ON THE IDENTIFIABILITY OF NONLINEAR ICA WITH UNCONDITIONAL PRIORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Nonlinear independent component analysis (ICA) aims to recover the underlying marginally independent latent sources from their observable nonlinear mixtures. The identifiability of nonlinear ICA is a major unsolved problem in unsupervised learning. Recent breakthroughs reformulate the standard marginal independence assumption of sources as conditional independence given some auxiliary variables (e.g., class labels) as weak supervision or inductive bias. However, the modified setting is not applicable in many scenarios that do not have auxiliary variables. We explore an alternative path and consider instead only assumptions on the mixing process, such as the pairwise orthogonality among the columns of the Jacobian of the mixing function. We show that marginally independent latent sources can be identified from strongly nonlinear mixtures up to a component-wise transformation and a permutation, thus providing, to the best of our knowledge, a first full identifiability result of nonlinear ICA without auxiliary variables. We provide an estimation method and validate the theoretical results experimentally.

1 INTRODUCTION

Nonlinear independent component analysis (ICA) is a fundamental problem in unsupervised representation learning. It generalizes linear ICA (Hyvärinen et al., 2001; Comon, 1994) to identify latent sources from observations, which are assumed to be an invertible nonlinear mixture of them. For an observed vector $\mathbf{x}$, nonlinear ICA expresses it as $\mathbf{x} = f(\mathbf{s})$, where f is an unknown invertible mixing function, and $\mathbf{s}$ is a latent random vector representing the marginally independent sources. The goal is to recover function f as well as sources $\mathbf{s}$ from the observed mixture $\mathbf{x}$. While there is an increasing interest in nonlinear ICA in various tasks, such as disentanglement (Lachapelle et al., 2021), its identifiability remains a major obstacle for providing a principled framework. The key difficulty is that, without additional assumptions, there exist infinite ways to transform the observations into independent components while still mixed w.r.t. the sources (Hyvärinen & Pajunen, 1999).

To deal with this challenge, existing works introduce the auxiliary variable $\mathbf{u}$ (e.g., label, domain index, timestamp) and assume that sources are conditionally independent given $\mathbf{u}$ (Hyvarinen & Morioka, 2017; Hyvarinen et al., 2019; Khemakhem et al., 2020). Most of them require the auxiliary variable to be observable, while Hyvarinen & Morioka (2016) and Hälvä & Hyvarinen (2020) are exceptions. These works impose mild restrictions on the mixing process but require many distinct values of $\mathbf{u}$ (e.g., number of classes) for full identifiability, which might restrict their practical applicability. Motivated by this, Yang et al. (2022) reduce the number of required distinct values by strengthening functional assumptions on the mixing process. As described, all these results are based on conditional independence, instead of the standard marginal independence assumption, of the sources in order to achieve identifiability in specific tasks, i.e., those with auxiliary variables as weak supervision. The identifiability of the original setting focusing on marginal independence, which is crucial for unsupervised learning, stays unsolved.

Instead of introducing auxiliary variables, there exists the possibility to rely on further restrictions on the mixing functions to achieve identifiability, but limited results are available in the literature over the past two decades and are often based on overly restrictive conditions. With the assumption of conformal transformation, Hyvärinen & Pajunen (1999) proved identifiability of the nonlinear ICA solution up to rotational indeterminacy for only two sources. Another result is shown on the mixture that consists of component-wise nonlinearities added to a linear mixture (Taleb & Jutten, 1999),

which relies on the linearity assumption. MND (Zhang & Chan, 2008) improves the performance of separation by preferring solutions with high linearity, while no identifiability result is given.

In this work, we aim to show full identifiability of nonlinear ICA with unconditional priors under specific conditions. We show that without any kind of auxiliary variables, the latent sources can be identified from nonlinear mixtures up to a component-wise invertible transformation and a permutation under assumptions on the mixing process. Specifically, we study the following conditions on the mixing process: (1) pairwise orthogonal columns of the Jacobian matrix, and (2) factorial Jacobian determinant (e.g., volume-preserving transformation). Different from existing works relying on weak supervision of auxiliary variables (observable or not), we establish a new identifiability result in a fully unsupervised setting, which is within the original setting of ICA that is based on marginal independence of the latent sources.

In particular, this is consistent with previous results on ruling out some spurious solutions (Gresele et al., 2021) and reducing the number of required labels (Yang et al., 2022) by restricting mixing function f. Besides, we show that the component-wise indeterminacy can be further reduced to the component-wise linear transformations if the true sources are assumed to be Gaussian. This work establishes a mathematical principle underlying unsupervised representation learning, opening up the possibility of identifiable nonlinear ICA under the standard marginal independence assumption, while not restricting it to the two-dimensional or the post-nonlinear case. We further develop an estimation method and validate our theoretical claims experimentally.

2 Nonlinear ICA with unconditional priors

2.1 Identifiability of nonlinear ICA

We consider the following data-generating process of nonlinear ICA

$$p_{\mathbf{s}}(\mathbf{s}) = \prod_{i=1}^{n} p_{\mathbf{s}_i}\left(\mathbf{s}_i\right), \tag{1}$$

$$\mathbf{x} = f(\mathbf{s}), \tag{2}$$

where $\mathbf{x}$ denotes the observed random vector, $\mathbf{s}$ is the latent random vector representing the marginally independent sources, $p_{\mathbf{s}_i}$ is the marginal probability density function (pdf) of $\mathbf{s}_i$, $p_{\mathbf{s}}$ is the joint pdf, and $f : \mathbf{s} \to \mathbf{x}$ denotes a nonlinear mixing function. Note that different from previous works for identifiable nonlinear ICA, sources $\mathbf{s}$ do not need to be mutually conditionally independent given an auxiliary variable. Instead, our setting is consistent with the traditional ICA problem and thus unifies ICA with auxiliary variables as a special case. Besides, no particular form of the source distribution is assumed at this point, which is similar to (Yang et al., 2022) and weaker than works assuming an exponential family (Hyvarinen et al., 2019; Hälvä & Hyvarinen, 2020; Khemakhem et al., 2020; Lachapelle et al., 2021). Following previous works, we assume the function f is invertible and twice differentiable. We further assume that the column vectors of the Jacobian matrix of f are pairwise orthogonal, i.e.,

$$\mathbf{J}_f(\mathbf{s}) = \mathbf{O}(\mathbf{s})\mathbf{D}(\mathbf{s}), \tag{3}$$

where $\mathbf{O}(\mathbf{s})$ is an orthogonal matrix and $\mathbf{D}(\mathbf{s})$ is a diagonal matrix.

Regarding the estimating model, similarly, we set the columns of Jacobian of estimating function $g^{-1} : \mathbf{x} \to \hat{\mathbf{s}}$ orthogonal to each other and with equivalent euclidean norm, i.e., $\mathbf{J}_{g^{-1}}(\mathbf{x}) = \mathbf{O}(\mathbf{x})\lambda(\mathbf{x})$, where $\lambda(\mathbf{x})$ is a scalar. Note that restricting $\mathbf{D}$ to λ in the estimating process does not affect the component-wise similarity between the recovered and true sources. We assume that the distribution of the estimated sources $\hat{\mathbf{s}}$ follows a factorial multivariate Gaussian:

$$p_{\hat{\mathbf{s}}}(\hat{\mathbf{s}}) = \prod_{i=1}^{n} \frac{1}{Z_i} \exp\left(-\theta_{i,1}\hat{\mathbf{s}}_i - \theta_{i,2}\hat{\mathbf{s}}_i^2\right), \tag{4}$$

where Z_i is a constant. The sufficient statistics $\theta_{i,1} = -\frac{\mu_i}{\sigma_i^2}$ and $\theta_{i,2} = \frac{1}{2\sigma_i^2}$ are assumed to be linearly independent. We set the variances σ_i^2 to be distinct without loss of generality, which can be placed as a constraint during the estimation process as the sufficient statistics of the estimated distribution are trainable. Note that the density $p_{\hat{\mathbf{s}}}(\hat{\mathbf{s}})$ must be a push-forward of $p_{\mathbf{s}}(\mathbf{s})$ through $g^{-1} \circ f$, where g^{-1} is the estimated unmixing function. This implicitly restricts $p_{\mathbf{s}}(\mathbf{s})$. The main result is as follows:

Theorem 2.1. *Assume*

i. Data points are sampled from a model defined in Eqs. (1-4), where columns of the Jacobian matrix $\mathbf{J}_f(\mathbf{s})$ are pairwise orthogonal.

ii. The mixing function f and its inverse are twice differentiable.

iii. The Jacobian determinant of mixing function can be factorized as $|det(\mathbf{J}_f(\mathbf{s}))| = \prod_{i=1}^{n} y_i(\mathbf{s}_i)$, where y_i is an arbitrary function with $y_i(\mathbf{s}_i) > 0$. Note that volume-preserving transformation is a special case when $y_i(\mathbf{s}_i) = 1$.

Then function $h := g^{-1} \circ f$ is a composition of a component-wise invertible transformation and a permutation.

The proof is provided in Appendix A.1. The assumption of pairwise orthogonal columns in the Jacobian matrix of the mixing function is also formalized as the criterion of IMA by Gresele et al. (2021), which, in other words, assumes that the partial derivatives $\partial \mathrm{f}/\partial \mathbf{s}_i$ are independent to each other in a *non-statistical* sense (Gresele et al., 2021) It has been shown theoretically that some spurious solutions can be ruled out by assuming IMA. Specifically, IMA is proved to be useful to distinguish specific models (e.g., conformal maps) from Darmois solutions (Darmois, 1951; Hyvärinen & Pajunen, 1999) and "rotated-Gaussian" measure-preserving automorphisms (MPAs) (Locatello et al., 2019). Inspired by this, we adopt Assumption (i) based on the pairwise orthogonality and further show that it is critical to the full identifiability of nonlinear ICA.

The fact that randomly chosen high-dimensional vectors from an isotropic prior are likely to be (near-)orthogonal (Ball et al., 1997) demonstrates that it may be a reasonable assumption in specific (but not rare) scenarios. To build intuition, one can consider a filmmaking process in a safari park, of which the main characters are wild animals. The task of boom operators here is to record the sound of animals. As there are many animals and microphones, the mixing (recording) process is strongly dependent on the relative positions between them. Thus, the animal (source $\mathbf{s}_i$) influences the recording (mixing function f) through its position. The moving direction can be represented as the i-th column vector of the Jacobian $\mathbf{J}(f)_{:,i}$. Assuming that there is a sufficiently large number of animals and that they are not cooperative enough to fine-tune their positions for better recording, the individual direction $\mathbf{J}(f)_{:,i}$ is likely to be almost random. By formalizing that randomness as randomly sampling from a standard Gaussian or other isotropic priors, the vector $\mathbf{J}(f)_{:,i}$ that represents its direction is likely to be orthogonal to the others. The connections between the orthogonality and other criteria have also been discussed by Janzing et al. (2010); Gresele et al. (2021) in the literature of causality. Other practical scenarios include fields that adopt orthogonal coordinate transformations (Gresele et al., 2021), such as dynamic control (Mistry et al., 2010) and structural geology (De Paor, 1983).

Having said that, a violation of the pairwise orthogonality assumption in real-world scenarios could be ascribed to a deliberate global adjustment of sources. For example, cinema audio systems are carefully adjusted to achieve a homogeneous sound effect on every audience. The position and orientation of each speaker are fine-tuned according to the others. In this case, it may be difficult to distinguish the influences from different speakers because of the fine-tuning. This leads to a high degree of multicollinearity across the columns of Jacobian, thus violating the assumption of pairwise orthogonality. Other possible violations can also be found in similar scenarios where the influences from different sources are fine-tuned in order to achieve homogeneity (e.g., the cocktail party problem described by Gresele et al. (2021)), or in an unlikely scenario where the mixing process is naturally component-wise (e.g., an autoregressive model (Papamakarios et al., 2021)).

Regarding Assumption (iii), we first note that mixing functions with factorial Jacobian determinants are of a much wider range as compared to component-wise transformations. To see this, a straightforward example is the volume-preserving transformation, which has been widely adopted in generative models like those based on flows (Zhang et al., 2021). In fact, all transformations with constant Jacobian determinant can be factorized w.r.t. the sources, which, generally speaking, are not component-wise. Besides, there exists other non-volume-preserving transformations with factorial Jacobian determinant.[1] The empirical results in Yang et al. (2022) indicate this, though our current knowledge of this type of function is rather limited. Moreover, the volume-preserving assumption

[1] A toy example: the Jacobian determinant of mixing function w.r.t. $(\mathbf{x}_1 = \frac{m\mathbf{s}_1\mathbf{s}_2}{n+m\mathbf{s}_2}, \mathbf{x}_2 = \frac{n\mathbf{s}_1}{n+m\mathbf{s}_2})$ is $|mn\mathbf{s}_1|$.

[2] has been demonstrated by Sorrenson et al. (2020); Yang et al. (2022) to be helpful to weaken the requirement of auxiliary variables, specifically by reducing the number of required labels. Our work takes this further by imposing a weaker version, i.e., Assumption (iii), together with other constraints, which completely removes the need for any auxiliary variable.

Theorem 2.1 opens up the possibility of the full identifiability of nonlinear ICA while maintaining the standard mutually marginal independence assumption of sources. This is consistent with the original setting of ICA and plays an important role in a more general range of unsupervised learning tasks, compared to previous works relying on conditional independence given auxiliary variables. In fact, one can easily modify Theorem 2.1 to include generating process based on auxiliary variables by imposing similar restrictions on the mixing process. Different from previous works (Hyvarinen & Morioka, 2016; 2017; Hyvarinen et al., 2019; Lachapelle et al., 2021; Yang et al., 2022), this extension removes common restrictions on the required auxiliary variables, such as the number of distinct values (e.g., labels), non-proportional variances, or sufficient variability. Therefore, although our main result (Theorem 2.1) focuses on nonlinear ICA with unconditional priors, it also provides potential insight on improving the flexibility of utilizing the auxiliary variable when it is available. As a trade-off, the additional assumptions on the mixing process might limit its usage in some scenarios.

2.2 Identifiability up to component-wise linear transformation

Most works in nonlinear ICA focus on identifiability up to a component-wise invertible transformation and a permutation (Hyvarinen & Morioka, 2016; Hyvarinen et al., 2019; Hälvä & Hyvarinen, 2020; Khemakhem et al., 2020), which is also called *permutation-identifiability* in the literature of disentanglement (Lachapelle et al., 2021). This indeterminacy is trivial compared to the fundamental nonuniqueness of nonlinear ICA (Hyvärinen & Pajunen, 1999) and is analogous to the indeterminacy involving rescaling and permutation in linear ICA (Taleb & Jutten, 1999). Recently, Yang et al. (2022) provided an identifiability result based on auxiliary variables that further reduces the indeterminacy of the component-wise nonlinear transformation to a linear one. Inspired by it, the following theorem provides identifiability conditions of nonlinear ICA up to a component-wise linear transformation with unconditional priors. This is interesting as it opens up the possibility for potential further optimization with only element-wise linear operations, such as the Hadamard product.

Theorem 2.2. *Besides the assumptions in Theorem 2.1, assume that the source distribution $p_{\mathbf{s}}(\mathbf{s})$ is a factorial multivariate Gaussian and function f is volume-preserving. Then $h := g^{-1} \circ f$ is a composition of a component-wise <u>linear</u> transformation and a permutation.*

3 Experiments

Setup. Based on the proposed theorems, we use a volume-preserving flow to recover the sources up to the indeterminacies mentioned above. Similar to (Gresele et al., 2021), we consider a regularized maximum likelihood estimation to enforce orthogonality between columns of the Jacobian. Differently, we propose a new objective to unify the norm of columns in the estimation, i.e.,

$$\mathcal{L}(\mathbf{g}; \mathrm{x}) = \mathbb{E}_{\mathbf{x}} \left[\log p_{\mathbf{g}}(\mathbf{x}) - \lambda_1 \left(n \log \left(\frac{1}{n} \sum_{i=1}^{n} \left\| \frac{\partial g^{-1}}{\partial \hat{s}_i} \right\| \right) - \log \left| \mathrm{J}_{g^{-1}}(\hat{s}) \right| \right) \right],$$

where λ_1 is a regularization parameter. The following proposition summarizes its related properties:

Proposition 3.1. *The following inequality holds*

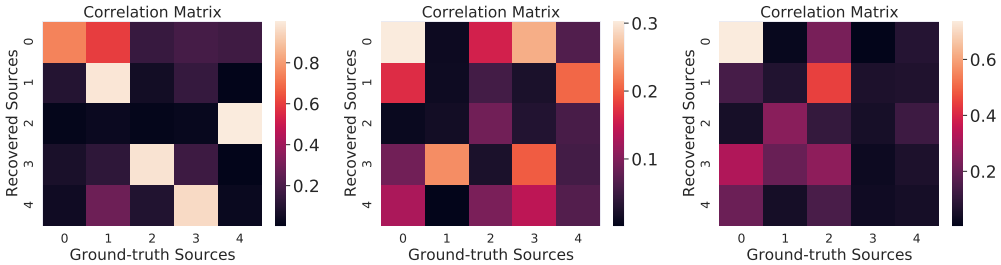
$$n \log \left(\frac{1}{n} \sum_{i=1}^{n} \left\| \frac{\partial g^{-1}}{\partial \hat{s}_i} \right\| \right) - \log \left| \mathrm{J}_{g^{-1}}(\hat{s}) \right| \geq 0,$$

with equality iff. $\mathrm{J}_{g^{-1}}(\hat{s}) = \mathbf{O}(\hat{s})\lambda(\hat{s})$, where $\mathbf{O}(\hat{s})$ is an orthogonal matrix and $\lambda(\hat{s})$ is a scalar.

The proof is provided in Appendix 3.1. We train a volume-preserving flow to maximize $\mathcal{L}$.

We first conduct an ablation study to verify the necessity of the proposed assumptions.

[2]For factorial Jacobian determinant, we majorly focus on volume-preserving transformations since the properties of other functions are not yet clear to us.



(a) *Full*. (b) *VP*. (c) *NVP*.

Figure 2: Pearson correlation matrices between the ground-truth and the recovered sources.

Specifically, we focus on the two major conditions, i.e., orthogonal columns and volume-preserving. For volume-preserving transformations, we train a volume-preserving flow to maximize the likelihood, which constraints the Jacobian determinant of GLOW (Kingma & Dhariwal (2018)). For non-volume-preserving transformations, we use GLOW instead. The observational data are generated according to the required assumptions, as detailed in Appendix B. We evaluate the models by the mean correlation coefficient (MCC) between ground truth and reconstructed sources.

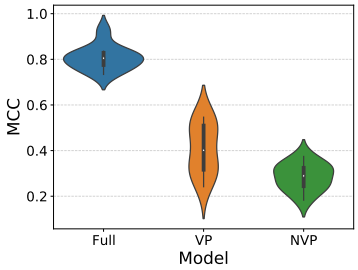


Figure 1: MCC when part of the conditions are not satisfied ($n = 5$).

Results. Results for each model are summarised in Figure 1. We construct each model as follows: *(Full)* All conditions satisfied; *(VP)* Compared to *Full*, columns of Jacobian of (un)mixing functions are not orthogonal; *(NVP)* Compared to *VP*, (un)mixing functions are not volume-preserving. One can observe that when the proposed assumptions are fully satisfied (*Full*), our model achieves the highest MCC on average. Besides, according to the decline of performance after violating each of these conditions, the necessity of them is supported. The visualization of the Pearson correlation matrices is shown in Figure 2. All results are of 10 trials.

4 Conclusion

We provide the full identifiability result for nonlinear ICA with unconditional priors. In particular, we prove that the i.i.d. latent sources can be recovered up to a component-wise invertible transformation and a permutation with only conditions on the nonlinear mixing process. To the best of our knowledge, this is the first identifiability result of nonlinear ICA with unconditional priors while not restricting it to the two-dimensional or post-nonlinear cases. Therefore it is consistent with the original setting of ICA that is based on the standard marginal independence assumption of latent sources, while previous works rely on conditional independence on auxiliary variables as weak supervision or inductive bias.

Besides, by further assuming the Gaussianity of sources, we prove that the indeterminacy of the component-wise transformation can be reduced to a component-wise linear transformation, which sheds light on potential further optimizations by element-wise linear operations. By including the auxiliary variable as a special case but without common constraints on it (e.g., number of class labels), our results may provide extra flexibility in specific scenarios when auxiliary variables are available.

Limitation and discussion. Although our identifiability result is in some sense more general than previous works as it removes the need for auxiliary variables, the assumptions on the mixing process may restrict its usage in specific scenarios as discussed in Section 2.1. This is inevitably a trade-off between introducing auxiliary variables and imposing restrictions on the mixing process in order to achieve full identifiability, whose practical usage depends on the scenario and information available. It is worth noting that Gresele et al. (2021) assumed orthogonality, conformal map, and the others, which are already rather restrictive, in order to rule out two specific types of spurious solutions (i.e., Darmois construction and "rotated-Gaussian" MPAs) without introducing auxiliary variables.

REFERENCES

Keith Ball et al. An elementary introduction to modern convex geometry. *Flavors of geometry*, 31 (1-58):26, 1997.

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

G Darmois. Analyse des liaisons de probabilité. In *Proc. Int. Stat. Conferences 1947*, pp. 231, 1951.

Declan G De Paor. Orthographic analysis of geological structures—i. deformation theory. *Journal of Structural Geology*, 5(3-4):255–277, 1983.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

Harley Flanders. Liouville's theorem on conformal mapping. *Journal of Mathematics and Mechanics*, 15(1):157–161, 1966.

Luigi Gresele, Julius von Kügelgen, Vincent Stimper, Bernhard Schölkopf, and Michel Besserve. Independent mechanism analysis, a new concept? *Advances in Neural Information Processing Systems*, 2021.

Hermanni Hälvä and Aapo Hyvarinen. Hidden markov nonlinear ica: Unsupervised learning from nonstationary time series. In *Conference on Uncertainty in Artificial Intelligence*, pp. 939–948. PMLR, 2020.

Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in Neural Information Processing Systems*, 29:3765–3773, 2016.

Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ica of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pp. 460–469. PMLR, 2017.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999.

Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. Independent component analysis. *John Wiley & Sons, Inc*, 2001.

Aapo Hyvarinen, 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.

D Janzing, P Hoyer, and B Schölkopf. Telling cause from effect based on high-dimensional observations. In *27th International Conference on Machine Learning (ICML 2010)*, pp. 479–486. Omnipress, 2010.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pp. 2207–2217. PMLR, 2020.

Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ica. *Causal Learning and Reasoning*, 2021.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pp. 4114–4124. PMLR, 2019.

Michael Mistry, Jonas Buchli, and Stefan Schaal. Inverse dynamics control of floating base systems using orthogonal decomposition. In *2010 IEEE international conference on robotics and automation*, pp. 3406–3412. IEEE, 2010.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Peter Sorrenson, Carsten Rother, and Ullrich Köthe. Disentanglement by nonlinear ica with general incompressible-flow networks (gin). *arXiv preprint arXiv:2001.04872*, 2020.

Anisse Taleb and Christian Jutten. Source separation in post-nonlinear mixtures. *IEEE Transactions on signal Processing*, 47(10):2807–2820, 1999.

Lloyd N Trefethen and David Bau III. *Numerical linear algebra*, volume 50. Siam, 1997.

Xiaojiang Yang, Yi Wang, Jiacheng Sun, Xing Zhang, Shifeng Zhang, Zhenguo Li, and Junchi Yan. Nonlinear ica using volume-preserving transformations. In *International Conference on Learning Representations*, 2022.

Kun Zhang and Laiwan Chan. Minimal nonlinear distortion principle for nonlinear independent component analysis. *Journal of Machine Learning Research*, 9(Nov):2455–2487, 2008.

Shifeng Zhang, Chen Zhang, Ning Kang, and Zhenguo Li. ivpf: Numerical invertible volume preserving flow for efficient lossless compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 620–629, 2021.

A PROOFS

A.1 PROOF OF THEOREM 2.1

Theorem 2.1. *Assume*

i. *Data points are sampled from a model defined in Eqs. (1-4), where columns of the Jacobian matrix $\mathbf{J}_f(\mathbf{s})$ are pairwise orthogonal.*

ii. *The mixing function f and its inverse are twice differentiable.*

iii. *The Jacobian determinant of mixing function can be factorized as $|det(\mathbf{J}_f(\mathbf{s}))| = \prod_{i=1}^{n} y_i(\mathbf{s}_i)$, where y_i is an arbitrary function with $y_i(\mathbf{s}_i) > 0$. Note that volume-preserving transformation is a special case when $y_i(\mathbf{s}_i) = 1$.*

Then function $h := g^{-1} \circ f$ is a composition of a component-wise invertible transformation and a permutation.

Proof. Our goal here is to show that function h is a permutation with a component-wise invertible transformation of sources.

By using chain rule repeatedly, we write the Jacobian of function $h := g^{-1} \circ f$ equivalently as

$$\begin{aligned}
\mathbf{J}_h(\mathbf{s}) &= \mathbf{J}_{g^{-1}\circ f}(\mathbf{s}) \\
&= \mathbf{J}_{g^{-1}}(f(\mathbf{s}))\mathbf{J}_f(\mathbf{s}) \\
&= \mathbf{J}_{g^{-1}}(\mathbf{x})\mathbf{J}_f(\mathbf{s}) \\
&= \hat{\mathbf{O}}(\mathbf{x})\hat{\lambda}(\mathbf{x})\mathbf{O}(\mathbf{s})\mathbf{D}(\mathbf{s}) \\
&= \hat{\lambda}(\mathbf{x})\hat{\mathbf{O}}(\mathbf{x})\mathbf{O}(\mathbf{s})\mathbf{D}(\mathbf{s}),
\end{aligned} \tag{5}$$

where $\hat{\lambda}(\mathbf{x})$ is a scalar, $\hat{\mathbf{O}}(\mathbf{x})$ and $\mathbf{O}(\mathbf{s})$ represent the corresponding orthogonal matrices, and $\mathbf{D}(\mathbf{s})$ denotes the corresponding diagonal matrix. The fourth equality follows from assumption i.

Since $\mathbf{J}_h(\mathbf{s}) = \hat{\lambda}(\mathbf{x})\hat{\mathbf{O}}(\mathbf{x})\mathbf{O}(\mathbf{s})\mathbf{D}(\mathbf{s})$ by Eq. 5 and $\hat{\mathbf{O}}(\mathbf{x})\mathbf{O}(\mathbf{s})$ is an orthogonal matrix, the columns of $\mathbf{J}_h(\mathbf{s})$ are orthogonal to each other, i.e.,

$$\mathbf{J}_h(\mathbf{s})_{:,j}^{\top}\mathbf{J}_h(\mathbf{s})_{:,k} = 0, \quad \forall j \neq k, \tag{6}$$

where $\mathbf{J}_h(\mathbf{s})_{:,j}$ is the j-th column of Jacobian matrix $\mathbf{J}_h(\mathbf{s})$. Then it follows that

$$\sum_{i=1}^{n} \left(\frac{\partial h_i}{\partial \mathbf{s}_j} \frac{\partial h_i}{\partial \mathbf{s}_k} \right) = 0, \quad \forall j \neq k. \tag{7}$$

By assumption, the components of the true sources $\mathbf{s}$ are mutually independent, while the estimated sources $\hat{\mathbf{s}}$ follows a multivariate Gaussian distribution:

$$\begin{aligned}
p_{\mathbf{s}}(\mathbf{s}) &= \prod_{i=1}^{n} p_{\mathbf{s}_i}(\mathbf{s}_i), \\
p_{\hat{\mathbf{s}}}(\hat{\mathbf{s}}) &= \prod_{i=1}^{n} \frac{1}{Z_i} \exp\left(-\theta_{i,1}\hat{\mathbf{s}}_i - \theta_{i,2}\hat{\mathbf{s}}_i^2\right),
\end{aligned} \tag{8}$$

where Z_i is a constant. The sufficient statistics $\theta_{i,1}$ and $\theta_{i,2}$ are assumed to be linearly independent.

Applying the change of variable rule, we have $p_{\mathbf{s}}(\mathbf{s}) = p_{\hat{\mathbf{s}}}(\hat{\mathbf{s}})|\det(\mathbf{J}_h(\mathbf{s}))|$. According to Eq. 8, we have the following equation after taking the logarithm on both sides:

$$\sum_{i=1}^{n} (\log p_{\mathbf{s}_i}(\mathbf{s}_i)) - \log|\text{det}(\mathbf{J}_h(\mathbf{s}))| = -\sum_{i=1}^{n} \left(\theta_{i,1}h_i(\mathbf{s}) + \theta_{i,2}h_i(\mathbf{s})^2 + \log Z_i\right). \tag{9}$$

According to assumption iii, one can factorize $|\det(\mathbf{J}_h(\mathbf{s}))|$ as $\prod_{i=1}^n y_i(\mathbf{s}_i)$, where y_i is an arbitrary function that $y_i(\mathbf{s}_i) > 0$.[3] This implies that

$$\sum_{i=1}^n (\log p_{\mathbf{s}_i}(\mathbf{s}_i) - y_i(\mathbf{s}_i)) = -\sum_{i=1}^n \left(\theta_{i,1} h_i(\mathbf{s}) + \theta_{i,2} h_i(\mathbf{s})^2 + \log Z_i\right). \tag{10}$$

We use second-order Taylor polynomial to expand h_i as h_i is assumed to be twice-differentiable:

$$h_i(\mathbf{s}_0 + \Delta\mathbf{s}) = h_i(\mathbf{s}_0) + \nabla h_i(\mathbf{s}_0)^\top \Delta\mathbf{s} + \Delta\mathbf{s}^\top \mathbf{H}_{h_i}(\mathbf{s}_0)\Delta\mathbf{s} + \mathcal{O}\left(\Delta\mathbf{s}^\top\Delta\mathbf{s}\right). \tag{11}$$

Same as the proof of Lemma 1 from Equation 17 to Equation 18 in (Yang et al., 2022), taking the Taylor expansion of $\mathbf{h}_i$ in Eq. 9, we have

$$\begin{aligned}\sum_{i=1}^n (\log p_{\mathbf{s}_i}(\mathbf{s}_i) - y_i(\mathbf{s}_i)) = -\sum_{i=1}^n \Bigg(&\theta_{i,1}\nabla h_i(s_0)^\top \Delta\mathbf{s} + \theta_{i,1}\Delta\mathbf{s}^\top \mathbf{H}_{h_i}(\mathbf{s}_0)\Delta\mathbf{s} \\ &+ \theta_{i,2}\left(\nabla h_i(\mathbf{s}_0)^\top \Delta\mathbf{s}\right)^2 + \mathcal{O}\left(\Delta\mathbf{s}^\top\Delta\mathbf{s}\right) + C_i\Bigg),\end{aligned} \tag{12}$$

where C_i is a constant. Because RHS has cross terms $\{\Delta s_j \Delta s_k\}_{j\neq k}$ while LHS does not, the coefficients of them are zeros:

$$\sum_{i=1}^n \left(\theta_{i,1}\frac{\partial^2 h_i}{\partial \mathbf{s}_j \partial \mathbf{s}_k} + 2\theta_{i,2}\frac{\partial h_i}{\partial \mathbf{s}_j}\frac{\partial h_i}{\partial \mathbf{s}_k}\right) = 0, \quad \forall j \neq k. \tag{13}$$

Note that the derivation of Eq. 13 is directly based on that of Lemma 1 in (Yang et al., 2022). The only difference is the term $y_i(\mathbf{s}_i)$, which does not affect the process as it has no cross terms.

Because $\theta_{i,1}$ and $\theta_{i,2}$ are sufficient statistics of a multivariate Gaussian, then $\theta_{i,1} = -\frac{\mu_i}{\sigma_i^2}$ and $\theta_{i,2} = \frac{1}{2\sigma_i^2}$, where μ_i and σ_i represent mean and standard deviation, respectively. We set $\theta_{i,1} = 0$. Plugging it into Eq. 13 yields

$$\sum_{i=1}^n \left(\frac{1}{\sigma_i^2}\frac{\partial h_i}{\partial \mathbf{s}_j}\frac{\partial h_i}{\partial \mathbf{s}_k}\right) = 0, \quad \forall j \neq k. \tag{14}$$

Define the following matrices:

$$\begin{aligned}\Lambda_1 &= \operatorname{diag}\left(\frac{1}{\sigma_1^2}, \cdots, \frac{1}{\sigma_n^2}\right), \\ \Lambda_2 &= \mathbb{I} \in \mathbb{R}^{n\times n}, \\ \Sigma(1,s) &= \operatorname{diag}\left(\sum_{i=1}^n \frac{1}{\sigma_i^2}\left(\frac{\partial h_i}{\partial \mathbf{s}_1}\right)^2, \cdots, \sum_{i=1}^n \frac{1}{\sigma_i^2}\left(\frac{\partial h_i}{\partial \mathbf{s}_n}\right)^2\right), \\ \Sigma(2,s) &= \operatorname{diag}\left(\sum_{i=1}^n \left(\frac{\partial h_i}{\partial \mathbf{s}_1}\right)^2, \cdots, \sum_{i=1}^n \left(\frac{\partial h_i}{\partial \mathbf{s}_n}\right)^2\right).\end{aligned} \tag{15}$$

Then we rewrite Eq. 14 and Eq. 7 as

$$\begin{cases} J_{\mathbf{h}}(\mathbf{s})\Lambda_1 J_{\mathbf{h}}(\mathbf{s}) = \Sigma(1,s), \\ J_{\mathbf{h}}(\mathbf{s})\Lambda_2 J_{\mathbf{h}}(\mathbf{s}) = \Sigma(2,s). \end{cases} \tag{16}$$

Because all elements in matrices defined in Eq. 15 are positive, we can take sqaure roots on both sides of equations in Eq. 16, yielding

$$\begin{cases} \Lambda_1^{1/2} J_{\mathbf{h}}(\mathbf{s}) = \mathbf{O}_1 \Sigma(1,s)^{1/2}, \\ \Lambda_2^{1/2} J_{\mathbf{h}}(\mathbf{s}) = \mathbf{O}_2 \Sigma(2,s)^{1/2}, \end{cases} \tag{17}$$

[3] Note that we use a volume-preserving flow as the estimation model, thus $|\det(\mathbf{J}_{g^{-1}}(\mathbf{x}))| = 1$

where $\mathbf{O}_1$ and $\mathbf{O}_2$ are the corresponding orthogonal matrices. Then we rewrite Eq. 17 as

$$\begin{cases} J_{\mathbf{h}}(\mathbf{s}) = \Lambda_1^{-1/2}\mathbf{O}_1\Sigma(1,s)^{1/2}, \\ J_{\mathbf{h}}(\mathbf{s}) = \Lambda_2^{-1/2}\mathbf{O}_2\Sigma(2,s)^{1/2}. \end{cases} \tag{18}$$

By reformulating,

$$\Sigma(1,s)^{1/2}\Sigma(2,s)^{-1/2} = \mathbf{O}_1^{-1}\Lambda_1^{1/2}\Lambda_2^{-1/2}\mathbf{O}_2. \tag{19}$$

Inspired by Yang et al. (2022), we consider both sides of Eq. 19 as a singular value decomposition of the LHS. Because $\sigma_j \neq \sigma_k, \forall j \neq k$, elements in $\Lambda_1^{1/2}\Lambda_2^{-1/2}$ are distinct, as well as non-degenerate and non-zero. Thus, based on the uniqueness of singular value decomposition (see e.g. Theorem 4.1 in Trefethen & Bau III (1997)), $\mathbf{O}_2$ is a composition of a permutation matrix and a signature matrix. Therefore, $J_{\mathbf{h}}(\mathbf{s}) = \Lambda_2^{-1/2}\mathbf{O}_2\Sigma(2,s)^{1/2}$ is a generalized permutation matrix. It thus follows that the function $h := g^{-1} \circ f$ is a composition of a component-wise invertible transformation and a permutation. □

A.2 Proof of Theorem 2.2

Theorem 2.2. *Besides the assumptions in Theorem 2.1, assume that the source distribution $p_{\mathbf{s}}(\mathbf{s})$ is a factorial multivariate Gaussian and function f is volume-preserving. Then $h := g^{-1} \circ f$ is a composition of a component-wise <u>linear</u> transformation and a permutation.*

Proof. Since $\mathbf{s}$ is from a factorial multivariate Gaussian distribution, we represent its density as

$$p_{\mathbf{s}}(\mathbf{s}) = \prod_{i=1}^{n} \frac{1}{Z_i} \exp\left(-\theta'_{i,1}\mathbf{s}_i - \theta'_{i,2}\mathbf{s}_i^2\right). \tag{20}$$

According to Theorem 2.1, function h is a component-wise invertible transformation of sources, i.e.,

$$h_i(\mathbf{s}) = h_i(\mathbf{s}_i). \tag{21}$$

Combining the above two equations with Eq. 9 and the condition that h is volume-preserving, it follows that

$$\sum_{i=1}^{n}\left(-\theta'_{i,1}\mathbf{s}_i - \theta'_{i,2}\mathbf{s}_i^2\right) = -\sum_{i=1}^{n}\left(\theta_{i,1}h_i(\mathbf{s}_i) + \theta_{i,2}h_i(\mathbf{s}_i)^2 + \log Z_i\right). \tag{22}$$

Then we have

$$\theta'_{i,1}\mathbf{s}_i + \theta'_{i,2}\mathbf{s}_i^2 + \log Z_i = \theta_{i,1}h_i(\mathbf{s}_i) + \theta_{i,2}h_i(\mathbf{s}_i)^2. \tag{23}$$

Clearly, $h_i(\mathbf{s}_i)$ is a linear function of $\mathbf{s}_i$. □

A.3 Proof of Proposition 3.1

Proposition 3.1. *The following inequality holds*

$$n\log\left(\frac{1}{n}\sum_{i=1}^{n}\left\|\frac{\partial g^{-1}}{\partial \hat{s}_i}\right\|\right) - \log\left|\mathrm{J}_{g^{-1}}(\hat{s})\right| \geq 0,$$

with equality iff. $\mathrm{J}_{g^{-1}}(\hat{s}) = \mathbf{O}(\hat{s})\lambda(\hat{s})$, where $\mathbf{O}(\hat{s})$ is an orthogonal matrix and $\lambda(\hat{s})$ is a scalar.

Proof. According to Hadamard's inequality, we have

$$\sum_{i=1}^{n}\log\left\|\frac{\partial g^{-1}}{\partial \hat{\mathbf{s}}_i}\right\| - \log\left|\mathrm{J}_{g^{-1}}(\hat{\mathbf{s}})\right| \geq 0, \tag{24}$$

with equality iff. vectors $\frac{\partial g^{-1}}{\partial \hat{\mathbf{s}}_i}(\hat{\mathbf{s}})$ are orthogonal. Then applying the inequality of arithmetic and geometric means yields

$$n\log\left(\frac{1}{n}\sum_{i=1}^{n}\left\|\frac{\partial g^{-1}}{\partial \hat{\mathbf{s}}_i}\right\|\right) - \log\left|\mathbf{J}_{g^{-1}}(\hat{\mathbf{s}})\right| \geq \sum_{i=1}^{n}\log\left\|\frac{\partial g^{-1}}{\partial \hat{\mathbf{s}}_i}\right\| - \log\left|\mathbf{J}_{g^{-1}}(\hat{s})\right| \geq 0, \tag{25}$$

with the first equality iff. for all i in $\{1, 2, \cdots, n\}$, $\left\|\frac{\partial g^{-1}}{\partial \hat{\mathbf{s}}_i}\right\|$ is equal. Because $\mathbf{J}_{g^{-1}}(\hat{\mathbf{s}})$ is non-singular, $\frac{\partial g^{-1}}{\partial \hat{\mathbf{s}}_i}$ are orthogonal and equal to each other for all i in $\{1, 2, \cdots, n\}$ iff. $\mathbf{J}_{g^{-1}} = \mathbf{O}(\hat{\mathbf{s}})\lambda(\hat{\mathbf{s}})$. □

B EXPERIMENTS

B.1 SYNTHETIC DATASETS AND IMPLEMENTATION DETAILS

In order to generate observational data satisfying required assumptions, we synthetic the sources and mixing process as follows:

Full. Here we describe the generating process that satisfies all our proposed assumptions. Following Gresele et al. (2021), we generate the mixing functions based on Möbius transformations with scaled sources and constant volume. According to Liouville's results (Flanders (1966), also summarized in Theorem F.2 in Gresele et al. (2021)), the Möbius transformation guarantees the orthogonality between the columns vectors of its Jacobian matrix, which achieves uncorrelatedness after centering. Different from Gresele et al. (2021), we scaled the sources while preserving volumes before Möbius transformation with distinct scalers to make sure that the generating process is not a conformal map.

VP. Here we describe the generating process that is volume-preserving but not necessarily with orthogonal columns of Jacobian. We use volume-preserving flow called GIN (Sorrenson et al. (2020)) to generate the mixing function. GIN is a volume-preserving version of RealNVP (Dinh et al. (2016)), which achieves volume preservation by setting the scaling function of the final component to the negative sum of previous ones. [4] We use the official implementation of GIN (Sorrenson et al. (2020)), which is also part of FrEIA.

NVP. Here we describe the generating process without assumptions on volume preservation and orthogonal columns of Jacobian. Following Sorrenson et al. (2020), we use GLOW (Kingma & Dhariwal, 2018) to generate the mixing function. The difference between the coupling block in GLOW and GIN is that the Jacobian determinant of the former is not constrained to be one. The implementation of GLOW is also included in the official implementation of GIN (Sorrenson et al. (2020)), which is also part of FrEIA.

Regarding the model evaluation, we use the mean correlation coefficient (MCC) between the ground-truth and recovered latent sources. We first compute pair-wise correlation coefficients between the true sources and recovered ones. Then we solve a linear sum assignment problem to match each recovered source to the ground truth with the highest correlation between them. MCC is a standard metric to measure the degree of identifiability up to component-wise transformation in the literature (Hyvarinen & Morioka, 2016).

Following previous works (Khemakhem et al. (2020)), the ground-truth sources ($n = 5$) are sampled from a multivariate Gaussian, with zero means and variances sampled from a uniform distribution on $[.5, 3]$[5]. The sample size for all experiments is 10000. The learning rate is 0.01 and batch size is 1000. The number of coupling layers for both GIN and GLOW is set as 24. The estimating model is built in the reverse direction as described above. The code of our experiments is in Python and developed based on GIN (Sorrenson et al. (2020)).

[4] In the official implementation of GIN, the volume-preservation is achieved in a slightly different way compared to that in its original paper for better stability of training, but there is no difference in the theoretical result w.r.t. volume-preservation.

[5] These are of the same values as previous works (Khemakhem et al. (2020); Sorrenson et al. (2020)).