KINKS VS. SMOOTHNESS: IDENTIFIABILITY OF REAL ANALYTIC NICA FOR LAPLACE-LIKE SOURCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Many machine learning systems try to explain complex data - like images or financial time series - in terms of hidden, independent factors that generated them. Recovering the true underlying factors, rather than some scrambled version of them, is the central challenge of nonlinear Independent Component Analysis (nICA). We prove identifiability (exact recovery) up to trivial ambiguities for real analytic generating functions when source probability density functions have a finite number of discontinuities in the first derivative. The Laplace distribution is the most prominent example satisfying this assumption. Our proof relies on the contrast between kinks in the source distribution and the smoothness of real analytic functions. Real analytic functions comprise a broad class of generating mechanisms, and can be approximated with Normalizing Flows or Variational Autoencoders with standard activation functions (e.g., tanh, softplus, GELU), so our result applies with minimal changes to existing training pipelines. We perform experiments on real and synthetic data with both Normalizing Flows and Variational Autoencoders demonstrating their identifiability properties. In experiments on CelebA data we recover several interpretable latent factors controlling unique attributes across the dataset.

1 INTRODUCTION

Nonlinear Independent Component Analysis (nICA) has foundational theoretical implications for unsupervised and representation learning. For observations $\boldsymbol{x}^{(i)} \in \mathbb{R}^m$, it seeks to recover statistically independent generative factors (also called sources) $\boldsymbol{s}^{(i)} \in \mathbb{R}^k$, such that

$$\boldsymbol{x}^{(i)} = \boldsymbol{f}(\boldsymbol{s}^{(i)}) + \boldsymbol{\epsilon}^{(i)} \leftrightarrow \boldsymbol{g}(\boldsymbol{x}^{(i)} - \boldsymbol{\epsilon}^{(i)}) = \boldsymbol{s}^{(i)} \,, \tag{1}$$

where $\boldsymbol{f}$ is a nonlinear diffeomorphism, with its inverse $\boldsymbol{g}$, and $\boldsymbol{\epsilon}^{(i)}$ is noise (see formal definitions in section 3). When $m = k$ the noise component is generally dropped. This model is in general not able to recover the generative factors $\boldsymbol{s}^{(i)}$ because of its non-identifiability (see Definition 4 for a formal definition). One of the central goals of nICA is to guarantee the recovery of the true latent factors given a set of assumptions. These guarantees are essential for interpretable latent representations. Additionally, they are useful for training stability and reproducibility (Hyvärinen et al., 2024; Kivva et al., 2022). Last but not least, D'Amour et al. (2022) point to non-identifiability as a cause of the drop in performance of deep learning systems at deployment.

In this paper, we prove identifiability of nICA without auxiliary observations when the source distribution has a discontinuity in the first derivative (for example Laplacian) and $\boldsymbol{f}$ and $\boldsymbol{g}$ are real analytic. This is a substantial contribution to nICA theory as the assumption of analyticity is mild. As a caveat we note in subsection 4.3 that while there are several distributions on compact supports that are theoretically identifiable according to Theorem 1, they are not practically identifiable due to the universal approximation ability of real analytic diffeomorphisms on compact sets.

Neural networks can easily be constrained to be real analytic by requiring the activation functions to be real analytic. Thus training for our method which we dub RAD (Real Analytic Decoders) requires minimal changes to standard techniques, enabling easy adoption. We demonstrate the validity of the theoretical results through extensive simulation experiments. RAD also demonstrates good stability in training on real world data and recovers several interpretable latent variables for the CelebA dataset (Liu et al., 2015).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2 Related Work

While nICA is still developing, approaches to guarantee identifiability generally fall into two categories. The first utilizes auxiliary observations such as class labels, time-series indices, or the preceding observation for auto-correlated data (Khemakhem et al., 2020; Hyvarinen et al., 2019; Sorrenson et al., 2020). The sources are assumed to be conditionally independent given the auxiliary variable. Hälvä & Hyvarinen (2020) extended this using a Hidden Markov Model so that the auxiliary variables can be unobserved too.

The second approach does not assume any auxiliary information, but restricts the function class of $\boldsymbol{f}$. Zheng et al. (2022); Zheng & Zhang (2023) proved identifiability assuming that the Jacobian of $\boldsymbol{f}$ had a certain sparsity pattern. Nguyen & Fu (2025) proposed the idea that in order to recover $\boldsymbol{s}^{(i)}$, each source should control a sufficiently distinct aspect of the observations. Mathematically, this was formulated as a sufficiently scattered condition on the columns of the Jacobian of $\boldsymbol{f}$. This requires the observations to be in a much higher dimension than the sources. In a similar vein Buchholz et al. (2022) showed that if the Jacobian was a scaled orthonormal matrix at every point (a conformal map), then identifiability could be achieved. Drawing from optimal transport theory, Huang et al. (2021) trained a unique neural network to recover the sources. The parameterization ensured that $\boldsymbol{f}$ was the gradient of a convex function (its Jacobian was symmetric positive semi-definite). Yang et al. (2022) shows identifiability for volume preserving flows when there are at least two distinct but overlapping classes in the data.

In the most comparable work to ours, Kivva et al. (2022) considered real analytic mixture distributions while constraining the form of $\boldsymbol{f}$ to be piecewise affine. In several theorems they prove identifiability of increasing strength under assumptions of increasing strength. The result most comparable to Theorem 1 shows that with a prior of independent Gaussian Mixture Models (GMMs) and a piecewise affine $\boldsymbol{f}$, $\boldsymbol{s}_i$ can be recovered up to permutation, scaling, and translation. While our proof does not resemble the proof presented in Kivva et al. (2022), the assumptions are interestingly duals of each other: we assume real analytic $\boldsymbol{f}$ and non-differentiable source distribution, while they assume piecewise affine $\boldsymbol{f}$ and real analytic source distribution.

While at first glance RAD may seem similar to Sparse Autoencoders (SAE) utilized in transformer interpretability because of the use of a Laplace source distribution, there are several important distinctions. First, SAEs generally only have three layers: linear encoder, ReLU, linear decoder (Huben et al., 2024). RAD in general has many more than this. Second, the latent dimension in SAEs is higher than the observation dimension, which is not allowed in RAD. Thus SAEs form a distinct line of research.

3 Preliminaries

To begin, we introduce some preliminary definitions. We require $\boldsymbol{f}$ and $\boldsymbol{g}$ to be real analytic diffeomorphisms.

Definition 1 (Diffeomorphism (Boumal, 2023))**.** A diffeomorphism is a bijective map $\boldsymbol{f} : \mathcal{U} \to \mathcal{V}$, where $\mathcal{U} \subseteq \mathbb{R}^k$, $\mathcal{V} \subseteq \mathbb{R}^m$ are open sets such that both $\boldsymbol{f}$ and $\boldsymbol{f}^{-1} = \boldsymbol{g}$ are smooth.

Note that diffeomorphisms do not require $m = k$. If $m > k$ (the data are higher dimensional than the generative factors), we simply require the data to live in a lower dimensional embedded submanifold (see Definition 3).

Definition 2 (Real Analytic Function (Krantz & Parks, 2002))**.** A function $\boldsymbol{f} : \mathcal{U} \subseteq R^m \to R^n$ is real analytic if for every $\boldsymbol{x}_0 \in \mathcal{U}$ each component $\boldsymbol{f}_i$ admits a convergent Taylor expansion about $\boldsymbol{x}_0$ equal to $\boldsymbol{f}_i$ in some neighborhood around $\boldsymbol{x}_0$.

The set of real analytic functions C^ω is a subset of infinitely differentiable functions C^∞. Many familiar functions, such as polynomials, $\exp(\cdot)$, and $\cos(\cdot)$, are real analytic. C^ω is closed under composition and inversion. If the functions u, v are real analytic at point x_0 then $u(x) + v(x)$ and $u(x)v(x)$ are real analytic at x_0. Additionally, $\frac{1}{u(x)}$ is real analytic at x_0 if $u(x_0) \neq 0$.

As mentioned earlier, in the case when $m > k$ we adopt the assumption that the data approximately lie on a low-dimensional submanifold embedded in a higher dimension. This is the standard assumption of representation learning methods, such as VAEs, when the embedded dimension is lower than

the observation dimension (Goodfellow et al., 2016). Because $\boldsymbol{f}$ is real analytic this submanifold must be real analytic also.

Definition 3 (Real Analytic Submanifold (Krantz & Parks, 2002))**.** The set $\mathcal{M} \subseteq \mathbb{R}^m$ is a k-dimensional real analytic submanifold if for each $p \in \mathcal{M}$ there exist an open $\mathcal{V}$ with $p \in \mathcal{V} \subseteq \mathbb{R}^m$, a convex open $\mathcal{U} \subseteq \mathbb{R}^k$, and real analytic maps $\boldsymbol{g} : \mathcal{V} \to \mathcal{U}$, $\boldsymbol{f} : \mathcal{U} \to \mathcal{V}$ such that $\mathcal{M} \cap \mathcal{V} = \text{image } \boldsymbol{f}$ and $\boldsymbol{g} \circ \boldsymbol{f}$ is the identity on $\mathcal{U}$.

This assumption implies that there is redundant information in the observations. For example, in an image each pixel value is not free to take on any value as it must be similar to those around it for the image to convey meaningful information. This is why compression is possible. Because in practice $\boldsymbol{x}^{(i)}$ almost never exactly lie on a submanifold, it is assumed that they are noisy. We define $p_{\boldsymbol{f},\boldsymbol{g}}(\boldsymbol{x})$ as the probability density of $\boldsymbol{x}$ given the model in Equation 1 parameterized by $\boldsymbol{f}$ and $\boldsymbol{g}$ with known source distribution ρ. We provide the standard definition of identifiability up to an ambiguity class below.

Definition 4 (Identifiability)**.** The model in Equation 1 is said to be identifiable up to equivalence class $\simeq$ if

$$\forall \boldsymbol{x} \in \mathbb{R}^m, p_{\tilde{\boldsymbol{f}},\tilde{\boldsymbol{g}}}(\boldsymbol{x}) = p_{\boldsymbol{f},\boldsymbol{g}}(\boldsymbol{x}) \implies \boldsymbol{g} \simeq \tilde{\boldsymbol{g}}.$$

4 Main Result

In this section we present the identifiability proof. First, using standard techniques, we transform the model in Equation 1 to be in terms of learning a self-map of the sources $\boldsymbol{h} = \tilde{\boldsymbol{g}} \circ \boldsymbol{f}$. This step also eliminates the noise from $\boldsymbol{\epsilon}$ in the case of $m > k$ and requires Assumption 1, which comes directly from Khemakhem et al. (2020). Then in Theorem 1 we prove that $\boldsymbol{h}$ must be a signed permutation (scaling and translation are embedded in the knowledge of the source distribution).

Assumption 1 ((Khemakhem et al., 2020))**.** The set $\{\boldsymbol{x} \in \mathbb{R}^m | \varphi_\epsilon(\boldsymbol{x}) = 0\}$ has measure zero, where φ_ϵ is the characteristic function of the density of ϵ.

Lemma 1. *Suppose that ϵ satisfies Assumption 1 and we learn real analytic $\tilde{\boldsymbol{f}}$ and $\tilde{\boldsymbol{g}}$ such that $\tilde{\boldsymbol{g}} \circ \tilde{\boldsymbol{f}}$ is the identity and $\forall \boldsymbol{x} \in \mathbb{R}^m, p_{\tilde{\boldsymbol{f}},\tilde{\boldsymbol{g}}}(\boldsymbol{x}) = p_{\boldsymbol{f},\boldsymbol{g}}(\boldsymbol{x})$. Then $\boldsymbol{h} = \tilde{\boldsymbol{g}} \circ \boldsymbol{f}$ is a real analytic diffeomorphism and $\boldsymbol{h}(\boldsymbol{s}) \sim \rho$.*

Proof. The first part of this proof is relatively trivial. From Definition 2.7.4 of (Krantz & Parks, 2002) it is clear that the composition $\tilde{\boldsymbol{g}} \circ \boldsymbol{f}$ is real analytic in $\mathbb{R}^k$. It is also clear from Definition 1 that the composition of diffeomorphisms is a diffeomorphism.

The second part removes the noise $\boldsymbol{\epsilon}$. Using Assumption 1, Khemakhem et al. (2020) provided the standard proof for this in Appendix B.2.2 Step I using a deconvolution argument. □

We utilize the contrast between real analytic generating functions and non-differentiable source distributions to prove identifiability. The non-differentiability of the source distribution ρ must satisfy the following assumption.

Assumption 2. ρ is a probability density function with connected support $\mathcal{D} \subseteq \mathbb{R}$. $\log \rho$ can be partitioned into a finite number of real analytic functions $\alpha_0, \ldots, \alpha_d$ at distinct breakpoints $\beta_1, \ldots, \beta_d$ for $d \geq 1$. The following must hold,

1. $\alpha'_{i-1}(\beta_i) \neq \alpha'_i(\beta_i)\ \forall i = 1, \ldots, d$
2. $\beta_i - \beta_{i-1} > 0\ \forall i = 1, \ldots, d+1$ with the convention that $\beta_0 = \inf \mathcal{D}$ and $\beta_{d+1} = \sup \mathcal{D}$
3. For d odd, there are analytic continuations of $\alpha_{\frac{d-1}{2}}$ and $\alpha_{\frac{d+1}{2}}$ on an open neighborhood around $\beta_{\frac{d+1}{2}}$. Similarly for d even, there are analytic continuations of $\alpha_{\frac{d}{2}-1}$ and $\alpha_{\frac{d}{2}+1}$ onto the interval $[\beta_{\frac{d}{2}}, \beta_{\frac{d}{2}+1}]$.

The last requirement in Assumption 2 is used in step 3 of the proof of Theorem 1 in order to use the Taylor expansion around a common point. Note that although distributions of the form $\rho(x) \propto \exp(-|x|^\alpha), \alpha \in (0, 1)$ satisfy requirements 1 and 2 of Assumption 2, they fail on requirement 3.

Several well known distributions satisfy this assumption including the Laplace (also called double exponential) distribution and the triangular distribution. The probability density function of the Laplace distribution is given by

$$\rho_{\mathcal{L}}(s) = \frac{1}{2b} \exp\left(-\frac{1}{b}|s-\mu|\right),$$

for location and scale parameters μ and b. The Laplace is well known because its maximum likelihood estimate uses the ℓ_1 norm penalty which induces sparsity while allowing outliers. Derivatives of the Laplace distribution like Laplace Mixture Model, Log Laplace, and asymmetric Laplace also satisfy Assumption 2. A Spike-and-Slab distribution built from Laplace distributions to induce sparsity as used by Moran et al. (2022) also satisfies the assumption.

The standard triangular distribution describes the probability density of the sum of two standard uniform random variables. However, in its general form it can also be asymmetric,

$$\rho_{\Delta}(x) = \begin{cases} 0, & x < a \\ \frac{2(x-a)}{(b-a)(c-a)}, & a \leq x \leq c \\ \frac{2(b-x)}{(b-a)(b-c)}, & c \leq x \leq b \\ 0, & b < x \end{cases}$$

While this distribution satisfies Assumption 2, it may be difficult to use in practice (see subsection 4.3). Exponential of many piecewise defined polynomials and other non-standard distributions also satisfy Assumption 2. The essential characteristic is the discontinuity in the derivative of the probability density function which implies a sharp change in the probability law of the source. Larger discontinuities are harder to approximate with real analytic functions and thus may make recovery easier. We now present our main result.

Theorem 1. *Suppose $\boldsymbol{h}$ is a real analytic diffeomorphism, $\boldsymbol{s}$ is iid according to ρ, and $\boldsymbol{h}(\boldsymbol{s})$ is also iid according to ρ. If ρ satisfies Assumption 2, then $\boldsymbol{h}$ must be the identity up to signed permutation and translation.*

Proof. The proof proceeds in three steps. In the first step we show that $\sum_i \log \rho(\boldsymbol{s}_i) - \sum_i \log \rho(\boldsymbol{h}_i(\boldsymbol{s}))$ must be real analytic. In the second step, we show that because the kinks in $\log \rho(\boldsymbol{s}_i)$ must be canceled out by the kinks in $\log \rho(\boldsymbol{h}_{\pi_1(i)}(\boldsymbol{s}))$ we can factorize $\boldsymbol{h}_{\pi_1(i)}$ as in Equation 2. Finally in step three, using Taylor expansion we show that $r_i(\boldsymbol{s}) = \pm 1$ in Equation 2.

Because $\boldsymbol{s} \sim \rho$ and $\boldsymbol{h}(\boldsymbol{s}) \sim \rho$, the change of variables formula gives

$$\prod_i \rho(\boldsymbol{s}_i) = |\det \boldsymbol{J}_h(\boldsymbol{s})| \prod_i \rho(\boldsymbol{h}_i(\boldsymbol{s}))$$

$$\sum_i \log \rho(\boldsymbol{s}_i) - \sum_i \log \rho(\boldsymbol{h}_i(\boldsymbol{s})) = \log |\det \boldsymbol{J}_h(\boldsymbol{s})|$$

Because $\boldsymbol{h}$ is invertible $\det \boldsymbol{J}_h(\boldsymbol{s}) \neq 0$. Further due to the continuity of $\det \boldsymbol{J}_h(\boldsymbol{s})$ it always has the same sign. Therefore, the absolute value reduces to a constant multiple. Since $\boldsymbol{h}$ is real analytic, $\boldsymbol{J}_h(\boldsymbol{s})$ is also real analytic. The determinant is a polynomial function, which is real analytic, as is the logarithm. Therefore $\log \det \boldsymbol{J}_h(\boldsymbol{s})$ is real analytic. This implies that $\gamma(\boldsymbol{s}) = \sum_i \log \rho(\boldsymbol{s}_i) - \sum_i \log \rho(\boldsymbol{h}_i(\boldsymbol{s}))$ must also be real analytic.

Let $\mathcal{A}_{ij} = \{\boldsymbol{s} | \boldsymbol{s}_i = \beta_j\}$ and $\mathcal{B}_{lm} = \{\boldsymbol{s} | \boldsymbol{h}_l(\boldsymbol{s}) = \beta_m\}$. At $\mathcal{A}_{ij}$ there is a non-differentiable point in $\log \rho(\boldsymbol{s})$ as you vary $\boldsymbol{s}_i$. It must be canceled out by $\sum_i \log \rho(\boldsymbol{h}_i(\boldsymbol{s}))$ to maintain the real analytic property. Thus $\mathcal{A}_{ij} \subset \bigcup_{l,m} \mathcal{B}_{lm}$ because there can only be non-analytic points in $\sum_i \log \rho(\boldsymbol{h}_i(\boldsymbol{s}))$ around $\bigcup_{l,m} \mathcal{B}_{lm}$ because $\boldsymbol{h}$ is real analytic.

Suppose that $\nexists\, l, m$, such that $\boldsymbol{h}_l \equiv \beta_m$ on an open neighborhood of $\boldsymbol{s} \in \mathcal{A}_{ij}$. For real analytic $\boldsymbol{h}_l - \beta_m$ the measure of the zero-set is zero (Mityagin, 2015). Thus $\mathcal{B}_{lm} \cap \mathcal{A}_{ij}$ has a measure of zero. But if all $\mathcal{B}_{lm}$ have measure zero in $\mathcal{A}_{ij}$ then we reach a contradiction in the measure of $\mathcal{A}_{ij}$. So $\exists\, l, m$ such that $\boldsymbol{h}_l \equiv \beta_m$ on an open neighborhood of $\mathcal{A}_{ij}$. By the real analytic identity theorem (Corollary 1.2.6 in (Krantz & Parks, 2002)) $\boldsymbol{h}_l \equiv \beta_m$ on all of $\mathcal{A}_{ij}$.

Because $\boldsymbol{h}_l$ is constant on $\mathcal{A}_{ij}$ the directional derivatives $\boldsymbol{v}^\top \nabla \boldsymbol{h}_l(\boldsymbol{s}) = 0, \forall \boldsymbol{s} \in \mathcal{A}_{ij}, \boldsymbol{v}_i = 0$ which implies $\nabla \boldsymbol{h}_l(\boldsymbol{s}) \in \text{span}\{\boldsymbol{e}_i\}$. Suppose that both $\boldsymbol{h}_l - \beta_j$ and $\boldsymbol{h}_k - \beta_j$ vanish on $\mathcal{A}_{ij}$, then $\nabla \boldsymbol{h}_l(\boldsymbol{s}), \nabla \boldsymbol{h}_k(\boldsymbol{s}) \in \text{span}\{\boldsymbol{e}_i\}, \forall \boldsymbol{s} \in \mathcal{A}_{ij}$. However, this contradicts the inverse function theorem which says that the Jacobian of $\boldsymbol{h}$ is invertible. Therefore $\mathcal{A}_{ij} = \mathcal{B}_{\pi(i,j)}$ for permutations π.

π can be further simplified. For $j \neq j'$, $\varnothing = \mathcal{A}_{ij} \cap \mathcal{A}_{ij'} = \mathcal{B}_{\pi(i,j)} \cap \mathcal{B}_{\pi(i,j')}$. However, $\mathcal{B}_{lm} \cap \mathcal{B}_{l'm'} = \varnothing$ only when $l = l'$ and $m' \neq m$. Therefore, we can factorize $\pi(i,j)$ into $\pi_1(i), \pi_2(j)$. This means that all kinks in $\rho(\boldsymbol{s}_i)$ are canceled by a single dimension of $\boldsymbol{h}$. For simplicity we decompose $\boldsymbol{s}$ into $u = \boldsymbol{s}_i$ and $\boldsymbol{v}$ be the rest of $\boldsymbol{s}$. For a fixed $\boldsymbol{v}$, $\boldsymbol{h}$ must pass through each kink exactly once or cancellation will not be achieved. Therefore because of continuity in $\boldsymbol{h}_{\pi_1(i)}(u, \boldsymbol{v})$ with respect to u π_2 can be reduced to the identity or the reverse order permutation.

Because $\boldsymbol{h}_{\pi_1(i)} - \beta_{\pi_2(j)}$ vanishes precisely when $u - \beta_j$ vanishes and $\nabla \boldsymbol{h}_{\pi_1(i)} \neq \boldsymbol{0}$, we can use the Weierstrass Preparation Theorem (Theorem 6.1.3 (Krantz & Parks, 2002)) to factorize $\boldsymbol{h}_{\pi_1(i)} - \beta_{\pi_2(j)}$

$$\boldsymbol{h}_{\pi_1(i)}(u, \boldsymbol{v}) - \beta_{\pi_2(j)} = r_i(u, \boldsymbol{v})(u - \beta_j) \tag{2}$$

where $r_i(u, \boldsymbol{v}) \neq 0$ and is real analytic. Note here the Weierstrass polynomial is of degree 1 because $\nabla \boldsymbol{h}_i \neq \boldsymbol{0}$.

Step 3 utilizes the uniqueness of the Taylor coefficients of $\gamma_i(u) = \log \rho(\boldsymbol{h}_{\pi_1(i)}(u, \boldsymbol{v})) - \log \rho(u)$ by expanding multiple segments of γ_i around a shared center point to prove that $r_i(\boldsymbol{s}) \equiv \pm 1$. Due to space constraints, we leave the full details for Appendix A.

Using $r_i \equiv \pm 1$ in Equation 2, we see that $\boldsymbol{h}$ is a signed permutation with a potential translation.

□

4.1 Model Architecture

To learn real analytic diffeomorphisms we consider two well known neural network architectures: normalizing flows (Kobyzev et al., 2021) and variational autoencoders (Kingma & Welling, 2019). Since affine transformations are real analytic, constraining a neural network to be real analytic only requires the activation functions to also be real analytic.

Normalizing Flows require invertible activation functions, which means that they must be monotonic. We classify several popular monotonic activation functions shown in Figure 1 by their membership in C^ω. The activation functions below are monotonic and real analytic,

1. $\text{Tanh}(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$
2. $\text{Sigmoid}(x) = \frac{1}{1+e^{-x}}$
3. $\text{Softplus}(x) = \log(1 + e^x)$

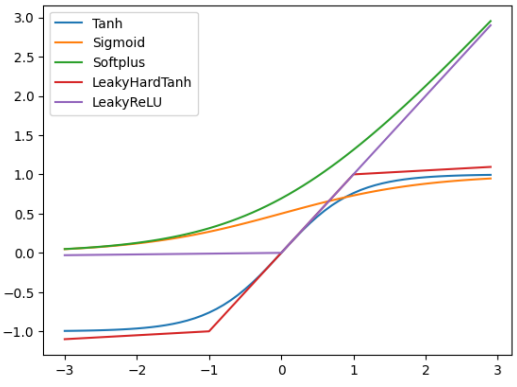


Figure 1: Common monotonic activation functions

The following are monotonic but not real analytic,

1. $\text{LeakyReLU}(x) = \begin{cases} x, & x \geq 0 \\ \alpha x, & x < 0 \end{cases}$
2. $\text{LeakyHardTanh}(x) = \begin{cases} \alpha x - 1, & x < -1 \\ x, & |x| \leq 1 \\ \beta x + 1, & x > 1 \end{cases}$

VAEs do not require invertible activation functions and so may also use more modern alternatives like Swish (Ramachandran et al., 2017) or GELU (Hendrycks & Gimpel, 2023), both of which are real analytic.

4.2 Assumption Tightness

In this section we show that both assumptions of real analyticity and non-differentiable source distributions are necessary for the identifiability in Theorem 1. Some of the most well known nICA identifiability counter-examples are the so called Measure Preserving Automorphisms (MPAs) (Hyvärinen & Pajunen, 1999). For any continuous source distribution ρ with mild regularity assumptions there exists a transformation $\boldsymbol{q}$ such that $\boldsymbol{q}(\boldsymbol{s}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. This $\boldsymbol{q}$ can be constructed by $q = \Phi^{-1} \circ R$ where Φ is the Gaussian Cumulative Distribution Function (CDF) and R is the CDF of ρ. Since the normal distribution is invariant under orthogonal transformation, $\boldsymbol{Q}\boldsymbol{q}(\boldsymbol{s}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Finally the sources can be transformed back to their original distribution so that $\boldsymbol{q}^{-1}(\boldsymbol{Q}\boldsymbol{q}(\boldsymbol{s}))$ has the same distribution as $\boldsymbol{s}$.

This provides a clear counter-example when we relax the requirement of $\boldsymbol{h}$ to be real analytic. However, $\boldsymbol{q}$ cannot be real analytic if ρ is not real analytic because R would not be real analytic. MPAs also provide a counter-example when $\boldsymbol{s}$ is distributed according to a real analytic distribution. This includes obvious distributions like Gaussian, but also when there is a discontinuity in the PDF between the 0 and non-zero parts such as for the Gamma or Beta distributions.

4.3 Limitations

The Whitney Approximation Theorem (Krantz & Parks, 2002) states that real analytic functions can approximate any C^r function for any $r \geq 0$ to arbitrary precision on a compact set (closed and bounded in euclidean space). Further these approximation results can be attained by neural networks. Ishikawa et al. (2023) provides general results on the universal approximation ability of invertible neural networks (Normalizing Flows). With our architecture we impose the constraint that $\boldsymbol{f} \in C^\omega \subset C^\infty$. So Ishikawa et al. (2023) demonstrates that our method is a universal function and universal density approximator. Puthawala et al. (2022) provides universal approximation results for injective flows where the observation dimension is greater than the source dimension.

Proving identifiability for universal function approximators is a two edged sword. On the one hand, they can approximate any generating function, but on the other hand they can also approximate identifiability counter-examples like Measure Preserving Automorphisms (MPAs). Therefore, while Theorem 1 indicates that RAD is identifiable for compactly supported source distributions, training a real analytic approximator is very difficult because it could approximate an MPA. Note that this does not apply to source distributions on non-compact supports like the Laplace.

In particular, a model with triangular or truncated Laplace source distribution, where the tails are cut off, while theoretically identifiable, may be practically non-identifiable. In subsection 5.1 we tested these source distributions. Table 2 indicates that we were not able to recover the sources, confirming our hypothesis about the universal approximation problem.

Incidentally, this is also a limitation of (Kivva et al., 2022) due to the universal approximation ability of piecewise affine functions. One cannot prove identifiability for a class of universal approximators and then rely on those functions to approximate any function not in the class. Rather the practitioner must know something substantive about the generating function in order to achieve identifiability.

5 Experiment

We run three experiments (Synthetic, Yahoo Stock, and CelebA) to verify the theory. The synthetic experiments have ground truth labels and thus are the primary means of identifiability verification. We utilize the standard Mean Correlation Coefficient (MCC) metric, which is the mean absolute correlation between learned and true sources after matching. Full details on experiments are in Appendix B.

5.1 Synthetic Data

In these experiments we generate synthetic data and train Normalizing Flows to recover the ground truth factors, in order to confirm Theorem 1. Experiments on alternate generating functions are recorded in Appendix B.

Table 1: Average MCC $\pm$ standard deviation for Laplace source model with different activation functions in the Normalizing Flow. Real analytic activation functions achieved substantially higher MCCs than non-real analytic activation functions.

	$k = 5$		$k = 10$	
Activation Function	Mixture A	Mixture B	Mixture A	Mixture B
Tanh	0.944 ± 0.017	0.945 ± 0.007	0.930 ± 0.023	0.920 ± 0.037
Sigmoid	0.895 ± 0.030	0.925 ± 0.031	0.842 ± 0.085	0.895 ± 0.042
Softplus	0.833 ± 0.075	0.916 ± 0.014	0.765 ± 0.110	0.859 ± 0.032
LeakyHardTanh	0.431 ± 0.198	0.577 ± 0.081	0.448 ± 0.057	0.502 ± 0.030
LeakyReLU	0.143 ± 0.159	0.067 ± 0.057	0.257 ± 0.107	0.192 ± 0.188

The sources were independent and identically generated from source distribution ρ. We used two nonlinear mixture functions which we term mixture A and mixture B.

Mixture A: $\boldsymbol{f}$ is a post-nonlinear function $\boldsymbol{x}^{(i)} = \boldsymbol{a}(\boldsymbol{A}\boldsymbol{s}^{(i)})$ where $\boldsymbol{a}$ is applied in a component-wise manner. For each component j we randomly generated the function $\boldsymbol{f}_j(y) = \frac{1}{4}\sin(y + 2\pi w_i) + \frac{y}{3}$ where $w_i \sim U(0, 1)$. For each trial a new random normal mixing matrix $\boldsymbol{A}$ was generated. To make training easier, we shrank the log singular values by 30%. This shrinks very high and very low singular values which were hard for the neural networks to learn. The training would significantly slow down when singular values were too high or low. Then we normalized by $\sqrt{k}$ so that the range of $\boldsymbol{y}$ would fall on the most nonlinear part of f_i.

Mixture B: $\boldsymbol{f}$ is a 3-layer invertible neural network with $\mathrm{Softplus}$ activations. The 3 weight matrices were random normal matrices. Again, for training ease we shrank the log singular values of the weight matrices by 40%.

In the experiments below we train Normalizing flows with 6 dense invertible layers. Unless otherwise stated, the final layer has no activation applied. For each experiment we trained on 5 different randomly generated datasets using the same seeds across experiments. We calculate Mean Correlation Coefficient (MCC) and report mean and standard deviation across trials.

In the first experiment, we compare identifiability of a Laplacian source model for different activation functions for both $m = k = 5$ and $m = k = 10$ with $n = 10,000$ samples. Table 1 contains the results. Real analytic activation functions resulted in markedly higher MCCs for both Mixture A and Mixture B as Theorem 1 would indicate.

In the second experiment, we compare identifiability of the model with different source distributions. We fixed the activation function as Tanh to keep $\boldsymbol{g}$ real analytic and because Tanh performed the best in the first experiment. Otherwise the architecture was the same. We considered higher dimensions $m = k = 15, 25$. Interestingly we found that a lower sample of $n = 5,000$ would still achieve high MCCs. We tested six different source distributions: Laplace, Laplace Mixture Model (LMM), Gaussian, Gamma, Triangular, and Truncated Laplace. For Laplace and Gaussian source distributions, we set the mean to 0 and scale to 1. For the Gamma source distribution we used $\alpha = 5, \beta = 5$ and added a $\mathrm{Softplus}$ activation to the final layer in $\tilde{\boldsymbol{g}}$ to ensure positive values. The LMM was a mixture of two Laplace distributions with scales 1 and means -0.5 and 0.5 with a $0.75, 0.25$ weighting respectively. Because the standard triangular distribution has support on $[-1, 1]$ we applied a Tanh activation to the final layer. Finally the truncated Laplace distribution was a Laplace distribution truncated to the support $[-2, 2]$. So we applied a Tanh to the final layer and then multiplied by 2. According to Theorem 1 the Laplace, LMM, Triangular, and truncated Laplace models are identifiable while the Gaussian and Gamma are not. However, as mentioned in subsection 4.3, the Triangular, and truncated Laplace models have compactly supported source distributions and thus are not expected to be able to recover the sources.

Results reported in Table 2 show high MCCs for the Laplace and LMM models, and low MCCs for Gaussian, Gamma, Triangular, and Truncated Laplace models. The high MCCs of the LMM model demonstrate that even when there are multiple non-differentiable points the model is still identifiable. Finally we provide an example correlation plot between true and learned sources in Figure 2 comparing an identifiable Laplace model to a non-identifiable Gamma model.

Table 2: Average MCC $\pm$ standard deviation for different source distributions. The Laplace source distribution results in high MCCs even for higher dimensions.

	$k = 15$		$k = 25$	
Distribution	Mixture A	Mixture B	Mixture A	Mixture B
Laplace	0.965 ± 0.004	0.925 ± 0.015	0.953 ± 0.006	0.916 ± 0.037
LMM	0.937 ± 0.021	0.919 ± 0.027	0.939 ± 0.013	0.904 ± 0.025
Gaussian	0.514 ± 0.011	0.503 ± 0.034	0.417 ± 0.011	0.423 ± 0.018
Gamma	0.511 ± 0.029	0.442 ± 0.015	0.432 ± 0.021	0.399 ± 0.020
Triangular	0.453 ± 0.010	0.461 ± 0.013	0.385 ± 0.005	0.395 ± 0.007
Trunc. Laplace	0.484 ± 0.033	0.469 ± 0.016	0.400 ± 0.006	0.394 ± 0.008

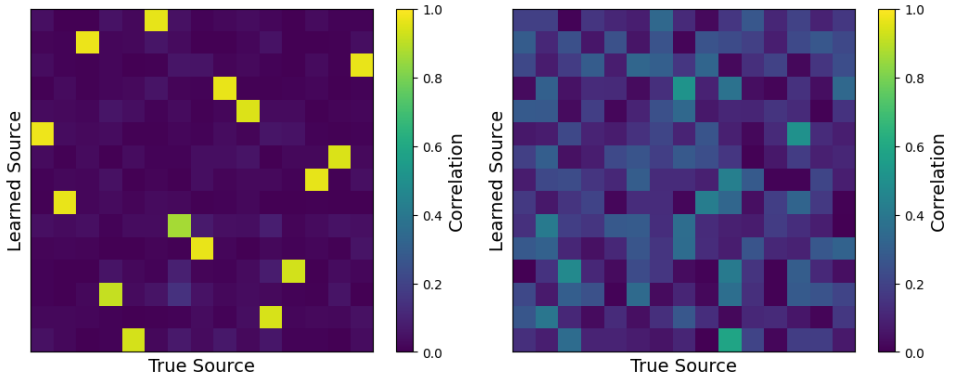


Figure 2: Correlation plot between learned and ground truth sources for Laplace source (left) and Gamma source (right). The identifiable Laplace model is able to recover the true sources, while the Gamma model is not.

5.2 Yahoo Stock Returns

Using yfinance[1], we downloaded daily closing stock prices for 15 companies for the period 1999-01-22 to 2026-08-31. See appendix for full data and training details. In order to obtain stationary data, we calculated the log returns for each stock, which is given by

$$r_s(t) = \log c_s(t) - \log c_s(t-1)$$

for closing price $c_s(t)$ of stock s at time t. We then trained Normalizing Flows with $\tanh$ activations to generate the returns. Because there is no ground truth to validate our results, following Kivva et al. (2022) we trained 10 models starting from distinct random seeds and compared latent sources between models. For all 45 distinct pairs of models we calculated the MCC between the learned sources. This was done for several source distributions as reported in Table 3. The results demonstrate that a Laplace prior made training much more stable across seeds, in accordance with Theorem 1.

Additionally, to compare against the results from Kivva et al. (2022) we trained a piecewise affine Normalizing Flow with independent Gaussian Mixture Model source distributions, which we term GMM. Each component had a mixture of two Gaussians that were jointly learned with the Normalizing Flow. We tried LeakyReLU and LeakyHardTanh activations with a 6-layer network. We report the LeakyReLU because it performed slightly better. The poor performance reported in Table 3 does not disprove the results from (Kivva et al., 2022), but it does demonstrate that this dataset does not satisfy the assumptions of (Kivva et al., 2022).

[1] https://ranaroussi.github.io/yfinance/

Table 3: Average $\pm$ standard error (SE) of MCC between model pairs trained on the Yahoo stock returns dataset for different source distributions. The Laplace distribution is the only one that satisfies Assumption 2 and thus has the highest average MCC.

Source Distribution	Avg MCC $\pm$ SE
Laplace	$\mathbf{0.871 \pm 0.012}$
Gaussian	0.516 ± 0.003
Exponential	0.447 ± 0.003
Gamma	0.492 ± 0.003
GMM	0.480 ± 0.044

5.3 CelebA

Lastly, we performed a larger scale experiment on the CelebA dataset (Liu et al., 2015) using the framework provided by Subramanian (2020). We trained a VAE with Laplacian prior and a latent space of 128. The encoder and decoder were CNNs with softplus and tanh activations to ensure the network is real analytic. See Appendix C for details.

The posterior distribution parameters are provided by $\tilde{\boldsymbol{g}}$. We modeled the posterior with a Laplace distribution since the prior distribution is also a Laplace. For sample i index j

$$p(\boldsymbol{s}_j^{(i)}|\boldsymbol{x}^{(i)}) = \text{Lap}(a_j^{(i)}, b_j^{(i)}) \propto \exp\left(-\frac{|\boldsymbol{s}_j^{(i)} - a_j^{(i)}|}{b_j^{(i)}}\right).$$

Nawa & Nadarajah (2024) provide the exact form of the KL Divergence between two Laplace distributions,

$$D_{KL}\left(\text{Lap}(a_0, b_0) \| \text{Lap}(a_1, b_1)\right) = \log\frac{b_1}{b_0} + \frac{b_0}{b_1}\exp\left(-\frac{|a_0 - a_1|}{b_0}\right) + \frac{|a_0 - a_1|}{b_1} - 1.$$

Following (Bojanowski et al., 2017) we augmented the standard mean squared error (MSE) reconstruction loss with Laplacian Pyramid loss, which helps avoid the excessive blurriness resulting from MSE reconstruction loss. Bojanowski et al. (2017) gives the Laplacian Pyramid loss as,

$$\text{Lap}_1(x, x') = \sum_j 2^{2j}|L^j(x) - L^j(x')|_1,$$

where L_j is the j-th level of the Laplacian pyramid (Ling & Okada, 2006). The results shown in Appendix C demonstrate that many of the resulting latent variables have a clear interpretation across different images.

6 Conclusion and Future Work

Real analytic functions comprise a large class of practically useful functions. Proving identifiability for real analytic transformations of sources with non-differentiable probability density functions is a significant step forward in nICA research. Additionally, unlike many nICA methods (Nguyen & Fu, 2025; Zheng et al., 2022; Gresele et al., 2021), RAD does not require expensive Jacobian regularization terms in the objective function. In fact, it can be trained with existing techniques such as Normalizing Flows and Variational Autoencoders.

However, there are still many open problems to be investigated, including the sample complexity of RAD and other nICA methods, and the trade-off between universal approximation and source recovery ability in nICA. Additionally, many nICA methods use VAEs, but assume that the likelihood can be perfectly learned. This is only possible when the KL-divergence between the true posterior and the variational distribution from the encoder is zero. If the encoder has infinite capacity this can happen, but in practice there is a gap between the ELBO and the likelihood. Future work would study how a non-zero gap affects identifiability results.

7 LLM USAGE DISCLOSURE

While all proofs were written by the authors, an LLM was employed to ideate on proof techniques for some parts of steps 2 and 3 of the proof of Theorem 1. The authors take responsibility for the correctness and comprehensibility of the proofs. LLMs were also used to provide feedback and suggest improvements for the work.

LLMs were not used to implement any significant portion of the experiments or propose novel ideas. Further LLMs were not used to write the paper apart from the feedback already mentioned.

8 REPRODUCIBILITY STATEMENT

Theorem 1, one of the main contributions, rests on the assumptions explicitly stated in section 4. The proof of the theorem is contained in section 4 and Appendix A. To reproduce experiments, we will release the code upon acceptance. Training details are provided in Appendix B and Appendix C.

REFERENCES

Piotr Bojanowski, Armand Joulin, David Lopez-Paz, and Arthur Szlam. Optimizing the latent space of generative networks. In *International Conference on Machine Learning*, 2017. URL https://api.semanticscholar.org/CorpusID:2019311.

Nicolas Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023.

Simon Buchholz, Michel Besserve, and Bernhard Schölkopf. Function classes for identifiable nonlinear independent component analysis. *Advances in Neural Information Processing Systems*, 35: 16946–16961, 2022.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *J. Mach. Learn. Res.*, 23(1), January 2022. ISSN 1532-4435.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Luigi Gresele, Julius Von Kügelgen, Vincent Stimper, Bernhard Schölkopf, and Michel Besserve. Independent mechanism analysis, a new concept? *Advances in neural information processing systems*, 34:28233–28248, 2021.

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

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), 2023. URL https://arxiv.org/abs/1606.08415.

Chin-Wei Huang, Ricky T. Q. Chen, Christos Tsirigotis, and Aaron Courville. Convex potential flows: Universal probability distributions with optimal transport and convex optimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=te7PVH1sPxJ.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd international conference on artificial intelligence and statistics*, pp. 859–868. Pmlr, 2019.

Aapo Hyvärinen, Ilyes Khemakhem, and Ricardo Monti. Identifiability of latent-variable and structural-equation models: from linear to nonlinear. *Annals of the Institute of Statistical Mathematics*, 76(1):1–33, 2024.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999. ISSN 0893-6080. doi: https://doi.org/10.1016/S0893-6080(98)00140-3. URL https://www.sciencedirect.com/science/article/pii/S0893608098001403.

Isao Ishikawa, Takeshi Teshima, Koichi Tojo, Kenta Oono, Masahiro Ikeda, and Masashi Sugiyama. Universal approximation property of invertible neural networks. *J. Mach. Learn. Res.*, 24(1), January 2023. ISSN 1532-4435.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International conference on artificial intelligence and statistics*, pp. 2207–2217. PMLR, 2020.

Diederik P Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Identifiability of deep generative models without auxiliary information. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Ivan Kobyzev, Simon J.D. Prince, and Marcus A. Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979, November 2021. ISSN 1939-3539. doi: 10.1109/tpami.2020.2992934. URL http://dx.doi.org/10.1109/TPAMI.2020.2992934.

Steven G. Krantz and Harold R. Parks. *A Primer of Real Analytic Functions*. Birkhäuser Boston, MA, 2nd edition, 2002. URL https://doi.org/10.1007/978-0-8176-8134-0.

Haibin Ling and K. Okada. Diffusion distance for histogram comparison. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pp. 246–253, 2006. doi: 10.1109/CVPR.2006.99.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Boris Mityagin. The zero set of a real analytic function, 2015. URL https://arxiv.org/abs/1512.07276.

Gemma Elyse Moran, Dhanya Sridhar, Yixin Wang, and David Blei. Identifiable deep generative models via sparse decoding. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=vd0onGWZbE.

Victor Nawa and Saralees Nadarajah. Exact expressions for kullback–leibler divergence for univariate distributions. *Entropy*, 26(11), 2024. ISSN 1099-4300. doi: 10.3390/e26110959. URL https://www.mdpi.com/1099-4300/26/11/959.

Hoang-Son Nguyen and Xiao Fu. Diverse influence component analysis: A geometric approach to nonlinear mixture identifiability, 2025. URL https://arxiv.org/abs/2510.17040.

Michael Puthawala, Matti Lassas, Ivan Dokmanic, and Maarten De Hoop. Universal joint approximation of manifolds and densities by simple injective flows. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 17959–17983. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/puthawala22a.html.

Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions, 2017. URL https://arxiv.org/abs/1710.05941.

Peter Sorrenson, Carsten Rother, and Ullrich Köthe. Disentanglement by nonlinear ica with general incompressible-flow networks (gin). In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygeHgSFDH.

A.K Subramanian. Pytorch-vae. https://github.com/AntixK/PyTorch-VAE, 2020.

Xiaojiang Yang, Yi Wang, Jiacheng Sun, Xing Zhang, Shifeng Zhang, Zhenguo Li, and Junchi Yan. Nonlinear ICA using volume-preserving transformations. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=AMpki9kp8Cn.

Yujia Zheng and Kun Zhang. Generalizing nonlinear ica beyond structural sparsity. *Advances in Neural Information Processing Systems*, 36:13326–13355, 2023.

Yujia Zheng, Ignavier Ng, and Kun Zhang. On the identifiability of nonlinear ica: Sparsity and beyond. *Advances in neural information processing systems*, 35:16411–16422, 2022.

A Main Theorem Continued Proof Step 3

We can also decompose γ as

$$\gamma(u, \boldsymbol{v}) = \gamma_i(u, \boldsymbol{v}) + \sum_{j \neq \pi_1(i)} \log \rho(\boldsymbol{h}_j(u, \boldsymbol{v})) - \sum_l \log \rho(\boldsymbol{v}_l)$$
$$\gamma_i(u, \boldsymbol{v}) = \log \rho(r_i(u, \boldsymbol{v})(u - \beta_j) + \beta_{\pi_2(j)}) - \log \rho(u).$$

Near $u = \beta_j, \boldsymbol{s} \in \mathcal{B}_{lm}, \forall (l,m) \neq (i,j)$, $\sum_{l \neq \pi_1(i)} \log \rho(\boldsymbol{h}_l(\boldsymbol{s})) - \sum_l \log \rho(\boldsymbol{v}_l)$ is real analytic, so γ_i must also be real analytic. At this point for notational simplicity, we focus on a fixed $\boldsymbol{v}$ and therefore drop $\boldsymbol{v}$ from the notation. We consider two cases which require slightly different techniques. The cases correspond to when $\boldsymbol{h}_{\pi_1(i)}$ is the identity of u and when $\boldsymbol{h}_{\pi_1(i)}$ reverses u.

Case 1: $r_i(u) > 0$ Because $\gamma_i(u)$ is real analytic, $\gamma_i'(u)$ must be continuous across $u = \beta_j$.

$$\lim_{u \to \beta_j^+} \gamma_i'(u) = \lim_{u \to \beta_j^+} \left(\alpha_{\pi_2(j)}(r_i(u)(u - \beta_j) + \beta_{\pi_2(j)}) - \alpha_j(u)\right)'$$
$$= \lim_{u \to \beta_j^+} \alpha'_{\pi_2(j)}(r_i(u)(u - \beta_j) + \beta_{\pi_2(j)})(r_i'(u)(u - \beta_j) + r_i(u)) - \alpha_j'(u)$$
$$= \alpha'_{\pi_2(j)}(\beta_{\pi_2(j)}) r_i(\beta_j) - \alpha_j'(\beta_j)$$

The same reasoning can be done for $u \to \beta_j^-$. Continuity at $u = \beta_j$ gives,

$$\alpha'_{\pi_2(j)}(\beta_{\pi_2(j)}) r_i(\beta_j) - \alpha_j'(\beta_j) = \alpha'_{\pi_2(j)-1}(\beta_{\pi_2(j)}) r_i(\beta_j) - \alpha'_{j-1}(\beta_j)$$
$$r_i(\beta_j) = \frac{\alpha_j'(\beta_j) - \alpha'_{j-1}(\beta_j)}{\alpha'_{\pi_2(j)}(\beta_{\pi_2(j)}) - \alpha'_{\pi_2(j)-1}(\beta_{\pi_2(j)})} \tag{3}$$

But since $\frac{\partial \boldsymbol{h}_l(\beta_j)}{\partial u} = r_i'(\beta_j)(\beta_j - \beta_j) + r_i(\beta_j) > 0$, π_2 is the identity from the continuity argument in step 2. Therefore $r_i(\beta_j) = 1$.

Let the power series of r_i around $u = \beta_j$ be $r_i(u) = 1 + \sum_{l=n}^{\infty} c_l \left(u - \beta_j\right)^l$, where coefficients $c_1, \ldots, c_{n-1} = 0$ for $n \geq 1$, where $n = 1$ means no coefficients are 0. Since α_j is real analytic it can be described by its power series $\alpha_j(x) = \sum_{m=1}^{\infty} a_m^{(j)} (x - \beta_j)^m$. Therefore

$$\alpha_j(\beta_j + (u - \beta_j) r_i(u)) = \sum_{m=1}^{\infty} a_m^{(j)} \left((u - \beta_j) + \sum_{l=n}^{\infty} c_l (u - \beta_j)^{l+1}\right)^m$$
$$\left(u - \beta_j + \sum_{l=n}^{\infty} c_l (u - \beta_j)^{l+1}\right)^m = (u - \beta_j)^m + m c_n (u - \beta_j)^{m+n} + \mathcal{O}((u - \beta_j)^{m+n+1})$$
$$\alpha_j(\beta_j + (u - \beta_j) r_i(u)) - \alpha_j(u) = a_1^{(j)} c_n (u - \beta_j)^{n+1} + \mathcal{O}((u - \beta_j)^{n+2})$$

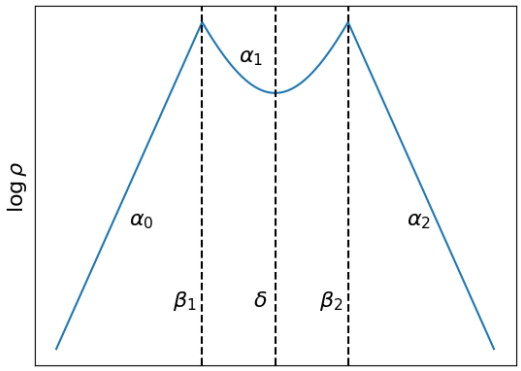


Figure 3: Example Log Density

We can do the same thing for α_{j-1} to get

$$\alpha_{j-1}(\beta_j + (u-\beta_j)r_i(u)) - \alpha_{j-1}(u) = a_1^{(j-1)} c_n (u-\beta_j)^{n+1} + \mathcal{O}((u-\beta_j)^{n+2})$$

Because $\gamma_i(u)$ is real analytic, as $u \to \beta_j$ the coefficients of $(u-\beta_j)^{n+1}$ must be equal. But, $a_1^{(j-1)} \neq a_1^{(j)}$, so $c_n = 0$. We can do this inductively to show that all Taylor coefficients of r_i are 0 except for the constant term. Since r_i is real analytic, this means that $r_i \equiv 1$.

Case 2: $r_i(u) < 0$

When d is odd, for $j = \frac{d+1}{2}$ $\pi_2(j) = j$ due to the fact that π_2 is a reverse ordering. The rest of the arguments mirror Case 1, resulting in $r_i \equiv -1$. This means that order reversal can only happen on symmetric densities.

However, for d even, the argument does not mirror exactly. Because $r_i \neq 0$ and r_i is continuous, it must have the same sign on its domain. There exists $\delta \in [\beta_{\frac{d}{2}}, \beta_{\frac{d}{2}+1}]$, such that $\boldsymbol{h}_{\pi_1(i)}(\delta, \boldsymbol{v}) = \delta$ (see Figure 3). So using the Weierstrass Preparation Theorem as before we factorize $\boldsymbol{h}_{\pi_1(i)}$

$$\boldsymbol{h}_{\pi_1(i)}(\boldsymbol{s}) - \delta = r_i(\boldsymbol{s})(\boldsymbol{s}_i - \delta)$$

The power series of r_i and α_j are $r_i(u) = c_0 + \sum_{l=n}^{\infty} c_l (u-\delta)^l$ and $\alpha_j(x) = \sum_{m=1}^{\infty} a_m^{(j)} (x-\delta)^m$.

$$\alpha_{\pi_2(j)}(r_i(u)(u-\delta)+\delta) = \sum_{m=1}^{\infty} a_m^{(\pi_2(j))} \left(c_0(u-\delta) + \sum_{l=n}^{\infty} c_l (u-\delta)^{l+1} \right)^m$$

$$\left(c_0(u-\delta) + \sum_{l=n}^{\infty} c_l (u-\delta)^{l+1} \right)^m = c_0^m (u-\delta)^m + m c_0^{m-1} c_n (u-\delta)^{n+m} + \mathcal{O}\left((u-\delta)^{n+m+1}\right)$$

$$\alpha_{\pi_2(j)}(r_i(u)(u-\delta)+\delta) - \alpha_{j-1}(u)$$

$$= \sum_{m=1}^{\infty} (a_m^{(\pi_2(j))} c_0^m - a_m^{(j-1)})(u-\delta)^m + a_1^{(\pi_2(j))} c_n (u-\delta)^{n+1} + \mathcal{O}\left((u-\delta)^{n+2}\right)$$

This is precisely γ_i for $u \in (\beta_{j-1}, \beta_j)$. For each interval we can do this same expansion. Since γ_i is real analytic, any power series expansion around δ must have the same coefficients. From the coefficients of the first order terms we have,

$$a_1^{(\pi_2(j))} c_0 - a_1^{(j-1)} = a_1^{(\pi_2(j)-1)} c_0 - a_1^{(j)}$$

$$c_0 = \frac{a_1^{(j)} - a_1^{(j-1)}}{a_1^{(\pi_2(j)-1)} - a_1^{(\pi_2(j))}}$$

With doing this same calculation for the intervals $(\beta_{\pi_2(j)-1}, \beta_{\pi_2(j)})$ and $(\beta_{\pi_2(j)}, \beta_{\pi_2(j)+1})$ we arrive at

$$c_0 = \frac{a_1^{(\pi_2(j)-1)} - a_1^{(\pi_2(j))}}{a_1^{(j)} - a_1^{(j-1)}}$$
$$c_0 = \frac{1}{c_0} \implies c_0 = -1$$

because c_0 cannot be positive due to the case $r_i < 0$. The $n+1$ order term must also have equal coefficients for all intervals. Thus the (β_j, β_{j+1}) interval produces the $n+1$ order taylor coefficient for γ_i,

$$a_{n+1}^{(\pi_2(j)-1)}(-1)^{n+1} - a_{n+1}^{(j)} + a_1^{(\pi_2(j)-1)} c_n \tag{4}$$

If n is odd we equate the $n+1$ coefficient for intervals (β_{j-1}, β_j) and (β_j, β_{j+1}) and then for intervals $(\beta_{\pi_2(j)-1}, \beta_{\pi_2(j)})$ and $(\beta_{\pi_2(j)}, \beta_{\pi_2(j)+1})$ to get the following two expressions for c_n,

$$c_n = \frac{(a_{n+1}^{(\pi_2(j)-1)} - a_{n+1}^{(\pi_2(j))}) - (a_{n+1}^{(j)} - a_{n+1}^{(j-1)})}{a_1^{(\pi_2(j))} - a_1^{(\pi_2(j)-1)}}$$
$$c_n = \frac{(a_{n+1}^{(j)} - a_{n+1}^{(j-1)}) - (a_{n+1}^{(\pi_2(j)-1)} - a_{n+1}^{(\pi_2(j))})}{a_1^{(j)} - a_1^{(j-1)}}$$

For the kinks in ρ and $\rho \circ \boldsymbol{h}$ to cancel each other when $r_i < 0$, the differences $a_1^{(\pi_2(j))} - a_1^{(\pi_2(j)-1)}$ and $a_1^{(j)} - a_1^{(j-1)}$ must have the same sign. However the numerators in the equation above are negatives of each other. Therefore $c_n = 0$.

If n is even we equate the coefficients in Equation 4 for intervals (β_{j-1}, β_j) and $(\beta_{\pi_2(j)}, \beta_{\pi_2(j)+1})$ where $j < \frac{d}{2}$ to get,

$$c_n = \frac{(a_{n+1}^{(j-1)} - a_{n+1}^{(\pi_2(j))})(-1) - (a_{n+1}^{(\pi_2(j))} - a_{n+1}^{(j-1)})}{a_1^{(\pi_2(j))} - a_1^{(j-1)}} = 0$$

$a_1^{(\pi_2(j))} - a_1^{(j-1)} \neq 0$ by the cancellation argument above.

B Experiment Details

B.1 Synthetic

For the first experiment with $m = k = 5, 10$ we generated 10,000 data points to train on for 500 epochs. For $m = k = 15, 25$ we generated 5,000 data points to train on for 1000 epochs. The optimizer was Adam with learning rate 5e-3 and batches of 128 samples. We employed gradient clipping for training stability.

We experimented with two other generating functions. Mixture C was a 3-layer neural network with orthogonal layers and $\tanh$ activation. Mixture D was an alternate post-nonlinear formulation with $a_i(y) = w_i y + \tanh(y)(1 - w_i)$.

B.2 Stock NF

The stock data was from the period 1999-01-22 to 2026-08-31. We chose large companies from a wide variety of market sectors (Table 5). From Figure 4 the auto-correlation does not seem to be

Table 4: Average MCC $\pm$ standard deviation for Laplace source model with different activation functions in the Normalizing Flow. Real analytic activation functions achieved substantially higher MCCs than non-real analytic activation functions.

	$k = 5$		$k = 10$	
Activation Function	Mixture A	Mixture B	Mixture A	Mixture B
Tanh	0.931 ± 0.015	0.993 ± 0.003	0.928 ± 0.008	0.983 ± 0.006
Sigmoid	0.926 ± 0.016	0.987 ± 0.002	0.860 ± 0.037	0.987 ± 0.006
Softplus	0.877 ± 0.063	0.991 ± 0.002	0.829 ± 0.042	0.980 ± 0.007
LeakyHardTanh	0.656 ± 0.080	0.660 ± 0.154	0.561 ± 0.092	0.540 ± 0.039
LeakyReLU	0.136 ± 0.116	0.066 ± 0.041	0.095 ± 0.105	0.038 ± 0.011

significant and the data appears to be relatively stationary. For each model we trained for 750 epochs with Adam optimizer, learning rate set at 1e-2, with a decline on plateau scheduler, and batches of 256 samples. We arbitrarily chose the seeds 9-18 to initialize the 10 models with.

Table 5: Company names of stock data in the Yahoo Stock experiment. For easier replication we provide the Yahoo Finance Symbol over the company names.

AAPL Apple Inc.	MSFT Microsoft Corp.	NVDA NVIDIA Corp.	AMZN Amazon Inc.
ALL The Allstate Corp.	JPM JPMorgan Chase & Co.	PFE Pfizer Inc.	JNJ Johnson & Johnson
CVX Chevron Corp.	WMT Walmart Inc.	CAT Caterpillar Inc.	DIS The Walt Disney Co.
XOM ExxonMobil Corp.	AMD Advanced Micro Devices Inc.	BRK-A Berkshire Hathaway Inc.	

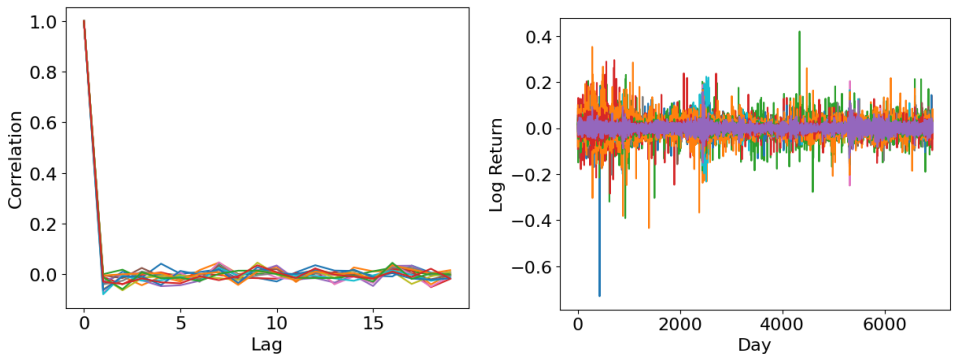


Figure 4: Auto-correlation and trace plots of stock returns data

C CelebA

The Variational Autoencoder architecture used convolutional layers for the encoder and convolutional transpose layers for the decoder. The kernel size was 3, with 32, 64, 128, 256, 512 filters for the 5 convolutional layers. We used the $\mathrm{Softplus}$ activation for hidden layers and the Tanh for the final layer of the decoder. We used a latent dimension of $k = 128$ and trained for 100 epochs with a learning rate of 0.005 on one B200 GPU.

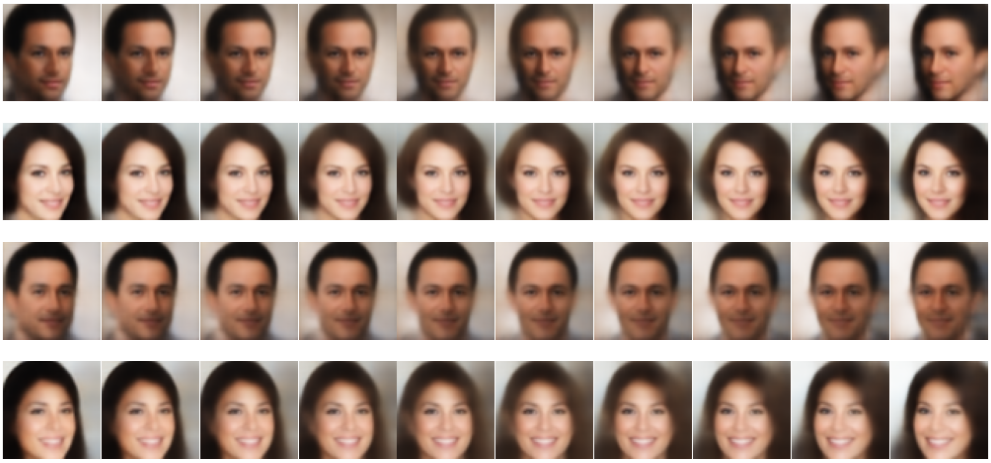


Figure 5: Head orientation

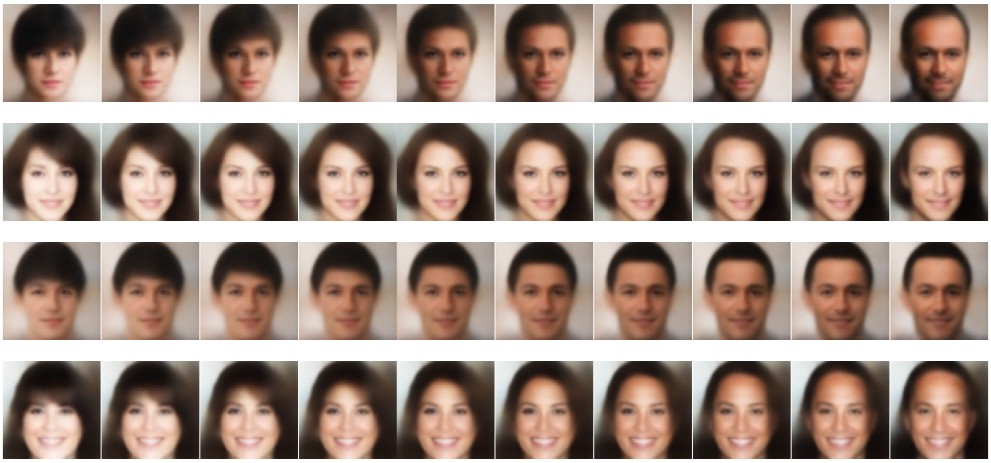


Figure 6: Bangs

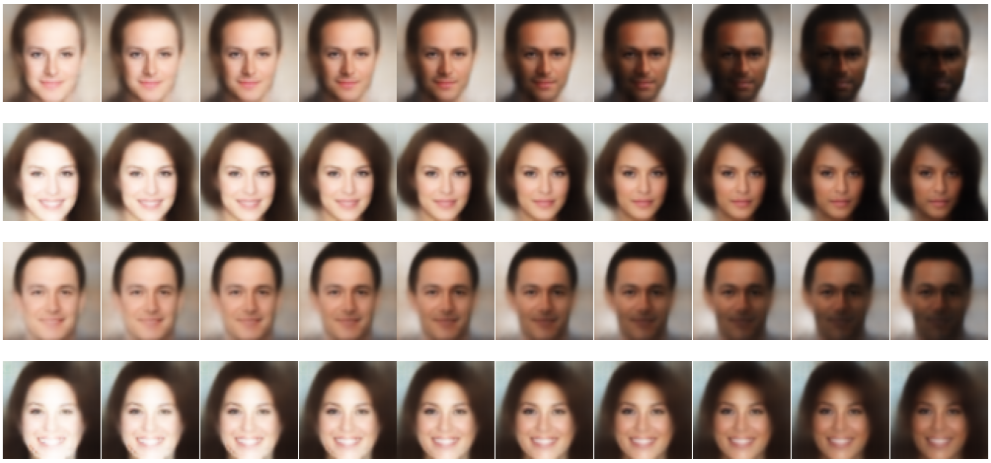


Figure 7: Skin Color

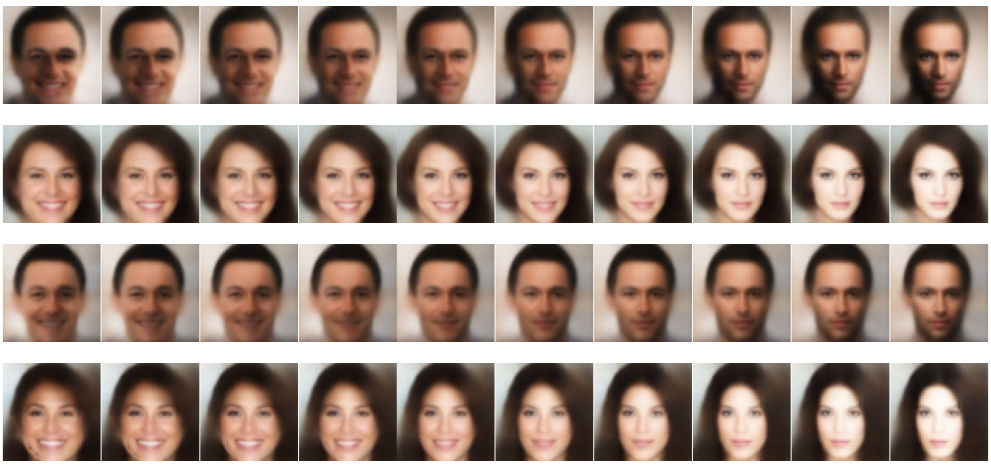


Figure 8: Smile/Broad face