On the Identifiability of Nonlinear ICA: Sparsity and Beyond

Yujia Zheng[1,2], **Ignavier Ng**[1], **Kun Zhang**[1,2]
[1] Carnegie Mellon University
[2] Mohamed bin Zayed University of Artificial Intelligence
{yujiazh, ignavierng, kunz1}@cmu.edu

Abstract

Nonlinear independent component analysis (ICA) aims to recover the underlying independent latent sources from their observable nonlinear mixtures. How to make the nonlinear ICA model identifiable up to certain trivial indeterminacies is a long-standing problem in unsupervised learning. Recent breakthroughs reformulate the standard independence assumption of sources as conditional independence given some auxiliary variables (e.g., class labels and/or domain/time indexes) as weak supervision or inductive bias. However, nonlinear ICA with unconditional priors cannot benefit from such developments. We explore an alternative path and consider only assumptions on the mixing process, such as *Structural Sparsity*. We show that under specific instantiations of such constraints, the independent latent sources can be identified from their nonlinear mixtures up to a permutation and a component-wise transformation, thus achieving nontrivial identifiability of nonlinear ICA without auxiliary variables. We provide estimation methods and validate the theoretical results experimentally. The results on image data suggest that our conditions may hold in a number of practical data generating processes.

1 Introduction

Nonlinear independent component analysis (ICA) is fundamental in unsupervised learning. It generalizes linear ICA (Comon, 1994) to identify latent sources from observations, which are assumed to be a nonlinear mixture of the sources. For an observed vector $\mathbf{x}$, nonlinear ICA expresses it as $\mathbf{x} = \mathbf{f}(\mathbf{s})$, where $\mathbf{f}$ 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 $\mathbf{f}$ as well as sources $\mathbf{s}$ from the observed mixture $\mathbf{x}$ up to certain indeterminacies. While nonlinear ICA is of general interest in a variety of tasks, such as disentanglement (Lachapelle et al., 2022) and unsupervised learning (Oja, 2002), its identifiability has been a long-standing problem for decades. 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 and Pajunen, 1999).

To deal with this challenge, existing works introduce the auxiliary variable $\mathbf{u}$ (e.g., class label, domain index, time index) and assume that sources are conditionally independent given $\mathbf{u}$ (Hyvärinen and Morioka, 2016, 2017; Hyvärinen et al., 2019; Khemakhem et al., 2020; Lachapelle et al., 2022). Most of them require the auxiliary variable to be observable, while clustering-based methods (Willetts and Paige, 2021) and those for time series (Hälvä and Hyvärinen, 2020; Hälvä et al., 2021; Yao et al., 2021) are exceptions. These works impose mild restrictions on the mixing process but require many distinct values of $\mathbf{u}$ for identifiability, which might restrict their practical applicability. Motivated by this, Yang et al. (2022) reduce the number of required distinct values by strengthening the functional assumptions on the mixing process. As described, all these results are based on conditional independence of the sources given auxiliary variable $\mathbf{u}$, instead of the standard marginal

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

independence assumption, to achieve identifiability in specific tasks, thanks to the weak supervision provided by the auxiliary variables. The identifiability of the original setting without auxiliary variables, 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. With the assumption of conformal transformation, Hyvärinen and Pajunen (1999) proved identifiability of the nonlinear ICA solution up to rotation indeterminacy for only two sources. Another result is shown for the mixture that consists of component-wise nonlinearities added to a linear mixture (Taleb and Jutten, 1999).

In this work, we aim to show the identifiability of nonlinear ICA with unconditional priors. We prove 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 only on the mixing process. Specifically, we mainly focus on the *structural sparsity* condition w.r.t. the structure from sources to observed variables, i.e., the support of the Jacobian matrix of the mixing function. Under structural conditions on the mixing process, one could identify the true sources by sparsity regularization during estimation (Thm. 1). By removing rotation indeterminacy, structural sparsity benefits the identifiability of linear ICA with Gaussian sources as well, which was previously thought to be unsolvable (Prop. 1, Thm. 2). Besides, we also develop another identifiability theory based on *Independent Influences*, which is an analogous concept of "sparsity" regarding the individual influencing process instead of the support (Prop. 2). We establish, to the best of our knowledge, one of the first identifiability results of the original nonlinear ICA model in a fully unsupervised setting.

Moreover, we also provide identifiability results for the underdetermined case, where the number of observed variables is larger than that of sources (Thm. 2, Thm. 3). This is of great practical interest and cannot be handled with previous conditions on nonlinear ICA, even with the help of auxiliary variables (Hyvärinen et al., 2019). We validate our theoretical claims experimentally, and the results on the image dataset suggest that our conditions appear to be reasonable for practical data generating processes.

2 Preliminaries

We consider the following data-generating process of ICA

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

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

where n denotes the number of latent sources, $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ denotes the observed random vector, $\mathbf{s} = (\mathbf{s}_1, \dots, \mathbf{s}_n)$ is the latent random vector representing the marginally independent sources, p_{s_i} is the marginal probability density function (PDF) of the i-th source s_i, $p_{\mathbf{s}}$ is the joint PDF of random vector $\mathbf{s}$, and $\mathbf{f} : \mathbf{s} \to \mathbf{x}$ denotes a nonlinear mixing function. For linear ICA, Eq. (2) is restricted as

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{3}$$

where $\mathbf{A} \coloneqq [\mathbf{A}_{:,1} \cdots \mathbf{A}_{:,n}]$ denotes the mixing matrix. The goal of ICA is to learn an estimated unmixing function $\hat{\mathbf{f}}^{-1} : \mathbf{x} \to \hat{\mathbf{s}}$ (or alternatively in the linear case, an estimated unmixing matrix $\hat{\mathbf{A}}^{-1}$), such that $\hat{\mathbf{s}} = (\hat{\mathbf{s}}_1, \dots, \hat{\mathbf{s}}_n)$ consists of independent estimated sources.

Notations. Throughout this work, for any matrix $\mathbf{M}$, we use $\mathbf{M}_{i,:}$ to refer to its ith row, and $\mathbf{M}_{:,j}$ to indicate its jth column. For any set of indices $\mathcal{S} \subset \{1, \dots, m\} \times \{1, \dots, n\}$, we analogously denote $\mathcal{S}_{i,:} \coloneqq \{j \mid (i,j) \in \mathcal{S}\}$ and $\mathcal{S}_{:,j} \coloneqq \{i \mid (i,j) \in \mathcal{S}\}$. We denote by $|\cdot|$ the cardinality of a set. We also define the following technical notations.

Definition 1. *Given a subset $\mathcal{S} \subseteq \{1, \dots, n\}$, the subspace $\mathbb{R}^n_{\mathcal{S}}$ is defined as*

$$\mathbb{R}^n_{\mathcal{S}} \coloneqq \{z \in \mathbb{R}^n \mid i \notin \mathcal{S} \Longrightarrow z_i = 0\}. \tag{4}$$

In other words, $\mathbb{R}^n_{\mathcal{S}}$ refers to the subspace of $\mathbb{R}^n$ indicated by an index set $\mathcal{S}$. In the following, we define the support of a matrix.

Definition 2. *The support of matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$ is defined as*

$$\operatorname{supp}(\mathbf{M}) \coloneqq \{(i,j) \mid \mathbf{M}_{i,j} \neq 0\}. \tag{5}$$

With slight abuse of notation, we also reuse the notation supp to denote the support of a matrix-valued function, depending on the context.

Definition 3. *The support of function* $\mathbf{M} : \Theta \to \mathbb{R}^{m \times n}$ *is defined as*

$$\mathrm{supp}(\mathbf{M}(\Theta)) \coloneqq \{(i,j) \mid \exists \theta \in \Theta, \mathbf{M}(\theta)_{i,j} \neq 0\}. \tag{6}$$

3 Identifiability with Structural Sparsity

3.1 Nonlinear ICA with Unconditional Priors: Structural Sparsity

In this section, we consider the identifiability of nonlinear ICA with unconditional priors, where the mixing function $\mathbf{f}$ in Eq. (2) is nonlinear. Different from recent breakthroughs in identifiable nonlinear ICA with side information (Hyvärinen et al., 2019; Lachapelle et al., 2022), in our setting, sources $\mathbf{s}$ in Eq. (1) do not need to be mutually conditionally independent given an auxiliary variable. Instead, our setting is consistent with the original ICA problem based on a general marginal independence assumption of sources. Besides, no particular form of the source distribution is assumed, which is different from most works that directly assume an exponential family (Hyvärinen et al., 2019). The goal of identifiable nonlinear ICA is to estimate the unmixing function $\hat{\mathbf{f}}^{-1}$ so that the independent sources are identified up to a certain indeterminacy, which is usually a composition of a component-wise invertible transformation and a permutation (Hyvärinen and Pajunen, 1999). We formulate the structural sparsity condition below and present the identifiability result with its proof shown in Appx. A.1. It is worth noting that Lachapelle et al. (2022) leverage sparse mechanism between latents and auxiliary variables for disentanglement, by which part of the assumptions and proof technique are inspired.

Theorem 1. *Let the observed data be sampled from a nonlinear ICA model as defined in Eqs. (1) and (2). Denote by* $\mathcal{F}$ *and* $\hat{\mathcal{F}}$ *as the support of the Jacobian* $\mathbf{J}_{\mathbf{f}}(\mathbf{s})$ *and* $\mathbf{J}_{\hat{\mathbf{f}}}(\hat{\mathbf{s}})$*, respectively. Suppose the following assumptions hold:*

i. Mixing function $\mathbf{f}$ *is invertible and smooth. Its inverse is also smooth.*

ii. For all $i \in \{1, \dots, n\}$*, there exist* $\{\mathbf{s}^{(\ell)}\}_{\ell=1}^{|\mathcal{F}_{i,:}|}$ *such that* $\mathrm{span}\{\mathbf{J}_{\mathbf{f}}(\mathbf{s}^{(\ell)})_{i,:}\}_{\ell=1}^{|\mathcal{F}_{i,:}|} = \mathbb{R}^n_{\mathcal{F}_{i,:}}$.

iii. $|\mathrm{supp}(\mathbf{J}_{\hat{\mathbf{f}}}(\hat{\mathbf{s}}))| \leq |\mathrm{supp}(\mathbf{J}_{\mathbf{f}}(\mathbf{s}))|$.

iv. <u>*(Structural Sparsity)*</u> *For all* $k \in \{1, \dots, n\}$*, there exists* $\mathcal{C}_k$ *such that*

$$\bigcap_{i \in \mathcal{C}_k} \mathrm{supp}(\mathbf{J}_{\mathbf{f}}(\mathbf{s})_{i,:}) = \{k\}.$$

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

Assumption *ii* in a generic sense rules out a set of specific parameters where the Jacobian is partially constant and can be contained in a proper affine subspace aligned by its support (i.e., $\mathbb{R}^n_{\mathcal{G}_{i,:}}$). Assumption *iii* corresponds to incorporating a suitable sparsity regularization term into the estimating process with no restriction on the ground truth. It helps to find the estimated mixing function with the minimal L_0 norm among all functions that allow the model to perfectly fit the true distribution of sources. Thus, it indicates that we should estimate the mixing process by maximizing its sparsity. Our discussion will focus on the assumption of structural sparsity, which is the core of our theory.

We start with the intuition of the proposed direction on exploiting sparsity on the mixing process for identifiability. As discussed in previous works (Hyvärinen and Pajunen, 1999), there exist infinitely many ways to preserve the independence among variables after mixing them up, such as Darmois construction (Darmois, 1951) and measure-preserving automorphism (MPA) (Hyvärinen and Pajunen, 1999). Besides, sources with non-Gaussian priors could be transferred to marginally independent Gaussians by trivial point-wise transformations (Hyvärinen et al., 2019). Thus, the existing strategy of exploiting independence and non-Gaussianity for identification fails in the nonlinear case. However, if the true structure is sparse enough, any alternative solution that produces any indeterminacy beyond component-wise transformations and permutations (e.g., mixtures or rotations) may correspond to a denser structure. In the light of that, instead of restricting the functional class (e.g., post nonlinear models (Taleb and Jutten, 1999) or conformal maps (Hyvärinen and Pajunen, 1999)) or introducing

auxiliary variables for extra structural dependencies between latent sources (Hyvärinen et al., 2019; Lachapelle et al., 2022), we leverage the sparsity pattern, i.e., the support of the Jacobian of the mixing function, to identify the sources.

The structural sparsity assumption (Assumption *iv* in Thm. 1) implies that, for every latent source $\mathbf{s}_i$, there exists a set of observed variable(s) such that $\mathbf{s}_i$ is the only latent source that participates in the generation of all observed variable(s) in the set. Graphically, for every latent source $\mathbf{s}_i$, there exists a set of observed variable(s) such that the intersection of their/its parent(s) is $\mathbf{s}_i$ (e.g., for $\mathbf{s}_1$ in Fig. 1, there exist $\mathbf{x}_1$ and $\mathbf{x}_4$ such that the intersection of their parents is $\mathbf{s}_1$.). This encourages the connections to be sparse enough to make the disentanglement of each source based on structural conditions possible. Together with the sparsity regularization during estimation (Assumption *iii*) and other mild conditions, Assumption *iv* illustrates a graphical pattern of nonlinear ICA models that could be identified by exploiting structural sparsity. It indicates a structural criterion of the sparsity that is needed for identifiability. Beyond that, structural sparsity is also of practical interest for the interpretability of results (Zhang et al., 2009) and identification of other latent variable models (Rhodes and Lee, 2021). Meanwhile, various versions of Occam's razor (e.g., faithfulness (Spirtes et al., 2000), minimality principle (Zhang, 2013), and frugality (Forster et al., 2020)) are also fundamental to the identifiability of the underlying causal structure. Readers may refer to additional discussion in Sec. 6.

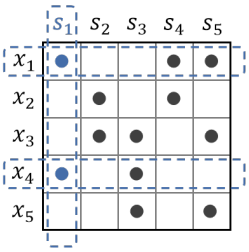


Figure 1: The structural sparsity assumption of Thm. 1, where the matrix represents $\text{supp}(\mathbf{J}_{\mathbf{f}}(\mathbf{s}))$.

With structural sparsity, Thm. 1 shows the 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. Moreover, one may modify Thm. 1 to consider the generating process based on auxiliary variables by imposing similar restrictions on the mixing process. Different from previous works (Hyvärinen et al., 2019; Yang et al., 2022), this extension removes common restrictions on the required auxiliary variables, such as the number of distinct values (e.g., class labels), non-proportional variances, or sufficient variability. Therefore, although our main result (Thm. 1) focuses on nonlinear ICA with unconditional priors, it also provides potential insight into improving the flexibility of utilizing the auxiliary variable when it is available. As a trade-off, the additional assumptions on the sparsity might limit its usage.

Besides, under additional assumptions, we could reduce the indeterminacy in Thm. 1 and further identify sources up to a component-wise *linear* transformation. Most works in nonlinear ICA focus on the identifiability up to a component-wise invertible transformation and a permutation (Hyvärinen and Morioka, 2016; Khemakhem et al., 2020), which is also called *permutation-identifiability* in the literature of disentanglement (Lachapelle et al., 2022). This indeterminacy is trivial compared to the fundamental nonuniqueness of nonlinear ICA and is analogous to the indeterminacy involving rescaling and permutation in linear ICA (Hyvärinen and Pajunen, 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 but without auxiliary variables, we show conditions for the identifiability of nonlinear ICA with unconditional priors up to a component-wise linear transformation, with the proof in Appx. A.2. The reduced indeterminacy may give rise to a more informative disentanglement and open up the possibility for further improving the quality of recovery with only element-wise linear operations, such as the Hadamard product.

Corollary 1. *Let the observed data be sampled from a nonlinear ICA model as defined in Eqs.* (1) *and* (2)*. Suppose the following assumptions hold:*

i. The function $\mathbf{h} := \hat{\mathbf{f}}^{-1} \circ \mathbf{f}$ *is a composition of a component-wise* <u>*invertible*</u> *transformation and a permutation.*

ii. The mixing function $\mathbf{f}$ *is volume-preserving.*

iii. The source distribution $p_{\mathbf{s}}(\mathbf{s})$ *is a factorial multivariate Gaussian.*

Then $\mathbf{h} := \hat{\mathbf{f}}^{-1} \circ \mathbf{f}$ *is a composition of a component-wise* <u>*linear*</u> *transformation and a permutation.*

3.2 Removing Rotation Indeterminacy with Structural Sparsity

Rotation indeterminacy is one of the major obstacles to the identifiability of ICA. Linear ICA exploits the maximization of the non-Gaussianity of the estimated sources (e.g., Kurtois (Hyvärinen and Oja, 1997)). As a direct result, the typical assumption is that at most one of the sources can have Gaussian distribution. In contrast, for the nonlinear case, a trivial point-wise function could transform sources to have any marginal distribution including Gaussian thus invalidating the effect of non-Gaussianity. Therefore, removing rotation indeterminacy remains critical for the identifiability in the nonlinear case. For instance, "rotated-Gaussian" MPA produces nonuniqueness in nonlinear ICA due to rotation indeterminacy (Hyvärinen and Pajunen, 1999). It first maps the distribution to an isotropic Gaussian, then applies a rotation and maps it back without affecting its original distribution. To address it for identifiability, Yang et al. (2022) assume that there do not exist two conditionally independent Gaussians such that their variances are proportional.

In the previous section, we have introduced the structural sparsity for the identifiability of nonlinear ICA. One may wonder whether it straightforwardly leads to the identifiability of linear ICA with Gaussian sources (Gaussian ICA), which was previously perceived to be impossible due to its rotation indeterminacy. The idea is simple–any rotation of the true mixing matrix $\mathbf{A}$ will be less sparse if $\mathbf{A}$ satisfies the proposed structural sparsity assumption. While the joint distribution of Gaussian sources stays invariant across rotations, the sparsity of the mixing matrix (i.e., L_0 norm) keeps changing. We first consider Assumption *iv* in Thm. 1 for Gaussian ICA as an extension. Because the Jacobian of linear mixing function is fixed, Assumption *ii* of Thm. 1 does not directly hold in the linear Gaussian case. As a result, we cannot directly apply Thm. 1 here and therefore propose some alternative conditions:

Proposition 1. *Let the observed data be sampled from a linear ICA model defined in Eqs.* (1) *and* (3) *with Gaussian sources. Suppose the following assumptions hold:*

i. Mixing matrix $\mathbf{A}$ *is invertible.*

ii. There exists a matrix $\hat{\mathbf{A}}$ *such that for all* $j \in \operatorname{supp}(\mathbf{A})_{i,:}$, $\operatorname{supp}(\hat{\mathbf{A}}\mathbf{A}^{-1})_{j,:} \in \mathbb{R}^n_{\operatorname{supp}(\hat{\mathbf{A}})_{i,:}}$.

iii. $|\operatorname{supp}(\hat{\mathbf{A}})| \leq |\operatorname{supp}(\mathbf{A})|$.

iv. *(Structural Sparsity) For all* $k \in \{1, \ldots, n\}$*, there exists* $\mathcal{C}_k$ *such that*

$$\bigcap_{i \in \mathcal{C}_k} \operatorname{supp}(\mathbf{A}_{i,:}) = \{k\}.$$

Then $\hat{\mathbf{A}} = \mathbf{ADP}$*, where* $\mathbf{D}$ *is a diagonal matrix and* $\mathbf{P}$ *is a column permutation matrix.*

The proof is provided in Appx. A.3. Similar to Assumption *ii* of Thm. 1, Assumption *ii* of Prop. 1 excludes a specific set of parameters that makes structure-based identification ill-posed.

The underdetermined case. For the full identifiability of nonlinear ICA, it is noteworthy that the invertibility of the mixing function is technically essential in our result (Thm. 1). This invertibility implies that the number of observed variables is equal to that of sources. Although that setting is rather common in the literature, especially for the theory of nonlinear cases (Hyvärinen and Pajunen, 1999; Hyvärinen et al., 2019), underdetermined ICA (i.e., there are more observed variables), may be of more practical interest and more challenging (Kim and Yoo, 2004). Therefore, we develop an alternative structural sparsity condition for the underdetermined case. As a trade-off, in the nonlinear case, the structural condition that removes the requirement of invertibility could only deal with the rotation indeterminacy. Thus we start from Gaussian ICA to introduce the condition.

Definition 4. *Given a matrix* $\mathbf{S} \in \mathbb{R}^{m \times n}$*, we define the function* $\operatorname{overlap} : \mathbb{R}^{m \times n} \to \{0, 1\}^{m \times n}$ *as*

$$(\operatorname{overlap}(\mathbf{S}))_{ij} = \begin{cases} 1 & \text{if } \mathbf{S}_{ij} = 1 \text{ and it is not the only nonzero entry in } \mathbf{S}_{i,:} \\ 0 & \text{otherwise.} \end{cases}$$

Theorem 2. *Let the observed data be sampled from a linear ICA model defined in Eqs.* (1) *and* (3) *with Gaussian sources. Differently, the number of observed variables (denoted as* m*) could be larger than that of the sources* n*, i.e.,* $m \geq n$*. Suppose the following assumptions hold:*

i. The nonzero coefficients of the mixing matrix $\mathbf{A}$ *are randomly drawn from a distribution that is absolutely continuous with respect to Lebesgue measure.*

ii. The estimated mixing matrix $\hat{\mathbf{A}}$ has the minimal L_0 norm during estimation.

iii. (Structural Sparsity) Given $\mathcal{C} \subseteq \{1, 2, \ldots, n\}$ where $|\mathcal{C}| > 1$, let $\mathbf{A}_{\mathcal{C}} \in \mathbb{R}^{m \times |\mathcal{C}|}$ represents a submatrix of $\mathbf{A} \in \mathbb{R}^{m \times n}$ consisting of columns with indices $\mathcal{C}$. Then, for all $k \in \mathcal{C}$, we have

$$\left| \bigcup_{k' \in \mathcal{C}} \operatorname{supp}(\mathbf{A}_{k'}) \right| - \operatorname{rank}(\operatorname{overlap}(\mathbf{A}_{\mathcal{C}})) > |\operatorname{supp}(\mathbf{A}_k)| \,.$$

Then $\hat{\mathbf{A}} = \mathbf{ADP}$ with probability one, where $\mathbf{D}$ is a diagonal matrix and $\mathbf{P}$ is a column permutation matrix.

The proof is provided in Appx. A.4. Intuitively, Assumption *iii* in Thm. 2 encourages the variability between the influences of each source. For each source, its influence on the observations should be as distinctive as possible, which is correlated with the sparsity of the mixing matrix. Fig. 2 is an illustration of the assumption. In that case, $k = 1$ and $\mathcal{C} = \{1, 2, 3\}$. Let $\mathbf{A}_{\mathcal{C}}$ denotes the submatrix shown in the figure, where $\left| \bigcup_{k' \in \mathcal{C}} \operatorname{supp}(\mathbf{A}_{k'}) \right| = 7$. Blue dotted square denotes $\operatorname{overlap}(\mathbf{A}_{\mathcal{C}})$ and we have $\operatorname{rank}(\operatorname{overlap}(\mathbf{A}_{\mathcal{C}})) = 2$. Thus, $\left| \bigcup_{k' \in \mathcal{C}} \operatorname{supp}(\mathbf{A}_{k'}) \right| - \operatorname{rank}(\operatorname{overlap}(\mathbf{A}_{\mathcal{C}})) = 5$ (the number of black dots), which is larger than $|\operatorname{supp}(\mathbf{A}_1)| = 4$. For the nonlinear case, we prove that it could help to distinguish spurious solutions due to the rotation indeterminacy, i.e., the "rotated-Gaussian" MPA (Defn. 2.5 in Gresele et al. (2021)). Gresele et al. (2021) prove that this class of MPAs could be ruled out with assumptions of conformal maps, non-Gaussianity and orthogonality of the Jacobian of the mixing function. Differently, we address it with the proposed structural sparsity condition (Assumption *iii* in Thm. 2). The corresponding theorem for the nonlinear case, with its proof given in Appx. A.5, is as follows:

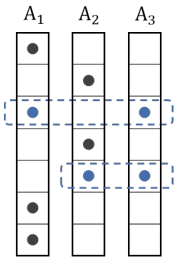


Figure 2: Assumption *iii* of Thm. 2.

Theorem 3. *Given a nonlinear ICA model defined in Eqs. (1) and (2), where $\mathbf{f}$ is the true mixing function. Consider $\hat{\mathbf{f}} = \mathbf{f} \circ \mathbf{G}^{-1} \circ \mathbf{U} \circ \mathbf{G}$, where $\mathbf{G}$ denotes an invertible Gaussianization*[1] *that maps the distribution to an standard isotropic (rotation-invariant) Gaussian, $\mathbf{U}$ denotes a rotation, and $\mathbf{G}^{-1}$ maps the distribution back to that before applying $\hat{\mathbf{U}} \circ \mathbf{G}$. If Assumptions i, ii and iii of Thm. 2 are satisfied by replacing $\mathbf{A}$ with $\mathbf{J}_{\mathbf{f}}(\mathbf{s})$ and $\hat{\mathbf{A}}$ with $\mathbf{J}_{\hat{\mathbf{f}}}(\mathbf{s})$, then function $\mathbf{h} := \hat{\mathbf{f}}^{-1} \circ \mathbf{f}$ is a composition of a component-wise invertible transformation and a permutation with probability one.*

Besides removing rotation indeterminacy, Thm. 3 also provides an extra insight for the full identifiability of nonlinear ICA, because all previous works assume that the nonlinear mixing function must be invertible (Hyvärinen and Morioka, 2016; Yang et al., 2022) even with the help of auxiliary variables.

4 Identifiability with Independent Influences

In the previous section, we have presented theoretical results based on conditions of structural sparsity. Alternatively, we show in this section that sparsity conditions on the support might not be necessary in some specific scenarios, and that one can leverage the independence (in a non-statistical sense) among the influences from different sources to the observations to achieve identifiability of nonlinear ICA. This implies sparse interactions between the individual influencing processes and could be viewed as an analogous concept of "sparsity" w.r.t. influences. We show that without any kind of auxiliary variables, the full identifiability of nonlinear ICA can also be provided given the condition of independent influences, thus complementing the results shown above from a different view of "sparsity".

Nonlinear ICA with unconditional priors: Independent Influences. We now specify the setting of the required estimating process. We sample the estimated sources $\hat{\mathbf{s}}$ from a multivariate Gaussian:

$$p_{\hat{\mathbf{s}}}(\hat{\mathbf{s}}) = \prod_{i=1}^{n} \frac{1}{Z_i} \exp\left(-\theta_{i,1}\hat{s}_i - \theta_{i,2}\hat{s}_i^2\right), \tag{7}$$

[1] One example is described in (Gresele et al., 2021), i.e., a composition of the element-wise CDFs of a smooth factorised density and a Gaussian, respectively.

where $Z_i > 0$ 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 constraint the variances σ_i^2 to be distinct without loss of generality.

Proposition 2. *Let the observed data be sampled from a nonlinear ICA model as defined in Eqs.* (1)*,* (2)*, and* (7)*. Suppose the following assumptions hold:*

i. *(Independent Influences): The influence of each source on the observed variables is independent of each other, i.e., each partial derivative $\partial \mathbf{f}/\partial s_i$ is independent of the other sources and their influences in a non-statistical sense.*

ii. *The mixing function $\mathbf{f}$ and its inverse are twice differentiable.*

iii. *The Jacobian determinant of mixing function can be factorized as $\det(\mathbf{J}_{\mathbf{f}}(\mathbf{s})) = \prod_{i=1}^{n} y_i(s_i)$, where y_i is a function that depends only on s_i. Note that volume-preserving transformation is a special case when $y_i(s_i) = 1, i = 1, \ldots, n$.*

iv. *During estimation, the columns of the Jacobian of the estimated unmixing function are regularized to be mutually orthogonal and with equal euclidean norm.*

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

The proof is included in Appx. A.6. Note that some part of the proof technique is inspired by or based on Lemma 1 in Yang et al. (2022). Assumption *iv* is a regularization during estimation and thus puts no restriction on the true generating process. We achieve it by optimizing an objective function with a designed regularization term, of which the property is as follows (proof in Appx. A.7):

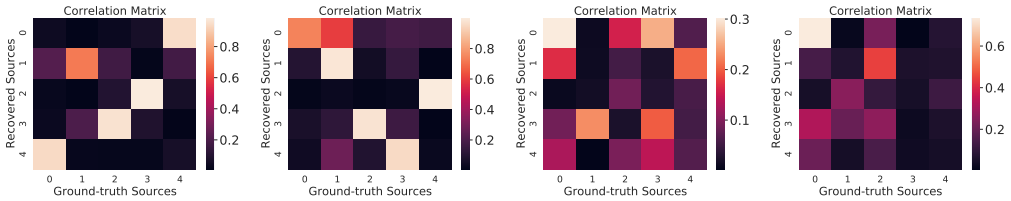
Proposition 3. *The following inequality holds*

$$n \log \left(\frac{1}{n} \sum_{i=1}^{n} \left\| \frac{\partial \hat{\mathbf{f}}^{-1}}{\partial x_i} \right\|_2 \right) - \log \left| \det(\mathbf{J}_{\hat{\mathbf{f}}^{-1}}(\mathbf{x})) \right| \geq 0, \tag{8}$$

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

Regarding the assumption of independent influences, to build intuition, one can consider a filmmaking process in a safari park, of which the main characters are wild animals. The task of the boom operators here is to record the sound of animals. As there are many animals and microphones, the mixing (recording) process is highly dependent on the relative positions between them. Thus, the animal (source s_i), loosely speaking, influences the recording (mixing function $\mathbf{f}$) through its position. The moving direction and speed of them can be loosely interpreted as the partial derivative w.r.t. the i-th source, i.e. $\partial \mathbf{f}/\partial s_i$. Assuming that the wild animals are not cooperative enough to fine-tune their positions and speeds for a better recording and the safari park is not crowded, the influences of animals on the recording are generated independently by them in the sense that they are not affected by the others while moving. As a result of that generating process, the column vectors of the Jacobian of the mixing function are uncorrelated with each other and the partial derivative w.r.t. to each source is independent of other sources. Other practical scenarios include fields that adopt independent influences in process like orthogonal coordinate transformations (Gresele et al., 2021), such as dynamic control (Mistry et al., 2010) and structural geology (De Paor, 1983). It is worth noting that this assumption of independent influences is different from the orthogonality condition proposed in (Gresele et al., 2021). For instance, if the sample mean is not zero, uncorrelatedness differs from orthogonality. And we argue that uncorrelatedness as a specific way to achieve independence might be more sensible as a condition on the influences. Besides, in (Gresele et al., 2021), orthogonality, together with conformal map and the others, are proposed as conditions to rule out two specific types of spurious solutions (i.e., Darmois construction (Darmois, 1951) and "rotated-Gaussian" MPAs (Locatello et al., 2019)) without an identifiability result, although its empirical results are encouraging.

Having said that, a violation of the assumption of independent influences 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 may lead to a high degree of multicollinearity across the columns of Jacobian, thus violating the assumption of independent influences.



(a) *SS*. (b) *II*. (c) *VP*. (d) *Base*.

Figure 3: Pearson correlation matrices between the ground-truth and the recovered sources.

Regarding Assumption *iii* in Prop. 2, 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 (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, consistent with the empirical results in Yang et al. (2022), there exists other non-volume-preserving transformations with factorial Jacobian determinant [2]. Moreover, the volume-preserving assumption has been demonstrated to be helpful to weaken the requirement of auxiliary variables, specifically by reducing the number of required labels (Sorrenson et al., 2020; Yang et al., 2022). Prop. 2 takes this further by imposing a weaker version, i.e., Assumption (iii), which, together with the other constraints, removes the need for any auxiliary variable.

5 Experiments

To validate the proposed theory of the identifiability of nonlinear ICA with unconditional priors, we conduct experiments based on our main condition, i.e., structural sparsity (Thm. 1), as well as the condition of the independent influences (Prop. 2).

Setup. For the required regularization during estimation, we consider a regularized maximum-likelihood approach with the following objective: $\mathcal{L}(\hat{\mathbf{f}}^{-1};\mathbf{x}) = \mathbb{E}_{\mathbf{x}}\left[\log p_{\hat{\mathbf{f}}^{-1}}(\mathbf{x}) - \lambda \mathbf{R}(\hat{\mathbf{f}}^{-1},\mathbf{x})\right]$, where λ is a regularization parameter and $\mathbf{R}(\hat{\mathbf{f}}^{-1},\mathbf{x})$ is the regularization term. For Thm. 1, we use L_1 norm as an approximation of the L_0 norm for efficiency (Donoho and Elad, 2003), therefore $\mathbf{R}(\hat{\mathbf{f}}^{-1},\mathbf{x}) := \|\mathbf{J}_{\hat{\mathbf{f}}^{-1}}(\mathbf{x})\|_1$; for Prop. 2, $\mathbf{R}(\hat{\mathbf{f}}^{-1},\mathbf{x})$ corresponds to LHS of Eq. (8). Following (Sorrenson et al., 2020; Yang et al., 2022), we train a GIN (Sorrenson et al., 2020) to maximize the the objective function $\mathcal{L}(\hat{\mathbf{f}}^{-1};\mathbf{x})$.

Ablation study. We conduct an ablation study to verify the necessity of the proposed assumptions. Specifically, we focus on the following models that correspond to different assumptions: *(SS)* The assumption of structural sparsity, as well as other assumptions in Thm. 1, are satisfied; *(II)* The assumption of independent influences, as well as other assumptions in Prop. 2, are satisfied; *(VP)* Compared to *II*, only the assumption of independent influences is violated while the other assumptions (e.g., factorizable Jacobian determinant) are still satisfied; *(Base)* The vanilla baseline. Compared to *VP*, the (un)mixing function is not restricted to having factorizable Jacobian determinants. The data are generated according to the required assumptions. We also conduct comparison between the assumption of independent influences and orthogonality in (Gresele et al., 2021), which are presented in Appx. B together with experimental settings. Results for each model are summarised in Fig. 4. For evaluation, we use the mean correlation coefficient (MCC) between the true sources and the estimated ones (Hyvärinen and Morioka, 2016). One could observe that when the proposed assumptions are fully satisfied (*SS* or *II*), our model achieves the highest MCC on average. This indicates that it is actually possible to identify sources from highly nonlinear mixtures up to trivial indeterminacies only based on restrictions on the mixing process. Moreover,

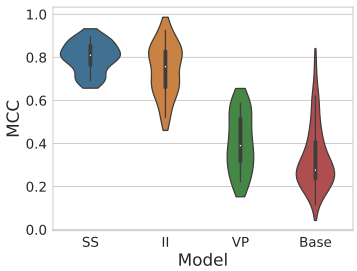


Figure 4: Ablation study.

[2] A toy example: the Jacobian determinant of mixing function w.r.t. $\left(x_1 = \frac{as_1s_2}{a+bs_2}, x_2 = \frac{bs_1}{b+as_2}\right)$ is abs_1, where $a, b \neq 0$ are some constants.

regarding Prop. 2, the importance of Assumption *iii* is supported by the higher performance of *VP* compared to that of *Base*. The visualization of the Pearson correlation matrices is shown in Fig. 3.

Stability. To study the stability of the performance of identification w.r.t. different datasets varying the number of sources n. We test the model *SS* (Thm. 1) with different n. Visually, we find that *SS* consistently outperforms *Base* (Fig. 5). Meanwhile, when the number of sources increases, one could observe that the MCC of *Base* decreases while that of *SS* stays stable. The stable empirical performance further validates our theoretical claims about the identifiability with structural sparsity.

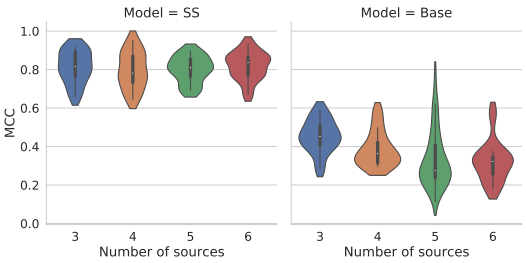


Figure 5: MCC w.r.t. different number of sources.

Image dataset. To study how reasonable the proposed theories are w.r.t. the practical generating process of observational data, we conduct experiments on the "Triangles" image dataset (Yang et al., 2022). The process of generating this dataset mimics the process of drawing triangles by humans: i) First, we sample the elements needed for humans to draw a monochrome triangle (i.e., rotation, width, height and grey level) from a factorial multivariate Gaussian distribution. Different from (Yang et al., 2022), we always sample from a single distribution in order to guarantee that all priors are unconditionally independent; ii) Then, for each pixel, we decide whether it locates inside the triangle based on the sampled elements and assign its gray level accordingly. Therefore, the process is similar to human drawing triangles. Even though each image is generated from these semantic elements, the true generating process and sources are still unknown (e.g., a pixel could be (indirectly) influenced by multiple elements in a complicated way). We apply GIN with sparsity regularization as the estimating method. The visualization of the identified sources (Fig. 6) indicates that our conditions may hold in practice.

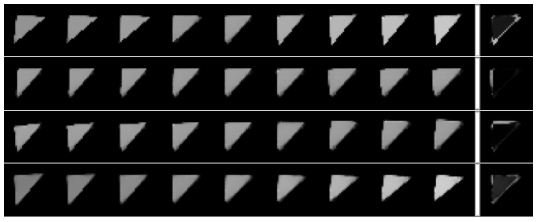


Figure 6: Identification results on Triangles. Each row represents a source identified by our model, with it varying from -2 to +2 SDs to illustrate its influence. The rightmost column is a heat map given by the absolute pixel difference between -1 and +1 SDs. Visually, the four rows correspond to rotation, width, height and gray level, respectively.

IMA. Recently, Gresele et al. (2021) assume orthogonality, conformal map and the others in order to rule out Darmois construction and "rotated-Gaussian" MPAs. They formalize the orthogonality between columns of the Jacobian of the mixing function as independent mechanism analysis (IMA) and show empirically that it helps the identification of latent sources. To further explore their exciting results, we generate datasets according to the generating process and regularization term described in (Gresele et al., 2021). We sample the sources from the same distributions described in Appx. B and use GIN for training. From Fig. 7, one could observe that both models outperform the baseline (*Base*) largely, which indicates that both conditions are empirically helpful to the identifiability of nonlinear ICA. At the same time, MCC of *II* appears to be higher than that of *IMA*, which might be the power of some additional constraints, such as the factorial volume. Meanwhile, it also supports the conjecture that IMA is a promising condition for identification, and the identifiability based on it may be achieved with some additional assumptions, such as conformal maps (Gresele et al., 2021).

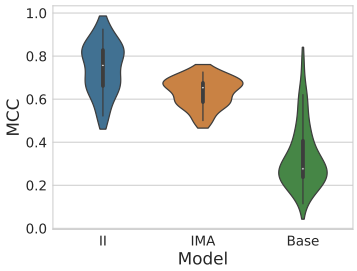


Figure 7: MCC w.r.t. independent influences (II), independent mechanism analysis (IMA), and the baseline model (Base).

6 Discussion and Conclusion

Sparsity Assumptions. Sparsity assumptions have been widely used in various fields. For latent variable models, sparsity in the generating process plays an important role in the disentanglement or identification of latent factors both empirically and theoretically (Bing et al., 2020; Rohe and Zeng, 2020; Moran et al., 2021; Rhodes and Lee, 2021; Lachapelle et al., 2022). In causality, various versions of

Occam's razor have been proposed to serve as fundamental assumptions for identifying the underlying causal structure (Spirtes et al., 2000; Zhang, 2013; Raskutti and Uhler, 2018; Forster et al., 2020).

Formulated as a measure of the density of dependencies, sparsity assumptions are more likely to be held when the observations are actually influenced by the sources in a "simple" way. For example, in biology, analyses on ecological, gene-regulatory, metabolic, and other living systems find that active interactions may often be rather sparse (Busiello et al., 2017), even when these systems evolve with an unlimited number of complicated external stimuli. In physics, it is an important heuristic that a relatively small set of laws govern complicated observed phenomena. For instance, Einstein's theory of special relativity contains parsimonious relations between substances as an important heuristic to shave away the influence of ether compared to Lorentz's theory (Einstein, 1905; Nash, 1963).

However, sparsity is not an irrefutable principle and our assumption may fail in a number of situations. The most direct one could be a scenario with heavily entangled relations between sources and observations. Let us consider the example of animal filmmaking in Sec. 4, where people are recording the sound of animals in a safari park. If the filming location is restricted to a narrow area of the safari park and multiple microphones are gathered together, the recording of each microphone will likely be influenced by almost all animals. In such a case, the dependencies between the recording of microphones and the animals are rather dense and our sparsity assumption is most likely not valid.

At the same time, even when the principle of simplicity holds, our formulation of sparsity, which is based on the sparse interactions between sources and observations, may still fail. One reason for this is the disparity between mechanism simplicity and structural sparsity. To illustrate this, one could consider the effect of sunlight on the shadow angles at the same location. In this case, the sun's rays and the shadow angles act as the sources and the observations, respectively. Because rays of sunlight, loosely speaking, may be parallel to each other, the processes of them influencing the shadow angles may be almost identical. Thus, the influencing mechanism could be rather simple. On the other hand, each shadow angle is influenced by an unlimited number of the sun's rays, which indicates that the interactions between them may not be sparse, therefore violating our assumption. This sheds light on one of the limitations of our sparsity assumption, because the principle of simplicity could be formulated in several ways. Besides, these different formulations also suggest various possibilities for identifiability based on simplicity assumptions. Another proposed assumption, i.e., independent influences, may be one of the alternative formulations, and more works remain to be explored in the future.

Conclusion. We provide identifiability results for nonlinear ICA with unconditional priors, which serve as one of the first steps to solve a long-standing problem in unsupervised learning. 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 (e.g., structural sparsity). Therefore it stays closer to the original notion of ICA that is based on the marginal independence assumption of latent sources, while previous works rely on conditional independence on auxiliary variables as weak supervision or inductive bias. Besides, by removing rotation indeterminacy, structural sparsity benefits the identifiability of Gaussian ICA as well, which was also thought to be unsolvable before. Moreover, the results on the underdetermined case are of great practical interest and introduce insight for extending identifiable nonlinear ICA to general real-world settings.

Our results on images illustrate the validity of the proposed conditions in practical data generating processes. In spite of this, it is possible that part of them is violated in several specific scenarios as discussed before. For example, the structural sparsity conditions do not apply to fully-connected structures, though the practical significance of such cases may be compromised by the lack of interpretability. As a complementary solution, we formulate the independent influences condition, which does not rely on the sparse structure of supports. While arguably natural in general, it could still be violated due to a deliberate global adjustment of sources for homogeneity. We argue that this is inevitably a trade-off between introducing auxiliary variables and imposing restrictions on the mixing process to achieve the identifiability, whose practical use depends on the scenario and information available. Future work includes further generalizing and validating our theory.

Acknowledgements

We thank the anonymous reviewers for their constructive comments. This work was partially supported by the National Institutes of Health (NIH) under Contract R01HL159805, by the NSF-Convergence Accelerator Track-D award #2134901, by a grant from Apple Inc., and by a grant from KDDI Research Inc..

References

X. Bing, F. Bunea, Y. Ning, and M. Wegkamp. Adaptive estimation in structured factor models with applications to overlapping clustering. *The Annals of Statistics*, 48(4):2055–2081, 2020.

D. M. Busiello, S. Suweis, J. Hidalgo, and A. Maritan. Explorability and the origin of network sparsity in living systems. *Scientific reports*, 7(1):1–8, 2017.

P. 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*, page 231, 1951.

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

L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using Real NVP. *arXiv preprint arXiv:1605.08803*, 2016.

D. L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via L_1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

A. Einstein. Does the inertia of a body depend upon its energy-content. *Annalen der Physik*, 18(13): 639–641, 1905.

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

M. Forster, G. Raskutti, R. Stern, and N. Weinberger. The frugal inference of causal relations. *The British Journal for the Philosophy of Science*, 2020.

L. Gresele, J. von Kügelgen, V. Stimper, B. Schölkopf, and M. Besserve. Independent mechanism analysis, a new concept? *Advances in Neural Information Processing Systems*, 2021.

H. Hälvä and A. Hyvärinen. Hidden markov nonlinear ICA: Unsupervised learning from nonstationary time series. In *Conference on Uncertainty in Artificial Intelligence*, pages 939–948. PMLR, 2020.

H. Hälvä, S. Le Corff, L. Lehéricy, J. So, Y. Zhu, E. Gassiat, and A. Hyvärinen. Disentangling identifiable features from noisy data with structured nonlinear ICA. *Advances in Neural Information Processing Systems*, 34, 2021.

A. Hyvärinen and H. Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. *Advances in Neural Information Processing Systems*, 29:3765–3773, 2016.

A. Hyvärinen and H. Morioka. Nonlinear ICA of temporally dependent stationary sources. In *International Conference on Artificial Intelligence and Statistics*, pages 460–469. PMLR, 2017.

A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural computation*, 9(7):1483–1492, 1997.

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

A. Hyvärinen, H. Sasaki, and R. Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, 2019.

I. Khemakhem, D. Kingma, R. Monti, and A. Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

S. G. Kim and C. D. Yoo. Underdetermined independent component analysis by data generation. In *International Conference on Independent Component Analysis and Signal Separation*, pages 445–452. Springer, 2004.

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

S. Lachapelle, P. R. López, Y. Sharma, K. Everett, R. L. Priol, A. Lacoste, and S. Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. *Conference on Causal Learning and Reasoning*, 2022.

F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019.

M. Mistry, J. Buchli, and S. Schaal. Inverse dynamics control of floating base systems using orthogonal decomposition. In *2010 IEEE international conference on robotics and automation*, pages 3406–3412. IEEE, 2010.

G. E. Moran, D. Sridhar, Y. Wang, and D. M. Blei. Identifiable variational autoencoders via sparse decoding. *arXiv preprint arXiv:2110.10804*, 2021.

L. K. Nash. The nature of the natural sciences. 1963.

E. Oja. Unsupervised learning in neural computation. *Theoretical computer science*, 287(1):187–207, 2002.

G. Raskutti and C. Uhler. Learning directed acyclic graph models based on sparsest permutations. *Stat*, 7(1):e183, 2018.

T. Rhodes and D. Lee. Local disentanglement in variational auto-encoders using jacobian l_1 regularization. *Advances in Neural Information Processing Systems*, 34:22708–22719, 2021.

K. Rohe and M. Zeng. Vintage factor analysis with varimax performs statistical inference. *arXiv preprint arXiv:2004.05387*, 2020.

P. Sorrenson, C. Rother, and U. Köthe. Disentanglement by nonlinear ICA with general incompressible-flow networks (GIN). *arXiv preprint arXiv:2001.04872*, 2020.

P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman. *Causation, prediction, and search*. MIT press, 2000.

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

L. N. Trefethen and D. Bau III. *Numerical linear algebra*, volume 50. Siam, 1997.

M. Willetts and B. Paige. I don't need **u**: Identifiable non-linear ICA without side information. *arXiv preprint arXiv:2106.05238*, 2021.

X. Yang, Y. Wang, J. Sun, X. Zhang, S. Zhang, Z. Li, and J. Yan. Nonlinear ICA using volume-preserving transformations. In *International Conference on Learning Representations*, 2022.

W. Yao, Y. Sun, A. Ho, C. Sun, and K. Zhang. Learning temporally causal latent processes from general temporal data. *arXiv preprint arXiv:2110.05428*, 2021.

J. Zhang. A comparison of three occam's razors for markovian causal models. *The British journal for the philosophy of science*, 64(2):423–448, 2013.

K. Zhang, H. Peng, L. Chan, and A. Hyvärinen. ICA with sparse connections: Revisited. In *International Conference on Independent Component Analysis and Signal Separation*, pages 195–202. Springer, 2009.

S. Zhang, C. Zhang, N. Kang, and Z. Li. iVPF: Numerical invertible volume preserving flow for efficient lossless compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 620–629, 2021.

A Proofs

A.1 Proof of Theorem 1

Theorem 1. *Let the observed data be sampled from a nonlinear ICA model as defined in Eqs. (1) and (2). Denote by $\mathcal{F}$ and $\hat{\mathcal{F}}$ as the support of the Jacobian $\mathbf{J}_{\mathbf{f}}(\mathbf{s})$ and $\mathbf{J}_{\hat{\mathbf{f}}}(\hat{\mathbf{s}})$, respectively. Suppose the following assumptions hold:*

i. Mixing function $\mathbf{f}$ is invertible and smooth. Its inverse is also smooth.

ii. For all $i \in \{1, \dots, n\}$, there exist $\{\mathbf{s}^{(\ell)}\}_{\ell=1}^{|\mathcal{F}_{i,:}|}$ such that $\operatorname{span}\{\mathbf{J}_{\mathbf{f}}(\mathbf{s}^{(\ell)})_{i,:}\}_{\ell=1}^{|\mathcal{F}_{i,:}|} = \mathbb{R}^n_{\mathcal{F}_{i,:}}$.

iii. $|\operatorname{supp}(\mathbf{J}_{\hat{\mathbf{f}}}(\hat{\mathbf{s}}))| \leq |\operatorname{supp}(\mathbf{J}_{\mathbf{f}}(\mathbf{s}))|$.

iv. <u>(Structural Sparsity)</u> For all $k \in \{1, \dots, n\}$, there exists $\mathcal{C}_k$ such that

$$\bigcap_{i \in \mathcal{C}_k} \operatorname{supp}(\mathbf{J}_{\mathbf{f}}(\mathbf{s})_{i,:}) = \{k\}.$$

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

Proof. Our goal here is to show that function $\mathbf{h} := \hat{\mathbf{f}}^{-1} \circ \mathbf{f}$ is a permutation with component-wise invertible transformation of sources, i.e., $\hat{\mathbf{f}} = \mathbf{f} \circ \mathbf{h}^{-1}(\mathbf{s})$. Let $\mathbf{D}(\mathbf{s})$ represents a diagonal matrix and $\mathbf{P}$ represent a permutation matrix. By using chain rule repeatedly, we write $\hat{\mathbf{f}} = \mathbf{f} \circ \mathbf{h}^{-1}(\mathbf{s})$ equivalently as

$$\begin{aligned}
\mathbf{J}_{\hat{\mathbf{f}}}(\hat{\mathbf{s}}) &= \mathbf{J}_{\mathbf{f} \circ \mathbf{h}^{-1}}(\mathbf{h}(\mathbf{s})) \\
&= \mathbf{J}_{\mathbf{f} \circ \mathbf{g}^{-1} \circ \mathbf{P}^{-1}}(\mathbf{P}\mathbf{g}(\mathbf{s})) \\
&= \mathbf{J}_{\mathbf{f} \circ \mathbf{g}^{-1}}\left(\mathbf{P}^{-1}\mathbf{P}\mathbf{g}(\mathbf{s})\right) \mathbf{J}_{\mathbf{P}^{-1}}(\mathbf{P}\mathbf{g}(\mathbf{s})) \\
&= \mathbf{J}_{\mathbf{f} \circ \mathbf{g}^{-1}}(\mathbf{g}(\mathbf{s})) \mathbf{J}_{\mathbf{P}^{-1}}(\mathbf{P}\mathbf{g}(\mathbf{s})) \\
&= \mathbf{J}_{\mathbf{f}}\left(\mathbf{g}^{-1}\mathbf{g}(\mathbf{s})\right) \mathbf{J}_{\mathbf{g}^{-1}}(\mathbf{g}(\mathbf{s})) \mathbf{J}_{\mathbf{P}^{-1}}(\mathbf{P}\mathbf{g}(\mathbf{s})) \\
&= \mathbf{J}_{\mathbf{f}}(\mathbf{s}) \mathbf{D}(\mathbf{s}) \mathbf{P},
\end{aligned} \tag{9}$$

where $\mathbf{g}$ is an invertible element-wise function. Thus our goal is equivalent to show that

$$\mathbf{J}_{\hat{\mathbf{f}}}(\hat{\mathbf{s}}) = \mathbf{J}_{\mathbf{f}}(\mathbf{s}) \mathbf{D}(\mathbf{s}) \mathbf{P}. \tag{10}$$

Because $\mathbf{J}_{\hat{\mathbf{f}}}(\hat{\mathbf{s}})$ and $\mathbf{J}_{\mathbf{f}}(\mathbf{s})$ are both invertible, we have the following equation

$$\mathbf{J}_{\hat{\mathbf{f}}}(\hat{\mathbf{s}}) = \mathbf{J}_{\mathbf{f}}(\mathbf{s}) \mathbf{T}(\mathbf{s}), \tag{11}$$

where $\mathbf{T}(\mathbf{s})$ is an invertible matrix.

For brevity of notation, we denote $\mathcal{F}$ as the support of $\mathbf{J}_{\mathbf{f}}(\mathbf{s})$, $\hat{\mathcal{F}}$ as the support of $\mathbf{J}_{\hat{\mathbf{f}}}(\hat{\mathbf{s}})$ and $\mathcal{T}$ as the support of $\mathbf{T}(\mathbf{s})$. Besides, we denote F, $\hat{\mathrm{F}}$, and T as the binary matrices with indices of ones corresponding to the supports $\mathcal{F}$, $\hat{\mathcal{F}}$, and $\mathcal{T}$, respectively. Let $b(\cdot)$ denotes the function that transfers a matrix to its binary form (i.e., setting all non-zero entries as ones). Then we have

$$b(\mathrm{F}_{i,:}\mathrm{T}) = \hat{\mathrm{F}}_{i,:}. \tag{12}$$

By Assumption ii, there exists $\{\mathbf{s}^{(\ell)}\}_{\ell=1}^{|\mathcal{F}_{i,:}|}$ such that $\operatorname{span}\{\mathbf{J}_{\mathbf{f}}(\mathbf{s}^{(\ell)})_{i,:}\}_{\ell=1}^{|\mathcal{F}_{i,:}|} = \mathbb{R}^n_{\mathcal{F}_{i,:}}$. By the definition of support (Defn. 3) and Eq. (12), we have

$$b(\mathbf{J}_{\mathbf{f}}(\mathbf{s}^{(\ell)})_{i,:}\mathrm{T}) \in \mathbb{R}^n_{\hat{\mathcal{F}}_{i,:}}, \tag{13}$$

which is equivalent to

$$\mathbf{J}_{\mathbf{f}}(\mathbf{s}^{(\ell)})_{i,:}\mathrm{T} \in \mathbb{R}^n_{\hat{\mathcal{F}}_{i,:}}. \tag{14}$$

Since $\{\mathbf{J}_{\mathbf{f}}(\mathbf{s}^{(\ell)})_{i,:}\}_{\ell=1}^{|\mathcal{F}_{i,:}|}$ forms a basis of $\mathbb{R}^n_{\mathcal{F}_{i,:}}$, for any $j_0 \in \mathcal{F}_{i,:}$, we are able to rewrite the one-hot vector $e_{j_0} \in \mathbb{R}^n_{\mathcal{F}_{i,:}}$ as

$$e_{j_0} = \sum_{\ell \in \mathcal{F}_{i,:}} \alpha_\ell \mathbf{J}_{\mathbf{f}}(\mathbf{s}^{(\ell)})_{i,:}, \tag{15}$$

where α_ℓ is corresponding coefficient. Then

$$\mathrm{T}_{j_0,:} = e_{j_0}\mathrm{T} = \sum_{\ell \in \mathcal{F}_{i,:}} \alpha_\ell \mathbf{J}_{\mathbf{f}}(\mathbf{s}^{(\ell)})_{i,:}\mathrm{T} \in \mathbb{R}^n_{\hat{\mathcal{F}}_{i,:}}, \tag{16}$$

where the final "$\in$" follows from Eq. (14) that each element in the summation belongs to $\mathbb{R}^n_{\hat{\mathcal{F}}_{i,:}}$. Thus

$$\forall j \in \mathcal{F}_{i,:},\ \mathrm{T}_{j,:} \in \mathbb{R}^n_{\hat{\mathcal{F}}_{i,:}}. \tag{17}$$

Then we are able to draw connections between these supports according to Defn. 3

$$\forall (i,j) \in \mathcal{F}, \{i\} \times \mathcal{T}_{j,:} \subset \hat{\mathcal{F}}. \tag{18}$$

According to Assumption i, $\mathbf{T}(\mathbf{s})$ is an invertible matrix, indicating that it has a non-zero determinant. Representing the determinant of the matrix $\mathbf{T}(\mathbf{s})$ as its Leibniz formula yields

$$\det(\mathbf{T}(\mathbf{s})) = \sum_{\sigma \in \mathcal{S}_n} \left(\mathrm{sgn}(\sigma) \prod_{i=1}^{n} \mathbf{T}(\mathbf{s})_{i,\sigma(i)} \right) \neq 0, \tag{19}$$

where $\mathcal{S}_n$ is the set of n-permutations. Then there exists at least one term of the sum that is non-zero, i.e.,

$$\exists \sigma \in \mathcal{S}_n,\ \forall i \in \{1,\dots,n\},\ \mathrm{sgn}(\sigma) \prod_{i=1}^{n} \mathbf{T}(\mathbf{s})_{i,\sigma(i)} \neq 0. \tag{20}$$

This is equivalent to

$$\exists \sigma \in \mathcal{S}_n,\ \forall i \in \{1,\dots,n\},\ \mathbf{T}(\mathbf{s})_{i,\sigma(i)} \neq 0. \tag{21}$$

Then we can see that this σ is in the support of $\mathbf{T}(\mathbf{s})$, which implies that

$$\forall j \in \{1,\dots,n\},\ \sigma(j) \in \mathcal{T}_{j,:}. \tag{22}$$

Together with Eq. (18), it follows that

$$\forall (i,j) \in \mathcal{F}, (i,\sigma(j)) \in \{i\} \times \mathcal{T}_{j,:} \subset \hat{\mathcal{F}}. \tag{23}$$

Denote

$$\sigma(\mathcal{F}) = \{(i,\sigma(j)) \mid (i,j) \in \mathcal{F}\}. \tag{24}$$

Then we have

$$\sigma(\mathcal{F}) \subset \hat{\mathcal{F}}. \tag{25}$$

According to Assumption iii, we can see that

$$|\hat{\mathcal{F}}| \leq |\mathcal{F}| = |\sigma(\mathcal{F})|. \tag{26}$$

Together with Eq. (25), we have

$$\sigma(\mathcal{F}) = \hat{\mathcal{F}}. \tag{27}$$

Suppose $\mathbf{T}(\mathbf{s}) \neq \mathbf{D}(\mathbf{s})\mathbf{P}$, then

$$\exists j_1 \neq j_2,\ \mathcal{T}_{j_1,:} \cap \mathcal{T}_{j_2,:} \neq \emptyset. \tag{28}$$

Besides, consider $j_3 \in \{1,\dots,n\}$ such that

$$\sigma(j_3) \in \mathcal{T}_{j_1,:} \cap \mathcal{T}_{j_2,:}. \tag{29}$$

Because $j_1 \neq j_2$, we can assume $j_3 \neq j_1$ without loss of generality. A similar strategy has previously been used in (Lachapelle et al., 2022). By Assumption iv, there exists $\mathcal{C}_{j_1} \ni j_1$ such that $\bigcap_{i \in \mathcal{C}_{j_1}} \mathcal{F}_{i,:} = \{j_1\}$. Because

$$j_3 \notin \{j_1\} = \bigcap_{i \in \mathcal{C}_{j_1}} \mathcal{F}_{i,:}, \tag{30}$$

there must exists $i_3 \in \mathcal{C}_{j_1}$ such that

$$j_3 \notin \mathcal{F}_{i_3,:}. \tag{31}$$

Because $j_1 \in \mathcal{F}_{i_3,:}$, we have $(i_3, j_1) \in \mathcal{F}$. Then according to Eq. (18), we have the following equation

$$\{i_3\} \times \mathcal{T}_{j_1,:} \subset \hat{\mathcal{F}}. \tag{32}$$

Notice that $\sigma(j_3) \in \mathcal{T}_{j_1,:} \cap \mathcal{T}_{j_2,:}$ implies

$$(i_3, \sigma(j_3)) \in \{i_3\} \times \mathcal{T}_{j_1,:}. \tag{33}$$

Then by Eqs. (33) and (32), we have

$$(i_3, \sigma(j_3)) \in \hat{\mathcal{F}}, \tag{34}$$

which implies $(i_3, j_3) \in \mathcal{F}$ by Eq. (27) and Eq. (24), therefore contradicting Eq. (31). Thus, we prove by contradiction that $\mathbf{T}(\mathbf{s}) = \mathbf{D}(\mathbf{s})\mathbf{P}$. Replacing $\mathbf{T}(\mathbf{s})$ with $\mathbf{D}(\mathbf{s})\mathbf{P}$ in Eq. (11), we prove Eq. (10), which is the goal. □

A.2 Proof of Corollary 1

Corollary 1. *Let the observed data be sampled from a nonlinear ICA model as defined in Eqs. (1) and (2). Suppose the following assumptions hold:*

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

ii. The mixing function $\mathbf{f}$ is volume-preserving.

iii. The source distribution $p_{\mathbf{s}}(\mathbf{s})$ is a factorial multivariate Gaussian.

Then $\mathbf{h} := \hat{\mathbf{f}}^{-1} \circ \mathbf{f}$ is a composition of a component-wise linear transformation and a permutation.

Proof. The proof technique of this corollary is inspired by Thm. 2 in Yang et al. (2022). According to Assumption iii, true sources $\mathbf{s}$ are from a factorial multivariate Gaussian distribution. Together with the estimated sources $\hat{\mathbf{s}}$ from the same type of distribution, we represent the densities of the true and estimated sources as

$$\begin{aligned} p_{\mathbf{s}}(\mathbf{s}) &= \prod_{i=1}^{n} \frac{1}{Z_i} \exp\left(-\theta'_{i,1} s_i - \theta'_{i,2} s_i^2\right), \\ p_{\hat{\mathbf{s}}}(\hat{\mathbf{s}}) &= \prod_{i=1}^{n} \frac{1}{Z_i} \exp\left(-\theta_{i,1} \hat{s}_i - \theta_{i,2} \hat{s}_i^2\right), \end{aligned} \tag{35}$$

where $Z_i > 0$ 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 $p_{\mathbf{s}}(\mathbf{s}) = p_{\hat{\mathbf{s}}}(\hat{\mathbf{s}}) |\det(\mathbf{J}_{\mathbf{h}}(\mathbf{s}))|$, which, by plugging in Eq. (35) and taking the logarithm on both sides, yields

$$\sum_{i=1}^{n} \log p_{s_i}(s_i) - \log |\det(\mathbf{J}_{\mathbf{h}}(\mathbf{s}))| = -\sum_{i=1}^{n} \left(\theta_{i,1} h_i(\mathbf{s}) + \theta_{i,2} h_i(\mathbf{s})^2 + \log Z_i\right). \tag{36}$$

Because of Assumption ii and the corresponding estimating process, $\det(\mathbf{J}_{\mathbf{h}}(\mathbf{s})) = 1$. Thus

$$\sum_{i=1}^{n} \log p_{s_i}(s_i) = -\sum_{i=1}^{n} \left(\theta_{i,1} h_i(\mathbf{s}) + \theta_{i,2} h_i(\mathbf{s})^2 + \log Z_i\right). \tag{37}$$

According to Assumption i, function $\mathbf{h}$ is a component-wise invertible transformation of sources, i.e.,

$$h_i(\mathbf{s}) = h_i(s_i). \tag{38}$$

Combining Eqs. (35) and (38) with Eq. (37), it follows that

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

Then we have

$$\theta'_{i,1} s_i + \theta'_{i,2} s_i^2 + \log Z_i = \theta_{i,1} h_i(s_i) + \theta_{i,2} h_i(s_i)^2.$$

Therefore, $h_i(s_i)$ is a linear function of s_i. □

A.3 Proof of Proposition 1

Proposition 1. *Let the observed data be sampled from a linear ICA model defined in Eqs.* (1) *and* (3) *with Gaussian sources. Suppose the following assumptions hold:*

i. Mixing matrix $\mathbf{A}$ *is invertible.*

ii. There exists a matrix $\hat{\mathbf{A}}$ *such that for all* $j \in \operatorname{supp}(\mathbf{A})_{i,:}$, $\operatorname{supp}(\hat{\mathbf{A}}\mathbf{A}^{-1})_{j,:} \in \mathbb{R}^n_{\operatorname{supp}(\hat{\mathbf{A}})_{i,:}}$.

iii. $|\operatorname{supp}(\hat{\mathbf{A}})| \leq |\operatorname{supp}(\mathbf{A})|$.

iv. *(Structural Sparsity) For all* $k \in \{1, \ldots, n\}$*, there exists* $\mathcal{C}_k$ *such that*

$$\bigcap_{i \in \mathcal{C}_k} \operatorname{supp}(\mathbf{A}_{i,:}) = \{k\}.$$

Then $\hat{\mathbf{A}} = \mathbf{ADP}$*, where* $\mathbf{D}$ *is a diagonal matrix and* $\mathbf{P}$ *is a column permutation matrix.*

Proof. Let $\hat{\mathbf{A}} = \mathbf{AT}$, where $\mathbf{T} \in \mathbb{R}^{n \times n}$ is an arbitrary matrix. Because our goal is to prove $\hat{\mathbf{A}} = \mathbf{ADP}$, it is equivalent to prove the following equation:

$$\mathbf{T} = \mathbf{DP}. \tag{39}$$

For brevity of notation, we denote $\mathcal{A}$ as the support of $\hat{\mathbf{A}}$, $\hat{\mathcal{A}}$ as the support of $\mathbf{A}$ and $\mathcal{T}$ as the support of $\mathbf{T}$.

According to Assumption ii, we have

$$\forall j \in \operatorname{supp}(\mathbf{A})_{i,:},\ \operatorname{supp}(\hat{\mathbf{A}}\mathbf{A}^{-1})_{j,:} \in \mathbb{R}^n_{\operatorname{supp}(\hat{A})_{i,:}}, \tag{40}$$

which is equivalent to

$$\forall j \in \mathcal{A}_{i,:},\ \mathbf{T}_{j,:} \in \mathbb{R}^n_{\hat{\mathcal{A}}_{i,:}}. \tag{41}$$

Then we have

$$\forall (i,j) \in \mathcal{A},\ \{i\} \times \mathcal{T}_{j,:} \subset \hat{\mathcal{A}}. \tag{42}$$

According to Assumption i, $\mathbf{T}$ is an invertible matrix, indicating that it has a non-zero determinant. Representing the determinant of the matrix $\mathbf{T}$ as its Leibniz formula yields

$$\det(\mathbf{T}) = \sum_{\sigma \in \mathcal{S}_n} \left(\operatorname{sgn}(\sigma) \prod_{i=1}^{n} \mathbf{T}_{i,\sigma(i)} \right) \neq 0, \tag{43}$$

where $\mathcal{S}_n$ is the set of n-permutations. Then there exists at least one term of the sum that is non-zero, i.e.,

$$\exists \sigma \in \mathcal{S}_n,\ \forall i \in \{1, \ldots, n\},\ \operatorname{sgn}(\sigma) \prod_{i=1}^{n} \mathbf{T}_{i,\sigma(i)} \neq 0. \tag{44}$$

This is equivalent to

$$\exists \sigma \in \mathcal{S}_n,\ \forall i \in \{1, \ldots, n\},\ \mathbf{T}_{i,\sigma(i)} \neq 0. \tag{45}$$

Then we can see that this σ is in the support of $\mathbf{T}$, which implies that

$$\forall j \in \{1, \ldots, n\},\ \sigma(j) \in \mathcal{T}_{j,:}. \tag{46}$$

Together with Eq. (42), it follows that

$$\forall (i,j) \in \mathcal{A}, (i, \sigma(j)) \in \{i\} \times \mathcal{T}_{j,:} \subset \hat{\mathcal{A}}. \tag{47}$$

Denote

$$\sigma(\mathcal{A}) = \{(i, \sigma(j)) \mid (i,j) \in \mathcal{A}\}. \tag{48}$$

Then we have

$$\sigma(\mathcal{A}) \subset \hat{\mathcal{A}}. \tag{49}$$

According to Assumption iii, we can see that

$$|\hat{\mathcal{A}}| \leq |\mathcal{A}| = |\sigma(\mathcal{A})|. \tag{50}$$

Together with Eq. (49), we have

$$\sigma(\mathcal{A}) = \hat{\mathcal{A}}. \tag{51}$$

Suppose $\mathbf{T} \neq \mathbf{DP}$, there must exists j_1 and j_2 such that

$$\exists j_1 \neq j_2,\ \mathcal{T}_{j_1,:} \cap \mathcal{T}_{j_2,:} \neq \emptyset. \tag{52}$$

Besides, consider $j_3 \in \{1, \ldots, n\}$ such that

$$\sigma(j_3) \in \mathcal{T}_{j_1,:} \cap \mathcal{T}_{j_2,:}. \tag{53}$$

Because $j_1 \neq j_2$, we can assume $j_3 \neq j_1$ without loss of generality. By Assumption iv, there exists $\mathcal{C}_{j_1} \ni j_1$ such that $\bigcap_{i \in \mathcal{C}_{j_1}} \mathcal{A}_{i,:} = \{j_1\}$. Because

$$j_3 \notin \{j_1\} = \bigcap_{i \in \mathcal{C}_{j_1}} \mathcal{A}_{i,:}, \tag{54}$$

there must exists $i_3 \in \mathcal{C}_{j_1}$ such that

$$j_3 \notin \mathcal{A}_{i_3,:}. \tag{55}$$

Because $j_1 \in \mathcal{A}_{i_3,:}$, we have $(i_3, j_1) \in \mathcal{A}$. Then according to Eq. (42), we have the following equation

$$\{i_3\} \times \mathcal{T}_{j_1,:} \subset \hat{\mathcal{A}}. \tag{56}$$

Notice that $\sigma(j_3) \in \mathcal{T}_{j_1,:} \cap \mathcal{T}_{j_2,:}$ implies

$$(i_3, \sigma(j_3)) \in \{i_3\} \times \mathcal{T}_{j_1,:}. \tag{57}$$

Then by Eqs. (57) and (56), we have

$$(i_3, \sigma(j_3)) \in \hat{\mathcal{A}}, \tag{58}$$

which implies $(i_3, j_3) \in \mathcal{A}$ by Eqs. (51) and (48), contradicting Eq. (55). Thus, we prove by contradiction that $\mathbf{T} = \mathbf{DP}$, which is the goal (i.e., Eq. (39)). □

A.4 Proof of Theorem 2

Theorem 2. *Let the observed data be sampled from a linear ICA model defined in Eqs. (1) and (3) with Gaussian sources. Differently, the number of observed variables (denoted as m) could be larger than that of the sources n, i.e., $m \geq n$. Suppose the following assumptions hold:*

i. The nonzero coefficients of the mixing matrix $\mathbf{A}$ are randomly drawn from a distribution that is absolutely continuous with respect to Lebesgue measure.

ii. The estimated mixing matrix $\hat{\mathbf{A}}$ has the minimal L_0 norm during estimation.

iii. (Structural Sparsity) Given $\mathcal{C} \subseteq \{1, 2, \ldots, n\}$ where $|\mathcal{C}| > 1$, let $\mathbf{A}_{\mathcal{C}} \in \mathbb{R}^{m \times |\mathcal{C}|}$ represents a submatrix of $\mathbf{A} \in \mathbb{R}^{m \times n}$ consisting of columns with indices $\mathcal{C}$. Then, for all $k \in \mathcal{C}$, we have

$$\left| \bigcup_{k' \in \mathcal{C}} \operatorname{supp}(\mathbf{A}_{k'}) \right| - \operatorname{rank}(\operatorname{overlap}(\mathbf{A}_{\mathcal{C}})) > |\operatorname{supp}(\mathbf{A}_k)|.$$

Then $\hat{\mathbf{A}} = \mathbf{ADP}$ with probability one, where $\mathbf{D}$ is a diagonal matrix and $\mathbf{P}$ is a column permutation matrix.

Proof. Because of Assumptions (ii), we consider the following combinatorial optimization

$$\hat{\mathbf{U}} \ := \ \underset{\substack{\mathbf{U} \in \mathbb{R}^{s \times s}; \\ \mathbf{U}\mathbf{U}^\top = \mathbf{I}_s}}{\arg\min} \ \|\mathbf{A}\mathbf{U}\|_0, \tag{59}$$

where $\mathbf{A}$ is the true mixing matrix and $\hat{\mathbf{U}}$ denotes the rotation matrix corresponding to the solution of the optimization problem. Let $\hat{\mathbf{A}} = \mathbf{A}\hat{\mathbf{U}}$.

Suppose $\hat{\mathbf{A}} \neq \mathbf{ADP}$, then $\hat{\mathbf{U}} \neq \mathbf{DP}$. This implies that there exists some $j' \in \{1, \ldots, s\}$ and its corresponding set of row indices $\mathcal{I}_{j'}$ ($|\mathcal{I}_{j'}| > 1$), such that $\hat{\mathbf{U}}_{i,j'} \neq 0$ for all $i \in \mathcal{I}'_j$, and

$\hat{\mathbf{U}}_{i,j'} = 0$ for all $i \notin \mathcal{I}_{j'}$. Because $\hat{U}$ is invertible and has full row rank, there exists one row index i' in $\mathcal{I}_{j'}$ that uniquely correspond to j', in order to avoid linear dependence among columns. Let $\hat{\mathbf{U}} := \left[\hat{\mathbf{U}}_1 \cdots \hat{\mathbf{U}}_s\right]$, we have

$$\left\|\hat{\mathbf{A}}_{j'}\right\|_0 = \left\|\mathbf{A}\hat{\mathbf{U}}_{j'}\right\|_0 = \left\|\sum_{i\in\mathcal{I}_{j'}} \mathbf{A}_i \hat{\mathbf{U}}_{i,j'}\right\|_0 . \tag{60}$$

Let $\mathbf{A}_{\mathcal{I}_{j'}} \in \mathbb{R}^{m\times|\mathcal{I}_{j'}|}$ represents a submatrix of $\mathbf{A}$ consisting of columns with indices $\mathcal{I}_{j'}$. Note that with a slight abuse of notation, $\mathbf{A}_i$ denotes i-th column of the matrix $\mathbf{A}$. According to Assumptions (i, iii), since $|\mathcal{I}_{j'}| > 1$ and $\hat{\mathbf{U}}_{i,j'} \neq 0$, we have

$$\left\|\sum_{i\in\mathcal{I}_{j'}} \mathbf{A}_i \hat{\mathbf{U}}_{i,j'}\right\|_0 \geq \left|\bigcup_{i\in\mathcal{I}_{j'}} \mathrm{supp}(\mathbf{A}_i)\right| - \mathrm{rank}(\mathrm{overlap}(\mathbf{A}_{\mathcal{I}_{j'}})) > |\mathrm{supp}(\mathbf{A}_{i'})| , \tag{61}$$

where $\mathrm{overlap}(\cdot)$ is defined as Defn. 4. Term $\mathrm{rank}(\mathrm{overlap}(\mathbf{A}_{\mathcal{I}_{j'}}))$ represents the maximal number of rows, in which all non-zero entries can be possibly cancelled out by the linear combination $\sum_{i\in\mathcal{I}_{j'}} \mathbf{A}_i$. Assumption i rules out a specific set of parameters that leads to a violation of that, e.g., two columns of $\mathbf{A}$ are identical in terms of element values and support. So it follows that

$$\left\|\sum_{i\in\mathcal{I}_{j'}} \mathbf{A}_i \hat{\mathbf{U}}_{i,j'}\right\|_0 > |\mathrm{supp}(\mathbf{A}_{i'})| = \|\mathbf{A}_{i'}\|_0 = \left\|\mathbf{A}_{i'}\hat{\mathbf{U}}_{i',j'}\right\|_0 . \tag{62}$$

Then we can construct $\tilde{\mathbf{U}} := \left[\tilde{\mathbf{U}}_1 \cdots \tilde{\mathbf{U}}_s\right]$. First, we set $\tilde{\mathbf{U}}_{i',j'}$ as a unique non-zero entry in column $\tilde{\mathbf{U}}_{j'}$. For simplicity, we can just set $\tilde{\mathbf{U}}_{i',j'} = 1$. For other column $\tilde{\mathbf{U}}_j$, where $j \neq j'$ and $\hat{\mathbf{U}}_{i,j} \neq 0$, we set $\tilde{\mathbf{U}}_{i,j} = 1$. Therefore

$$\begin{cases} \left\|\mathbf{A}\hat{\mathbf{U}}_j\right\|_0 > \left\|\mathbf{A}\tilde{\mathbf{U}}_j\right\|_0, & j = j', \\ \left\|\mathbf{A}\hat{\mathbf{U}}_j\right\|_0 = \left\|\mathbf{A}\tilde{\mathbf{U}}_j\right\|_0, & j \neq j'. \end{cases} \tag{63}$$

Since Assumption iii covers all columns, Eq. (62) holds for any $j' \in \{1, \ldots, s\}$. If there are multiple columns of $\hat{U}$ with more than one nonzero entry, we derive Eq. (62) for each of them. We denote the set of different target column indices j' as $\mathbf{J}$. For $j \in \mathbf{J}$, we set $\tilde{\mathbf{U}}_{i_j,j} = 1$, where i_j is the unique index of the corresponding non-zero entry in column $\hat{U}_j$. For other column $\tilde{\mathbf{U}}_j$, where $j \notin \mathbf{J}$ and $\hat{\mathbf{U}}_{i,j} \neq 0$, we set $\tilde{\mathbf{U}}_{i,j} = 1$. Then we have

$$\begin{cases} \left\|\mathbf{A}\hat{\mathbf{U}}_j\right\|_0 > \left\|\mathbf{A}\tilde{\mathbf{U}}_j\right\|_0, & j \in \mathbf{J}, \\ \left\|\mathbf{A}\hat{\mathbf{U}}_j\right\|_0 = \left\|\mathbf{A}\tilde{\mathbf{U}}_j\right\|_0, & j \notin \mathbf{J}. \end{cases} \tag{64}$$

As noted previously, every column index j corresponds to a unique row index. $\tilde{\mathbf{U}}$ is a permutation matrix and $\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top = \mathbf{I}_s$. It then follows that $\left\|\mathbf{A}\hat{\mathbf{U}}\right\|_0 > \left\|\mathbf{A}\tilde{\mathbf{U}}\right\|_0$, which contradicts the definition of $\hat{\mathbf{U}}$. □

A.5 Proof of Theorem 3

Theorem 3. *Given a nonlinear ICA model defined in Eqs. (1) and (2), where* $\mathbf{f}$ *is the true mixing function. Consider* $\hat{\mathbf{f}} = \mathbf{f} \circ \mathbf{G}^{-1} \circ \mathbf{U} \circ \mathbf{G}$*, where* $\mathbf{G}$ *denotes an invertible Gaussianization*[3] *that maps the distribution to an standard isotropic (rotation-invariant) Gaussian,* $\mathbf{U}$ *denotes a rotation, and* $\mathbf{G}^{-1}$ *maps the distribution back to that before applying* $\hat{\mathbf{U}} \circ \mathbf{G}$*. If Assumptions i, ii and iii of Thm. 2 are satisfied by replacing* $\mathbf{A}$ *with* $\mathbf{J}_{\mathbf{f}}(\mathbf{s})$ *and* $\hat{\mathbf{A}}$ *with* $\mathbf{J}_{\hat{\mathbf{f}}}(\mathbf{s})$*, then function* $\mathbf{h} := \hat{\mathbf{f}}^{-1} \circ \mathbf{f}$ *is a composition of a component-wise invertible transformation and a permutation with probability one.*

Proof. Let $\mathbf{D}(\mathbf{s})$ represents a diagonal matrix and $\mathbf{P}$ represent a permutation matrix. By using the chain rule repeatedly, we write $\hat{\mathbf{f}} = \mathbf{f} \circ \mathbf{h}(\mathbf{s})$ equivalently as

$$\begin{aligned}
\mathbf{J}_{\hat{\mathbf{f}}}(\hat{\mathbf{s}}) &= \mathbf{J}_{\mathbf{f}\circ\mathbf{h}}(\mathbf{h}(\mathbf{s})) \\
&= \mathbf{J}_{\mathbf{f}\circ\mathbf{g}^{-1}\circ\mathbf{P}^{-1}}(\mathbf{P}\mathbf{g}(\mathbf{s})) \\
&= \mathbf{J}_{\mathbf{f}\circ\mathbf{g}^{-1}}\left(\mathbf{P}^{-1}\mathbf{P}\mathbf{g}(\mathbf{s})\right)\mathbf{J}_{\mathbf{P}^{-1}}(\mathbf{P}\mathbf{g}(\mathbf{s})) \\
&= \mathbf{J}_{\mathbf{f}\circ\mathbf{g}^{-1}}(\mathbf{g}(\mathbf{s}))\mathbf{J}_{\mathbf{P}^{-1}}(\mathbf{P}\mathbf{g}(\mathbf{s})) \\
&= \mathbf{J}_{\mathbf{f}}\left(\mathbf{g}^{-1}\mathbf{g}(\mathbf{s})\right)\mathbf{J}_{\mathbf{g}^{-1}}(\mathbf{g}(\mathbf{s}))\mathbf{J}_{\mathbf{P}^{-1}}(\mathbf{P}\mathbf{g}(\mathbf{s})) \\
&= \mathbf{J}_{\mathbf{f}}(\mathbf{s})\mathbf{D}(\mathbf{s})\mathbf{P},
\end{aligned} \tag{65}$$

where $\mathbf{g}$ is an invertible element-wise function. Thus our goal is equivalent to show that

$$\mathbf{J}_{\hat{\mathbf{f}}}(\hat{\mathbf{s}}) = \mathbf{J}_{\mathbf{f}}(\mathbf{s})\mathbf{D}(\mathbf{s})\mathbf{P}. \tag{66}$$

We prove it by contrapositive.

Because $\hat{\mathbf{f}} = \mathbf{f} \circ \mathbf{G}^{-1} \circ \mathbf{U} \circ \mathbf{G}$, we write

$$\begin{aligned}
\mathbf{J}_{\hat{\mathbf{f}}}(\mathbf{s}) &= \mathbf{J}_{\mathbf{f}\circ\mathbf{G}^{-1}\circ\hat{\mathbf{U}}\circ\mathbf{G}}(\mathbf{s}) \\
&= \mathbf{J}_{\mathbf{f}\circ\mathbf{G}^{-1}\circ\hat{\mathbf{U}}}(\mathbf{G}(\mathbf{s}))\mathbf{J}_{\mathbf{G}}(\mathbf{s}) \\
&= \mathbf{J}_{\mathbf{f}\circ\mathbf{G}^{-1}}(\hat{\mathbf{U}}\mathbf{G}(\mathbf{s}))\mathbf{J}_{\hat{U}}(\mathbf{G}(\mathbf{s}))\mathbf{J}_{\mathbf{G}}(\mathbf{s}) \\
&= \mathbf{J}_{\mathbf{f}}(\mathbf{G}^{-1}\hat{\mathbf{U}}\mathbf{G}(\mathbf{s}))\mathbf{J}_{\mathbf{G}^{-1}}(\hat{\mathbf{U}}\mathbf{G}(\mathbf{s}))\mathbf{J}_{\hat{U}}(\mathbf{G}(\mathbf{s}))\mathbf{J}_{\mathbf{G}}(\mathbf{s}) \\
&= \mathbf{J}_{\mathbf{f}}(\mathbf{s})\mathbf{J}_{\mathbf{G}^{-1}}(\hat{\mathbf{U}}\mathbf{G}(\mathbf{s}))\mathbf{J}_{\hat{\mathbf{U}}}(\mathbf{G}(\mathbf{s}))\mathbf{J}_{\mathbf{G}}(\mathbf{s}) \\
&= \mathbf{J}_{\mathbf{f}}(\mathbf{s})\mathbf{D}_1(\mathbf{s})\hat{\mathbf{U}}\mathbf{D}_2(\mathbf{s}),
\end{aligned} \tag{67}$$

where we have used the chain rule repeatedly. Because $\mathbf{G}$ is an invertible element-wise transformation, $\mathbf{D}_1(\mathbf{s})$ and $\mathbf{D}_2(\mathbf{s})$ are both diagonal matrices.

Because Assumption ii of Thm. 2 is satisfied for $\mathbf{J}_{\mathbf{f}}(\mathbf{s})$, we consider the following combinatorial optimization problem

$$\hat{\mathbf{U}} \;:=\; \underset{\substack{\mathbf{U}\in\mathbb{R}^{s\times s};\\ \mathbf{U}\mathbf{U}^T=\mathbf{I}_s}}{\arg\min}\; \|\mathbf{J}_{\mathbf{f}}(\mathbf{s})\mathbf{U}\|_0. \tag{68}$$

Let $\bar{\mathbf{U}} = \mathbf{D}_1(\mathbf{s})\hat{\mathbf{U}}\mathbf{D}_2(\mathbf{s})$. Because $\mathbf{J}_{\hat{\mathbf{f}}}(\mathbf{s}) = \mathbf{J}_{\mathbf{f}}(\mathbf{s})\bar{\mathbf{U}}$, then if $\mathbf{J}(\hat{\mathbf{f}}) = \mathbf{J}_{\mathbf{f}}(\mathbf{s})\mathbf{D}(\mathbf{s})\mathbf{P}$, we have $\bar{\mathbf{U}} = \mathbf{D}(\mathbf{s})\mathbf{P}$. Thus, for every $j \in \{1, \ldots, s\}$, there exists a corresponding i', such that $\bar{\mathbf{U}}_{i',j} \neq 0$, and $\bar{\mathbf{U}}_{i,j} = 0$ for all $i \neq i'$. Because the columns of the matrix of an orthogonal transformation form an orthogonal set, columns of $\bar{\mathbf{U}}$ are linearly independent. Thus, different j cannot correspond to the same i', otherwise it is possible for these columns to be linearly dependent.

Suppose $\mathbf{J}_{\hat{\mathbf{f}}}(\mathbf{s}) \neq \mathbf{J}_{\mathbf{f}}(\mathbf{s})\mathbf{P}$, then $\bar{\mathbf{U}} \neq \mathbf{P}$. There exists $j' \in \{1, \ldots, s\}$ and its corresponding set of row indices $\mathcal{I}_{j'}$ ($|\mathcal{I}_{j'}| > 1$), such that $\bar{\mathbf{U}}_{i,j'} \neq 0$ for all $i \in \mathcal{I}'_j$, and $\bar{\mathbf{U}}_{i,j'} = 0$ for all $i \notin \mathcal{I}_{j'}$. Similarly, there exists one row index i' in $\mathcal{I}_{j'}$ that uniquely correspond to j', in order to avoid linear dependence among columns. Let $\bar{\mathbf{U}} := \left[\bar{\mathbf{U}}_1 \cdots \bar{\mathbf{U}}_s\right]$, we have

[3] One example is described in (Gresele et al., 2021), i.e., a composition of the element-wise CDFs of a smooth factorised density and a Gaussian, respectively.

$$\left\|\mathbf{J}_{\hat{\mathbf{f}}}(\mathbf{s})_{j'}\right\|_0 = \left\|\mathbf{J}_{\mathbf{f}}(\mathbf{s})\bar{\mathbf{U}}_{j'}\right\|_0 = \left\|\sum_{i\in\mathcal{I}_{j'}} \mathbf{J}_{\mathbf{f}}(\mathbf{s})_i \bar{\mathbf{U}}_{i,j'}\right\|_0 . \tag{69}$$

Let $\mathbf{J}_{\mathbf{f}}(\mathbf{s})_{\mathcal{I}_{j'}} \in \mathbb{R}^{m\times|\mathcal{I}_{j'}|}$ represents a submatrix of $\mathbf{J}_{\mathbf{f}}(\mathbf{s})$ consisting of columns with indices $\mathcal{I}_{j'}$. Note that with a slight abuse of notation, $\mathbf{J}_{\mathbf{f}}(\mathbf{s})_i$ denotes i-th column of the matrix $\mathbf{J}_{\mathbf{f}}(\mathbf{s})$. Because Assumptions (i, iii) of Thm. 2 hold for $\mathbf{J}_{\mathbf{f}}(\mathbf{s})$, with $|\mathcal{I}_{j'}| > 1$ and $\bar{\mathbf{U}}_{i,j'} \neq 0$, we have

$$\left\|\sum_{i\in\mathcal{I}_{j'}} \mathbf{J}_{\mathbf{f}}(\mathbf{s})_i \bar{\mathbf{U}}_{i,j'}\right\|_0 \geq \left|\bigcup_{i\in\mathcal{I}_{j'}} \operatorname{supp}(\mathbf{J}_{\mathbf{f}}(\mathbf{s})_i)\right| - \operatorname{rank}(\operatorname{overlap}(\mathbf{J}_{\mathbf{f}}(\mathbf{s})_{\mathcal{I}_{j'}})) > \left|\operatorname{supp}(\mathbf{J}_{\mathbf{f}}(\mathbf{s})_{i'})\right| , \tag{70}$$

where $\operatorname{overlap}(\cdot)$ is defined as Defn. 4. Note that here we slightly abuse the notation $\operatorname{overlap}(\cdot)$ to make it apply for the matrix-valued function $\mathbf{J}_{\mathbf{f}}(\mathbf{s})$. Term $\operatorname{rank}(\operatorname{overlap}(\mathbf{J}_{\mathbf{f}}(\mathbf{s})_{\mathcal{I}_{j'}}))$ generally represents the maximal number of rows, in which all non-zero entries can be cancelled out by the linear combination $\sum_{i\in\mathcal{I}_{j'}} \mathbf{J}_{\mathbf{f}}(\mathbf{s})_i$. Also with a slight abuse of the notation, Assumption i of Thm. 2 rules out a specific set of parameters that leads to a violation of that, e.g., two columns of $\mathbf{J}_{\mathbf{f}}(\mathbf{s})$ are identical in terms of element values and support. So it follows that

$$\left\|\sum_{i\in\mathcal{I}_{j'}} \mathbf{J}_{\mathbf{f}}(\mathbf{s})_i \bar{\mathbf{U}}_{i,j'}\right\|_0 > \left|\operatorname{supp}(\mathbf{J}_{\mathbf{f}}(\mathbf{s})_{i'})\right| = \left\|\mathbf{J}_{\mathbf{f}}(\mathbf{s})_{i'}\right\|_0 = \left\|\mathbf{J}_{\mathbf{f}}(\mathbf{s})_{i'} \bar{\mathbf{U}}_{i',j'}\right\|_0 . \tag{71}$$

Then we can construct $\tilde{\mathbf{U}} := \left[\tilde{\mathbf{U}}_1 \cdots \tilde{\mathbf{U}}_s\right]$. First, we set $\tilde{\mathbf{U}}_{i',j'}$ as a unique non-zero entry in column $\tilde{\mathbf{U}}_{j'}$. For simplicity, we can just set $\tilde{\mathbf{U}}_{i',j'} = 1$. For other column $\tilde{\mathbf{U}}_j$, where $j \neq j'$ and $\bar{\mathbf{U}}_{i,j} \neq 0$, we set $\tilde{\mathbf{U}}_{i,j} = 1$. Therefore

$$\begin{cases} \left\|\mathbf{J}_{\mathbf{f}}(\mathbf{s})\bar{\mathbf{U}}_j\right\|_0 > \left\|\mathbf{J}_{\mathbf{f}}(\mathbf{s})\tilde{\mathbf{U}}_j\right\|_0, & j = j', \\ \left\|\mathbf{J}_{\mathbf{f}}(\mathbf{s})\bar{\mathbf{U}}_j\right\|_0 = \left\|\mathbf{J}_{\mathbf{f}}(\mathbf{s})\tilde{\mathbf{U}}_j\right\|_0, & j \neq j'. \end{cases} \tag{72}$$

Since Assumption iii of Thm. 2 covers all columns, Eq. (71) holds for any $j' \in \{1, \ldots, s\}$. If there are multiple columns of $\bar{U}$ with more than one nonzero entry, we derive Eq. (71) for each of them. We denote the set of different target column indices j' as $\mathbf{J}$. For $j \in \mathbf{J}$, we set $\tilde{\mathbf{U}}_{i_j,j} = 1$, where i_j is the unique index of the corresponding non-zero entry in column $\hat{U}_j$. For other column $\tilde{\mathbf{U}}_j$, where $j \notin \mathbf{J}$ and $\bar{\mathbf{U}}_{i,j} \neq 0$, we set $\tilde{\mathbf{U}}_{i,j} = 1$. Then we have

$$\begin{cases} \left\|\mathbf{J}_{\mathbf{f}}(\mathbf{s})\bar{\mathbf{U}}_j\right\|_0 > \left\|\mathbf{J}_{\mathbf{f}}(\mathbf{s})\tilde{\mathbf{U}}_j\right\|_0, & j \in \mathbf{J}, \\ \left\|\mathbf{J}_{\mathbf{f}}(\mathbf{s})\bar{\mathbf{U}}_j\right\|_0 = \left\|\mathbf{J}_{\mathbf{f}}(\mathbf{s})\tilde{\mathbf{U}}_j\right\|_0, & j \notin \mathbf{J}. \end{cases} \tag{73}$$

As noted previously, every column index j corresponds to a unique row index i. $\tilde{\mathbf{U}}$ is a permutation matrix and $\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top = \mathbf{I}_s$. It then follows that $\left\|\mathbf{J}_{\mathbf{f}}(\mathbf{s})\bar{\mathbf{U}}\right\|_0 > \left\|\mathbf{J}_{\mathbf{f}}(\mathbf{s})\tilde{\mathbf{U}}\right\|_0$, which contradicts the definition of $\bar{\mathbf{U}}$. □

A.6 Proof of Proposition 2

Proposition 2. *Let the observed data be sampled from a nonlinear ICA model as defined in Eqs.* (1), (2)*, and* (7)*. Suppose the following assumptions hold:*

i. *(Independent Influences): The influence of each source on the observed variables is independent of each other, i.e., each partial derivative $\partial \mathbf{f}/\partial s_i$ is independent of the other sources and their influences in a non-statistical sense.*

ii. *The mixing function $\mathbf{f}$ and its inverse are twice differentiable.*

iii. *The Jacobian determinant of mixing function can be factorized as $\det(\mathbf{J}_{\mathbf{f}}(\mathbf{s})) = \prod_{i=1}^{n} y_i(s_i)$, where y_i is a function that depends only on s_i. Note that volume-preserving transformation is a special case when $y_i(s_i) = 1, i = 1, \ldots, n$.*

iv. *During estimation, the columns of the Jacobian of the estimated unmixing function are regularized to be mutually orthogonal and with equal euclidean norm.*

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

Proof. We first restate the setup here. Consider the following data-generating process of nonlinear ICA

$$p_{\mathbf{s}}(\mathbf{s}) = \prod_{i=1}^{n} p_{s_i}(s_i), \tag{74}$$

$$\mathbf{x} = \mathbf{f}(\mathbf{s}), \tag{75}$$

where $\mathbf{x}$ denotes the observed random vector, $\mathbf{s}$ is the latent random vector representing the marginally independent sources, p_{s_i} is the marginal PDF of s_i, $p_{\mathbf{s}}$ is the joint PDF, and $\mathbf{f} : \mathbf{s} \to \mathbf{x}$ denotes a nonlinear mixing function. We preprocess the observational data by centering the columns of the Jacobian of the mixing function. This could be achieved by applying a fixed transformation on the observational data that is equivalent to left multiplying the Jacobian of the mixing function with the centering matrix $\mathbf{U} := \mathbf{I}_n - (1/n)\mathbf{1}\mathbf{1}^\top$, where $\mathbf{I}_n$ is the identity matrix of size n and $\mathbf{1}$ is the n-dimensional vector of all ones. After centering, according to the assumption of independent influences, the columns of the Jacobian matrix are uncorrelated and of zero-means, and thus orthogonal to each other [4]. Since the preprocessing transformation $\mathbf{U}$ is fixed w.r.t. no variables other than n, which is the fixed number of dimensions, it could always be recovered after reconstruction.

We assume that the distribution of the estimated sources $\hat{\mathbf{s}}$ follows a factorial multivariate Gaussian:

$$p_{\hat{\mathbf{s}}}(\hat{\mathbf{s}}) = \prod_{i=1}^{n} \frac{1}{Z_i} \exp\left(-\theta_{i,1}\hat{s}_i - \theta_{i,2}\hat{s}_i^2\right), \tag{76}$$

where $Z_i > 0$ 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.

Our goal here is to show that function $\mathbf{h} = \hat{\mathbf{f}}^{-1} \circ \mathbf{f}$ is a composition of a permutation and a component-wise invertible transformation of sources. By using chain rule repeatedly, we write the Jacobian of function $\mathbf{h} := \hat{\mathbf{f}}^{-1} \circ \mathbf{f}$ as

$$\begin{aligned}
\mathbf{J}_{\mathbf{h}}(\mathbf{s}) &= \mathbf{J}_{\hat{\mathbf{f}}^{-1} \circ \mathbf{f}}(\mathbf{s}) \\
&= \mathbf{J}_{\hat{\mathbf{f}}^{-1}}(\mathbf{f}(\mathbf{s}))\mathbf{J}_{\mathbf{f}}(\mathbf{s}) \\
&= \mathbf{J}_{\hat{\mathbf{f}}^{-1}}(\mathbf{x})\mathbf{J}_{\mathbf{f}}(\mathbf{s}) \\
&= \hat{\mathbf{O}}(\mathbf{x})\hat{\lambda}(\mathbf{x})\mathbf{O}(\mathbf{s})\mathbf{D}(\mathbf{s}) \\
&= \hat{\lambda}(\mathbf{x})\hat{\mathbf{O}}(\mathbf{x})\mathbf{O}(\mathbf{s})\mathbf{D}(\mathbf{s}),
\end{aligned} \tag{77}$$

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 $\mathbf{J}_{\mathbf{h}}(\mathbf{s}) = \hat{\lambda}(\mathbf{x})\hat{\mathbf{O}}(\mathbf{x})\mathbf{O}(\mathbf{s})\mathbf{D}(\mathbf{s})$ by Eq. (77) and $\hat{\mathbf{O}}(\mathbf{x})\mathbf{O}(\mathbf{s})$ is an orthogonal matrix, the columns of $\mathbf{J}_{\mathbf{h}}(\mathbf{s})$ are orthogonal to each other, i.e.,

$$\mathbf{J}_{\mathbf{h}}(\mathbf{s})_{:,j}^\top \mathbf{J}_{\mathbf{h}}(\mathbf{s})_{:,k} = 0, \quad \forall j \neq k, \tag{78}$$

[4] Similarly, Gresele et al. (2021) directly formalize the notion of Independent Mechanism Analysis as orthogonality of the columns of the Jacobian.

where $\mathbf{J_h}(\mathbf{s})_{:,j}$ is the j-th column of matrix $\mathbf{J_h}(\mathbf{s})$. This can be rewritten as

$$\sum_{i=1}^{n} \left(\frac{\partial h_i}{\partial s_j} \frac{\partial h_i}{\partial s_k} \right) = 0, \quad \forall j \neq k. \tag{79}$$

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:

$$\begin{aligned} p_{\mathbf{s}}(\mathbf{s}) &= \prod_{i=1}^{n} p_{\mathbf{s}_i}(s_i), \\ p_{\hat{\mathbf{s}}}(\hat{\mathbf{s}}) &= \prod_{i=1}^{n} \frac{1}{Z_i} \exp\left(-\theta_{i,1}\hat{s}_i - \theta_{i,2}\hat{s}_i^2\right), \end{aligned} \tag{80}$$

where $Z_i > 0$ 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 $p_{\mathbf{s}}(\mathbf{s}) = p_{\hat{\mathbf{s}}}(\hat{\mathbf{s}})|\det(\mathbf{J_h}(\mathbf{s}))|$, which, by plugging in Eq. (80) and taking the logarithm on both sides, yields

$$\sum_{i=1}^{n} \log p_{s_i}(s_i) - \log|\det(\mathbf{J_h}(\mathbf{s}))| = -\sum_{i=1}^{n} \left(\theta_{i,1}h_i(\mathbf{s}) + \theta_{i,2}h_i(\mathbf{s})^2 + \log Z_i\right). \tag{81}$$

According to assumption iii, we have the factorization $\det(\mathbf{J_h}(\mathbf{s})) = \prod_{i=1}^{n} y_i(s_i)$, where y_i is a function that depends only on s_i. This implies that

$$\sum_{i=1}^{n} \left(\log p_{s_i}(s_i) - |y_i(s_i)|\right) = -\sum_{i=1}^{n} \left(\theta_{i,1}h_i(\mathbf{s}) + \theta_{i,2}h_i(\mathbf{s})^2 + \log Z_i\right). \tag{82}$$

Then, based on Lemma 1 by Yang et al. (2022), we have the following equation:

$$\sum_{i=1}^{n} \left(\frac{1}{\sigma_i^2} \frac{\partial h_i}{\partial s_j} \frac{\partial h_i}{\partial s_k} \right) = 0, \quad \forall j \neq k. \tag{83}$$

Specifically, Eq. (82) is similar to Eq. (17') and Eq. (17) in Yang et al. (2022). Eq. (83) is similar to Eq. (20) in Yang et al. (2022). The only difference is that there is no auxiliary variable in Eqs. (82) and (83), which does not influence the deriviation of Lemma 1 by Yang et al. (2022). The assumptions of the lemma are satisfied here, with the alternation that Eq. (12) in Yang et al. (2022) is weakened to Eq. (17') in Yang et al. (2022), which does not influence the deriviation according to Appx. B in Yang et al. (2022). Thus, based on Lemma 1 in (Yang et al., 2022) but without auxiliary variables, we could obtain Eq. (83).

Then, we define the following matrices:

$$\begin{aligned} \mathbf{\Lambda}_1 &= \operatorname{diag}\left(\frac{1}{\sigma_1^2}, \cdots, \frac{1}{\sigma_n^2}\right), \\ \mathbf{\Lambda}_2 &= \mathbf{I}_n, \\ \mathbf{\Sigma}(1, \mathbf{s}) &= \operatorname{diag}\left(\sum_{i=1}^{n} \frac{1}{\sigma_i^2} \left(\frac{\partial h_i}{\partial s_1}\right)^2, \cdots, \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \left(\frac{\partial h_i}{\partial s_n}\right)^2\right), \\ \mathbf{\Sigma}(2, \mathbf{s}) &= \operatorname{diag}\left(\sum_{i=1}^{n} \left(\frac{\partial h_i}{\partial s_1}\right)^2, \cdots, \sum_{i=1}^{n} \left(\frac{\partial h_i}{\partial s_n}\right)^2\right). \end{aligned} \tag{84}$$

We rewrite Eq. (83) and Eq. (79) as

$$\begin{cases} \mathbf{J_h}(\mathbf{s})^\top \mathbf{\Lambda}_1 \mathbf{J_h}(\mathbf{s}) = \mathbf{\Sigma}(1, \mathbf{s}), \\ \mathbf{J_h}(\mathbf{s})^\top \mathbf{\Lambda}_2 \mathbf{J_h}(\mathbf{s}) = \mathbf{\Sigma}(2, \mathbf{s}). \end{cases} \tag{85}$$

Because all elements in matrices defined in Eq. (84) are positive, we can take sqaure roots on both sides of equations in Eq. (85), yielding

$$\begin{cases} \mathbf{\Lambda}_1^{1/2}\mathbf{J}_{\mathbf{h}}(\mathbf{s}) = \mathbf{O}_1\mathbf{\Sigma}(1,\mathbf{s})^{1/2}, \\ \mathbf{\Lambda}_2^{1/2}\mathbf{J}_{\mathbf{h}}(\mathbf{s}) = \mathbf{O}_2\mathbf{\Sigma}(2,\mathbf{s})^{1/2}, \end{cases} \tag{86}$$

where $\mathbf{O}_1$ and $\mathbf{O}_2$ are the corresponding orthogonal matrices ($\mathbf{O}_1^\intercal\mathbf{O}_1 = \mathbf{O}_2^\intercal\mathbf{O}_2 = \mathbf{I}_n$) and $\mathbf{\Lambda}^{1/2}$ denotes the matrix by applying entry-wise square root to the entries of diagonal matrix $\mathbf{\Lambda}$. Then we rewrite Eq. (86) as

$$\begin{cases} \mathbf{J}_{\mathbf{h}}(\mathbf{s}) = \mathbf{\Lambda}_1^{-1/2}\mathbf{O}_1\mathbf{\Sigma}(1,\mathbf{s})^{1/2}, \\ \mathbf{J}_{\mathbf{h}}(\mathbf{s}) = \mathbf{\Lambda}_2^{-1/2}\mathbf{O}_2\mathbf{\Sigma}(2,\mathbf{s})^{1/2}. \end{cases} \tag{87}$$

By reformulating, we have

$$\mathbf{\Sigma}(1,\mathbf{s})^{1/2}\mathbf{\Sigma}(2,\mathbf{s})^{-1/2} = \mathbf{O}_1^{-1}\mathbf{\Lambda}_1^{1/2}\mathbf{\Lambda}_2^{-1/2}\mathbf{O}_2. \tag{88}$$

We consider both sides of Eq. (88) as a singular value decomposition of the LHS. Since $\sigma_j \neq \sigma_k, \forall j \neq k$, elements in $\mathbf{\Lambda}_1^{1/2}\mathbf{\Lambda}_2^{-1/2}$ are distinct. Therefore, non-zero elements in $\mathbf{\Lambda}_1^{1/2}\mathbf{\Lambda}_2^{-1/2}$ are also distinct because $\mathbf{\Lambda}_2$ is an identity matrix.

Because the mixing function is assumed to be invertible, the Jacobian of the mixing function $\mathbf{J}_{\mathbf{f}}(\mathbf{s}) = \mathbf{O}(\mathbf{s})\mathbf{D}(\mathbf{s})$ is of full-rank. Therefore, all diagonal elements in the diagonal matrix $\mathbf{\Lambda}_1^{1/2}\mathbf{\Lambda}_2^{-1/2}$ are non-zero.

Then, based on the uniqueness of singular value decomposition (see e.g. Proposition 4.1 in Trefethen and Bau III (1997)), $\mathbf{O}_2$ is a composition of a permutation matrix and a signature matrix. Therefore, $\mathbf{J}_{\mathbf{h}}(\mathbf{s}) = \mathbf{\Lambda}_2^{-1/2}\mathbf{O}_2\mathbf{\Sigma}(2,\mathbf{s})^{1/2}$ is a generalized permutation matrix. It thus follows that the function $\mathbf{h} := \hat{\mathbf{f}}^{-1} \circ \mathbf{f}$ is a composition of a component-wise invertible transformation and a permutation. □

A.7 Proof of Proposition 3

Proposition 3. *The following inequality holds*

$$n \log\left(\frac{1}{n}\sum_{i=1}^{n}\left\|\frac{\partial \hat{\mathbf{f}}^{-1}}{\partial x_i}\right\|_2\right) - \log\left|\det(\mathbf{J}_{\hat{\mathbf{f}}^{-1}}(\mathbf{x}))\right| \geq 0, \tag{8}$$

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

Proof. According to Hadamard's inequality, we have

$$\sum_{i=1}^{n}\log\left\|\frac{\partial \hat{\mathbf{f}}^{-1}}{\partial x_i}\right\|_2 - \log\left|\det(\mathbf{J}_{\hat{\mathbf{f}}^{-1}}(\mathbf{x}))\right| \geq 0, \tag{89}$$

with equality iff. vectors $\frac{\partial \hat{\mathbf{f}}^{-1}}{\partial x_i}, i = 1, 2, \ldots, n$ are orthogonal. Then applying the inequality of arithmetic and geometric means yields

$$\begin{aligned} & n \log\left(\frac{1}{n}\sum_{i=1}^{n}\left\|\frac{\partial \hat{\mathbf{f}}^{-1}}{\partial x_i}\right\|_2\right) - \log\left|\det(\mathbf{J}_{\hat{\mathbf{f}}^{-1}}(\mathbf{x}))\right| \\ \geq & \sum_{i=1}^{n}\log\left\|\frac{\partial \hat{\mathbf{f}}^{-1}}{\partial x_i}\right\|_2 - \log\left|\det(\mathbf{J}_{\hat{\mathbf{f}}^{-1}}(\mathbf{x}))\right| \\ \geq & 0, \end{aligned} \tag{90}$$

with the first equality iff. for all $i = 1, 2, \ldots, n$, $\left\|\frac{\partial \hat{\mathbf{f}}^{-1}}{\partial x_i}\right\|_2$ is equal. Because $\mathbf{J}_{\hat{\mathbf{f}}^{-1}}(\mathbf{x})$ is non-singular, $\frac{\partial \hat{\mathbf{f}}^{-1}}{\partial x_i}$ are orthogonal and equal to each other for all $i = 1, 2, \ldots, n$ iff. $\mathbf{J}_{\hat{\mathbf{f}}^{-1}}(\mathbf{x}) = \mathbf{O}(\mathbf{x})\lambda(\mathbf{x})$. □

B Experiments

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

SS. To guarantee that the ground-truth nonlinear mixing process satisfies the structured sparsity condition (Assumption iv in Thm. 1), we generate observed variables with "structured" multi-layer perceptrons (MLPs): Each observed variables is only a nonlinear mixture of its own parents. For example, if the observed variable $\mathbf{x}_1$ has parents $\mathbf{s}_1$ and $\mathbf{s}_2$, then $\mathbf{x}_1 = \mathrm{MLP}(\mathbf{s}_1, \mathbf{s}_2)$. The MLP here could be replaced by any nonlinear functions.

II. 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. 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. We center the columns of the Jacobian as described in Appx. A.6.

VP. Here we describe the generating process with a factorizable Jacobian determinant but not necessarily with orthogonal columns of Jacobian. We use a 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.[5] We use the official implementation of GIN (Sorrenson et al., 2020), which is part of FrEIA.

Base. Here we describe the generating process without restrictions on having a factorizable Jacobian determinant and orthogonal columns of the Jacobian. Following (Sorrenson et al., 2020), we use GLOW (Kingma and 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.

The ground-truth sources are sampled from a multivariate Gaussian, with zero means and variances sampled from a uniform distribution on $[0.5, 3]$[6]. It is worth noting that we sample sources from a single multivariate Gaussian so that all sources are marginally independent, which is different from all previous works assuming conditional independence given auxiliary variables.

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 an 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 (Hyvärinen and Morioka, 2016). All results are of 10 trials with different random seeds.

The sample size for the synthetic datasets is 10000. For experiments conducted on them, 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. Regarding the image dataset, we have 25000 32×32 images of the drawn triangle. The statistic of the dataset is described in (Yang et al., 2022), with the difference that we only use one class of triangles for unconditional priors. For experiments conducted on images, the learning rate is 3×10^{-4} and batch size is 100. The number of coupling layers for the estimating method GIN is set as 10. The experiments are directly conducted with the official implementation of GIN (Sorrenson et al., 2020) [7] with additional regularization terms and on 4 CPU cores with 16 GB RAM.

[5] 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. There is no difference in the theoretical result w.r.t. volume-preservation.

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

[7] https://github.com/VLL-HD/GIN