z_d^{(2)})$$

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

$$\begin{aligned}p_{\mathbf{Z}^{(1)}|s_0}(z_1^{(1)},\dots,z_D^{(1)}|s_0)=p_{\mathbf{Z}^{(1)}|s_0}(\alpha(\sigma)(s_0+\epsilon_{10}),\dots,\alpha(\sigma)(s_0+\epsilon_{D0})|s_0)&=\prod_{d=1}^{D}p_{Z_d^{(1)}|S_0}(\alpha(\sigma)(s_0+\epsilon_{d0})|s_0)\\ =\prod_{d=1}^{D}\mathcal{N}(\alpha(\sigma)(s_0+\epsilon_{d0});s_0,\sigma^2\mathbb{I}_c)&=\prod_{d=1}^{D}\mathcal{N}(z_d^{(1)};\alpha(\sigma)s_0,\alpha(\sigma)^2\sigma^2\mathbb{I}_c)\end{aligned}$$

It follows that

$$
\begin{aligned}
p_{\mathbf{Z}^{(1)}}(z_1^{(1)},\dots,z_D^{(1)}) &= \int p_{\mathbf{Z}^{(1)}|s_0}(z_1^{(1)},\dots,z_D^{(1)}|s_0)p_{S_0}(s_0)ds_0 \\
&= \int \Big(\prod_{d=1}^{D}\mathcal{N}(z_d^{(1)};\alpha(\sigma)s_0,\alpha(\sigma)^2\sigma^2\mathbb{I}_c)\Big)p_{S_0}(s_0)ds_0 \\
&\propto \int \exp\Big(-\sum_{d=1}^{D}\frac{\|z_d^{(1)}-\alpha(\sigma)s_0\|^2}{2\alpha(\sigma)^2\sigma^2}\Big)p_{S_0}(s_0)ds_0 \\
&= \int \exp\Big(-\frac{D\|\alpha(\sigma)s_0-\bar{s}_0\|^2+\sum_{d=1}^{D}\|z_d^{(1)}-\bar{s}_0\|^2}{2\alpha(\sigma)^2\sigma^2}\Big)p_{S_0}(s_0)ds_0 \\
&= \exp\Big(-\frac{\sum_{d=1}^{D}\|z_d^{(1)}-\bar{s}_0\|^2}{2\alpha(\sigma)^2\sigma^2}\Big)\int \exp\Big(-\frac{D\|\alpha(\sigma)s_0-\bar{s}_0\|^2}{2\alpha(\sigma)^2\sigma^2}\Big)p_{S_0}(s_0)ds_0
\end{aligned}
$$

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

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

It be easily derived from 4 by making the following observations resulting from whitening

- $N\sum_{d=1}^{D}\log|W_d| = ND$ since $\forall d\; W_d$ is orthogonal
- $\mathrm{trace}(Z_d^{(1)}Z_d^{(1)\top}) = c$ due to Lemma A.1

□

Remark. Let $D = 1$ and we have the following simple BSS model for $k < p$:

$$x = As \text{ with } x \in \mathbb{R}^p, s \in \mathbb{R}^k, A \in \mathbb{R}^{p\times k}$$

It follows that $s = (A^\top A)^{-1}A^\top x$. Define $W = (A^\top A)^{-1}A^\top$. The density function of $p_s(s)$ is given by (, https://math.stackexchange.com/users/491644/maxim)

$$p_s(s) = \det(WW^\top)^{-\frac{1}{2}}\int_{Wx=s} p_x(x)dS(x),$$

where we integrate over a $p-k$ dimensional surface. Thus, in that case we use this representation for our optimization.

B Real Data Experiment

B.1 Data Acquisition and Preprocessing

Our analysis is primarily based on two 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.

We removed genes with missing values from Dataset 1 and we selected 3994 genes that are present in both datasets. To evaluate our results, we collect a ground truth network from the online database *Subti*Wiki [4] which consists of 5,952 pairs of regulator and regulated gene. Since our method predicts pairs of co-regulated genes, we transform the ground truth network into an undirected graph that links genes with a common regulator. Thus, the ground truth network is stored in the form of an adjacency matrix with entries 1 if the genes are co-regulated and 0 otherwise.

B.2 Gene-gene Interaction Pipeline

The main steps of our method are presented in Algorithm 1. We infer latent components from the data as described in Appendix B.2.1. Afterward, we learn a sparse undirected graph from the estimated independent components (see Appendix B.2.2).

B.2.1 Data Integration

Let $X \in \mathbb{R}^{n \times p}$ be a transcriptome data matrix with n samples (or experimental outcomes) and p genes. We assume that the transcriptome matrix follows a linear latent model, i.e. there exist a matrix $A \in \mathbb{R}^{n \times k}$ and a matrix $S \in \mathbb{R}^{k \times p}$ such that $X = AS$. The k components can be represent gene expression. If a group of genes is either over or under-expressed in a specific component they are usually assumed to share a functional property in the genome. Additionally, if the components are independent (i.e. a BSS model) we assume that the components represent independent gene pathways, i.e. the components' groups of over/under-expressed genes act independently from each other given the experimental conditions.

PLS (OmicsPLS) This baseline is not a BSS model, i.e. the estimated components are not necessarily independent. We make an additional assumption that the view-specific sources are orthogonal to the other views. The model is defined by

$$X_1 = A_1 Y_1 + B_1 Z_1 + E_1$$
$$X_2 = A_2 Y_2 + B_2 Z_2 + E_2,$$

where $Y_1 \in \mathbb{R}^{c \times n}$ $Y_2 \in \mathbb{R}^{c \times n}$ are the latent variables that are responsible for the joint variation between X_1 and X_2, i.e. Y_1 and Y_2 are obtained by solving a CCA problem, and $Z_i \in \mathbb{R}^{k_i - c \times n}$ represent the components that are orthogonal to X_j with $j \neq i$, and E_i is the noise (or residuals). In our application we define $S_i = (Y_i, Z_i)$ for the downstream task of interest.

B.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 which glasso solves can be formalized as follows

$$\min_{\Theta \succ 0} - \log \det(\Theta) + \mathrm{tr}(\hat{\Sigma}\Theta) + \lambda \|\Theta\|_1, \tag{6}$$

[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] See http://www.subtiwiki.uni-goettingen.de/v4/exports

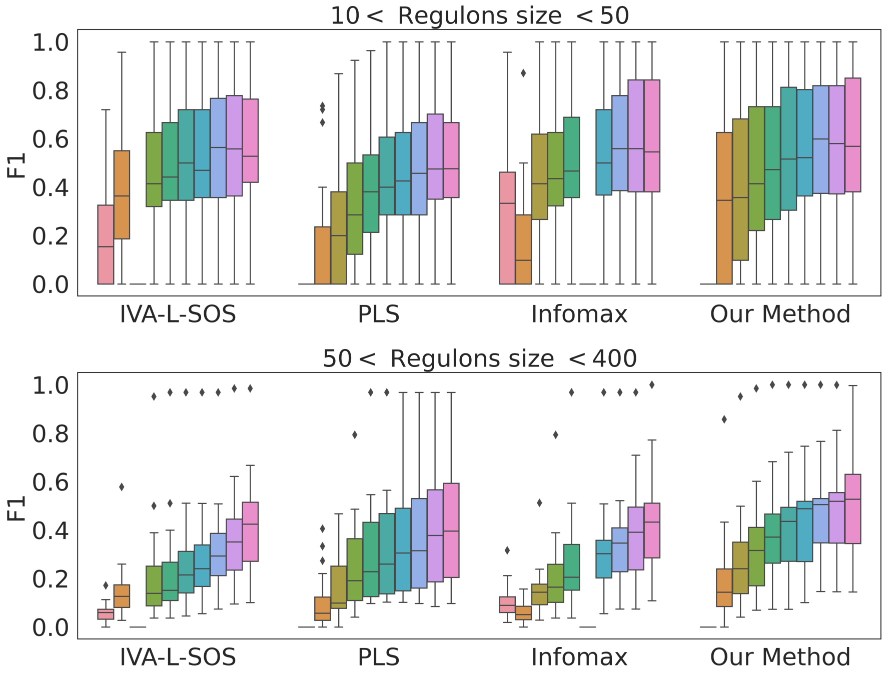


Figure 4: Boxplots of the regulons F1 score for two groups of regulons depending on their size (between 10 and 50, and between 50 and 400).

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 graphical lasso.

B.2.3 Extended EBIC

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

$$-\log\det(\Theta(E)) + \mathrm{tr}(\hat{\Sigma}\Theta(E)) + |E|\log n + 4|E|\gamma\log p, \tag{7}$$

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 & 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 $\gamma = 0.5$.

B.2.4 Precision and Recall

To evaluate the proposed method, we use two different evaluation strategies. First, we count the true positive and false positive (or unknown) edges from the output undirected graph. Edges are annotated as true positive if they connect pairs of co-regulated genes. In the second part of our evaluation, we are interested in the regulon prediction power of our method. For each known regulon, we compute precision and recall score in the following way:

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

$$\text{Precision}(R) = \frac{\sum_{g \in R} |N(g) \cap N^{gt}(g)|}{\sum_{g \in R} |N(g)|},$$
$$\text{Recall}(R) = \frac{\max_{C \in \mathcal{C}(R)} |C|}{|R|}$$

where R denotes the set of regulon genes, $N(g)$ and $N^{gt}(g)$ are the sets of all neighbours of gene g in the output network and ground truth network, respectively, and $\mathcal{C}(R)$ the set of all connected components in the induced graph with vertices in R. From that we can compute the F1 score per regulon. We evaluated each procedure and we aggregated all F1 scores in a boxplot Figure 4.

B.2.5 Method

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

Algorithm 1 Algorithmic description of the data integration task.

```
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
    $\Lambda$ is a set of regularization parameters
    $\gamma$ EBIC selection parameter (7)
Perform a data integration method to obtain $S_1, \in \mathbb{R}^{k_1 \times p}, S_2 \in \mathbb{R}^{k_2 \times p}$
Concatenate $S = (S_1, S_2)^\top \in \mathbb{R}^{k_1 + k_2 \times p}$
Compute the Pearson correlation matrix $\hat{\Sigma} \in \mathbb{R}^{p \times p}$ of $S$.
Estimate the precision matrices $\{\hat{\Theta}^\lambda\}_{\lambda \in \Lambda}$ which solves 6 for each $\lambda$ from the set $\Lambda$
Select the final $\hat{\Theta}^{out} \in \{\hat{\Theta}^\lambda\}_{\lambda \in \Lambda}$ according to EBIC($\gamma$) (see 7)
Output:
    the selected $\hat{\Theta}^{out}$
```

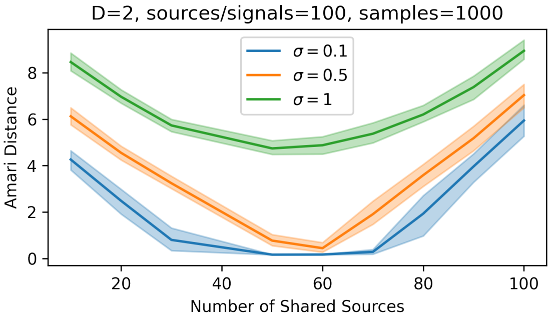
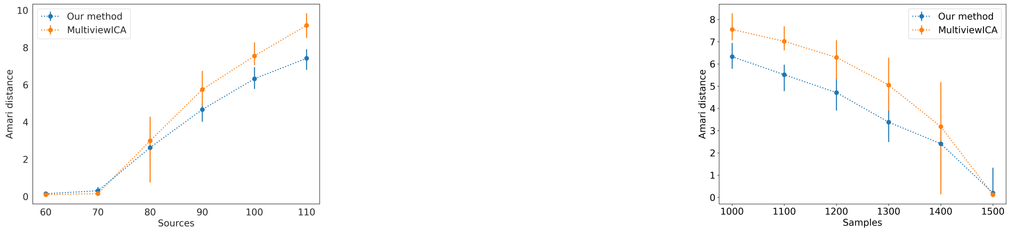


Figure 5: We have the two view case again with number of total sources and observed signals 100 and number of samples 1000. We consider three cases of noise standard deviation: $\sigma = 0.1, 0.5, 1$. As soon as enough shared sources are present (around 60) our method lower value of Amari distance (the lower the better) in all cases. In the 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.



(a)

(b)

Figure 6: Comparison of MultiViewICA and our method on a two-view shared response model setting. In Figure 6a we fix the sample size and measure the Amari distance for sources $60, 70, \dots 110$. In Figure 6b the number of sources is set to 100 and we conduct the experiments for different sample sizes (x-axis). It seems that our method outperforms MultiViewICA in both scenarios.

C Synthetic Experiments

C.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

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

C.2 Additional Experiments on Synthetic Data

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 5. In the first two cases the results are close to the one discussed in the main paper. As expected, by adding noise with high variance ($\sigma = 1$) our method does not converge and affects the quality of the estimated mixing matrices measured with the Amari distance. The whole procedure is repeated 50 times, and the error bars are the 95% confidence intervals based on the independent runs.

Objective function motivation. In the following experiment, we compare MultiViewICA and our method when the observed data is high-dimensional on a two-view shared response model applications, i.e. no individual sources. The experimental setup allows for comparing standard MLE (MultiViewICA) and MLE after whitening (Our Method). Figure 6a compares the two methods for fixed sample size 1000. In Figure 6b we fixed the number of sources to be 100 and vary the sample size.

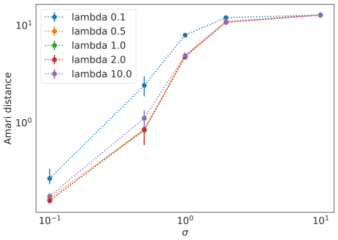


Figure 7: Choice of Hyperparameter λ. The data comes from a two-view model with 50 shared and 50 individual sources per view. The x-axis is represents the noise standard deviation and the y-axis the Amari distance.

Study	Application	Observed Signals	Latent Sources	Views
Salman et al. (2019)	Identifying biomarkers	fMRI data	brain functional networks	multiple subjects
(Durieux & 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	heterogenous 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 interpetation.

For all experiments the noise standard deviation is 0.01. It seems that our method 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.

Choice of λ For this experiment we used data generated 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 7 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.

C.3 Implementation

The code for GroupICA, ShICA, MultViewICA is distributed with BSD 3-Clause License. The `OmicsPLS` R library has a GPL-3 license, the `scikit-learn` library is distributed with BSD 2-Clause License. Our code is available under `https://anonymous.4open.science/r/shindica-C497/`

D Model Justification

Multiview 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 multiview 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 multiview 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 multiview ICA models were used in Table 1. We also provided an interpretation of 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 above-mentioned areas of applications (Dubois & Adolphs, 2016) , such as (Seghier & 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 we 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 our method is a valuable addition to this set of tools. In the independent component analysis context, we are unaware of a similar model that both provides identifiability results and an optimization procedure that maximizes the direct data log-likelihood for given source priors.

Linearity assumption in the biomedical domain. The linear assumption can be explained by the nature of the data in the targeted domains. 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 & 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 often we 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. Event though a non-linear multiview version will be a valuable addition to the current active research on non-linear ICA, e.g. (Hyvärinen & 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. Moreover, there are other non-linear multi-view versions, as stated in our work, that lack identifiability.

E 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 & 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).