Robust contrastive learning and nonlinear ICA in the presence of outliers

Hiroaki Sasaki[1], **Takashi Takenouchi**[1,2], **Ricardo Monti**[3], **Aapo Hyvärinen**[4,5]
[1]Future University Hakodate, Hokkaido, Japan [2]RIKEN AIP, Tokyo, Japan [3]University College London, UK
[4]Université Paris-Saclay, Inria, CEA, France [5]University of Helsinki, Finland

Abstract

Nonlinear independent component analysis (ICA) is a general framework for unsupervised representation learning, and aimed at recovering the latent variables in data. Recent practical methods perform nonlinear ICA by solving classification problems based on logistic regression. However, it is well-known that logistic regression is vulnerable to outliers, and thus the performance can be strongly weakened by outliers. In this paper, we first theoretically analyze nonlinear ICA models in the presence of outliers. Our analysis implies that estimation in nonlinear ICA can be seriously hampered when outliers exist on the tails of the (non-contaminated) target density, which happens in a typical case of contamination by outliers. We develop two robust nonlinear ICA methods based on the γ-divergence, which is a robust alternative to the KL-divergence in logistic regression. The proposed methods are theoretically shown to have desired robustness properties in the context of nonlinear ICA. We also experimentally demonstrate that the proposed methods are very robust and outperform existing methods in the presence of outliers. Finally, the proposed method is applied to ICA-based causal discovery and shown to find a plausible causal relationship on fMRI data.

1 Introduction

Nonlinear independent component analysis (ICA) is a principled framework for unsupervised representation learning which has generated a large amount of recent interest in learning deep neural networks. Unlike most of unsupervised methods, nonlinear ICA is based on a clearly defined statistical estimation task. The problem is rigorously formulated by defining a generative model for the data, and the goal is to recover (or identify) the latent source components of which data is observed as a general nonlinear mixing. Nonlinear ICA includes a number of potential applications such as causal analysis [Monti et al., 2019] and transfer learning [Noroozi and Favaro, 2016].

In contrast to the success of *linear* ICA [Hyvärinen and Oja, 2000], nonlinear ICA has not received so much attention until recently because the problem is fundamentally ill-posed in its basic form: There exist an infinite number of decompositions of a random vector into mutually independent variables [Hyvärinen and Pajunen, 1999, Locatello et al., 2019], while the identifiability proof is established in linear ICA [Comon, 1994]. Thus, in general, we cannot recover the source components under the same conditions as linear ICA.

Novel identifiability proofs for nonlinear ICA have been recently established [Sprekeler et al., 2014, Hyvärinen and Morioka, 2016, 2017, Hyvärinen et al., 2019]. The main idea is to introduce some auxiliary variables given which the latent source components are conditionally independent. For instance, *time contrastive learning* divides time series data into a number of time segments and uses the time segment label as the auxiliary variable [Hyvärinen and Morioka, 2016]; in *permutation contrastive learning*, the auxiliary variable is the history of time-series data [Hyvärinen and Morioka, 2017]. Interestingly, a heuristic yet successful approach called *self-supervised learning* [Noroozi and Favaro, 2016, Larsson et al., 2017, Oord et al., 2018] also takes the same approach of solving unsupervised learning problems through classification. Thus, the theory of nonlinear ICA might shed light on the principles underlying self-supervised learning.

In order to solve the nonlinear ICA problem in practice, logistic regression has been employed [Hyvärinen and Morioka, 2016, 2017, Hyvärinen et al., 2019], which is

Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR volume 124, 2020.

based on (conditional) maximum likelihood estimation (MLE). MLE has a number of useful properties, but it is well-known to be vulnerable to outliers. Thus, the performance of the existing nonlinear ICA methods might be strongly degraded by outliers. This is a very important problem because outliers are ubiquitous on real-world datasets. For instance, outliers have been often observed in functional MRI data to which ICA methods have been applied [Monti et al., 2019].

In this paper, we first define a contaminated density model of sources as a mixture of the (noncontaminated) target and outlier densities, and then theoretically analyze how outliers hamper estimation in nonlinear ICA. Our analysis implies that estimation in nonlinear ICA might be degraded particularly when the ratio of the outlier density to the target density can take a very large value. This large ratio happens when the outlier density lies on the tails of the target density as in a typical case of contamination by outliers.

Next, we propose two robust methods for nonlinear ICA. Our methods also solve classification problems, but are based on the γ-divergence [Fujisawa and Eguchi, 2008]. γ-divergence is a generalization of KL-divergence and has a favorable robustness property expressed as the *super robustness* [Cichocki and Amari, 2010, Amari, 2016]: The latent bias caused from outliers can be small even in the case of heavy contamination. This is in stark contrast with the density power divergence [Basu et al., 1998], which is often proved to be robust under small contamination of outliers. We theoretically show that the super robustness holds in the context of nonlinear ICA. Furthermore, the proposed method is proved to have desirable robustness properties in terms of influence function as well [Hampel et al., 2011]. We experimentally demonstrate that the proposed methods are much more robust against outliers than existing methods. Finally, our robust method is applied to causal analysis and demonstrated to find a plausible causal relationship on fMRI data.

2 Background

ICA is a rigorous framework for unsupervised learning, and assumes that the d_{x}-dimensional vectors of observed data $\boldsymbol{x}(t) := (x_1(t), \ldots, x_{d_{\mathrm{x}}}(t))^\top,\ t = 1, \ldots, T$ are generated from a nonlinear mixing of the source vectors $\boldsymbol{s}(t) = (s_1(t), \ldots, s_{d_{\mathrm{x}}}(t))^\top$ as

$$\boldsymbol{x}(t) = \boldsymbol{f}(\boldsymbol{s}(t)), \tag{1}$$

where $\boldsymbol{f}(\boldsymbol{s}) = (f_1(\boldsymbol{s}), \ldots, f_{d_{\mathrm{x}}}(\boldsymbol{s}))^\top$, and f_i denotes a smooth and invertible nonlinear function. The goal is to recover (or identify) the sources from data only.

Nonlinear ICA has been hampered by the fact that the problem is seriously ill-posed and the original sources cannot be recovered (i.e., not identifiable) under the same independence assumption as linear ICA, although some empirical success has been achieved by heuristic methods [Wiskott and Sejnowski, 2002, Harmeling et al., 2003]. Recently, novel identifiability proofs have been established together with practical algorithms [Sprekeler et al., 2014, Hyvärinen and Morioka, 2016, 2017, Hyvärinen et al., 2019]. Time contrastive learning (TCL) [Hyvärinen and Morioka, 2016] divides time series data $\{\boldsymbol{x}(t)\}_{t=1}^T$ into K time segments, and then a time segment label $u(t) = k,\ k = 1, \ldots, K$ is assigned to $\boldsymbol{x}(t)$ in the k-th segment. A nonlinear feature $\boldsymbol{h}(\boldsymbol{x}) = (h_1(\boldsymbol{x}), \ldots, h_{d_{\mathrm{x}}}(\boldsymbol{x}))^\top$ modelled by a neural network is learned via multinomial logistic regression to the *artificial* supervised dataset $\{(u(t), \boldsymbol{x}(t))\}_{t=1}^T$. For the identifiability, when the conditional density of $\boldsymbol{s}$ given a time segment label u is conditionally independent and belongs to an exponential family as,

$$\log p^\star(\boldsymbol{s}|u) = \sum_{j=1}^{d_{\mathrm{x}}} \lambda_{u,j} q_j^\star(s_j) - \log Z(\boldsymbol{\lambda}_u), \tag{2}$$

where $q_j^\star$ is a scalar function, $\lambda_{u,j}$ denotes a parameter depending on the time-segment label u, $\boldsymbol{\lambda}_u := (\lambda_{u,1}, \ldots, \lambda_{u,d_{\mathrm{x}}})^\top$, and $Z(\boldsymbol{\lambda}_y)$ is the partition function, then Theorem 1 in Hyvärinen and Morioka [2016] states that the learned $\boldsymbol{h}(\boldsymbol{x})$ asymptotically equals to $\boldsymbol{q}^\star(\boldsymbol{s}) = (q_1^\star(s_1), \ldots, q_{d_{\mathrm{x}}}^\star(s_{d_{\mathrm{x}}}))^\top$ up to a linear transformation.

A more general theory without the exponential family assumption (2) was established in Hyvärinen et al. [2019]. Suppose that some auxiliary data $\boldsymbol{u}(t) = (u_1(t), \ldots, u_{d_{\mathrm{u}}}(t))^\top$ is available in addition to $\boldsymbol{x}(t)$. For instance, the time segment label in TCL can be interpreted as auxiliary data, and another existing method, permutation contrastive learning, employs the past information of $\boldsymbol{x}(t)$ (e.g, $\boldsymbol{u}(t) = \boldsymbol{x}(t-1)$) [Hyvärinen and Morioka, 2017]. In the general theory, a nonlinear feature $\boldsymbol{h}(\boldsymbol{x})$ is learned by solving the following binary classification problem based on logistic regression:

$$\mathcal{D} := \{(\boldsymbol{x}(t), \boldsymbol{u}(t)\}_{t=1}^T \text{ vs. } \mathcal{D}_{\mathrm{p}} := \{(\boldsymbol{x}(t), \boldsymbol{u}_{\mathrm{p}}(t))\}_{t=1}^T, \tag{3}$$

where $\boldsymbol{u}_{\mathrm{p}}(t)$ is a random permutation of $\boldsymbol{u}(t)$ with respect to t, that is, a time-shuffled version $\boldsymbol{u}(t)$. Eq.(3) indicates $\mathcal{D}$ is drawn from the joint density of $\boldsymbol{x}(t)$ and $\boldsymbol{u}(t)$, while the underlying density of $\mathcal{D}_{\mathrm{p}}$ can be regarded as the product of marginal densities of $\boldsymbol{x}(t)$ and $\boldsymbol{u}(t)$. Under the even more general conditional independence assumption,

$$\log p^\star(\boldsymbol{s}|\boldsymbol{u}) = \sum_{i=1}^{d_{\mathrm{x}}} q^\star(s_i|\boldsymbol{u}) - \log Z(\boldsymbol{u}), \tag{4}$$

where $Z(\boldsymbol{u})$ denotes the partition function and $q^\star$ is a twice differential function, it was proved that the learned $\boldsymbol{h}$ equals to $\boldsymbol{s}$ up to an invertible function when $p^\star(\boldsymbol{s}|\boldsymbol{u})$ is sufficiently diverse and complex [Hyvärinen et al., 2019, Theorem 1].[1]

In practice, the nonlinear ICA methods above employ logistic regression to learn $\boldsymbol{h}(\boldsymbol{x})$, which is based on the (conditional) maximum likelihood estimation (MLE). MLE has a number of useful properties such as asymptotic efficiency [Wasserman, 2006]; on the other hand, it is well-known to be vulnerable against outliers. Very recently, Khemakhem et al. [2019] proposed an alternative nonlinear ICA method, which is based on maximizing a lower bound of the likelihood of a density model. Thus, these nonlinear ICA methods might be sensitive to outliers. Next, we first theoretically investigate how outliers hamper estimation in nonlinear ICA, and then propose robust practical methods.

3 Influence of outliers in nonlinear ICA

This section theoretically investigates the influence of outliers in existing methods for nonlinear ICA, which motivates us to develop robust methods.

3.1 Contaminated density model by outliers

Here, we assume the following contaminated conditional density model of $\boldsymbol{s}$ given auxiliary variables $\boldsymbol{u}$:

$$p(\boldsymbol{s}|\boldsymbol{u}) = (1-\epsilon(\boldsymbol{u}))p^\star(\boldsymbol{s}|\boldsymbol{u}) + \epsilon(\boldsymbol{u})\delta(\boldsymbol{s}|\boldsymbol{u}), \quad (5)$$

This equation means that the original sources in $p^\star(\boldsymbol{s}|\boldsymbol{u})$ are *contaminated* by outliers generated from the outlier density $\delta(\boldsymbol{s}|\boldsymbol{u})$. Here, $\epsilon(\boldsymbol{u})$ is a contamination ratio in $[0,1)$. We call $p^\star(\boldsymbol{s}|\boldsymbol{u})$ the *target* density throughout this paper because it generates the target sources which we want to recover from data $\boldsymbol{x}$. The density model (5) is very general and called *heterogeneous* contamination because $\epsilon(\boldsymbol{u})$ can be dependent on $\boldsymbol{u}$. Next, we investigate how estimation in nonlinear ICA is hampered under the outlier model (5).

3.2 Influence of outliers in conditionally exponential case

We first consider the method in Hyvärinen et al. [2019, Section 4.3], thus focusing on the following conditionally independent and exponential family density, which generalizes the exponential family (2) in TCL with one-dimensional discrete variable (i.e, time-segment label):

$$\log p^\star(\boldsymbol{s}|\boldsymbol{u}) = \sum_{j=1}^{d_{\mathrm{x}}} \lambda_j(\boldsymbol{u}) q_j^\star(s_j) - \log Z(\boldsymbol{\lambda}(\boldsymbol{u})), \quad (6)$$

where $\boldsymbol{\lambda}(\boldsymbol{u}) := (\lambda_1(\boldsymbol{u}), \ldots, \lambda_{d_{\mathrm{x}}}(\boldsymbol{u}))^\top$ and $\boldsymbol{u}$ denotes the auxiliary variables. Then, to investigate how the outlier density $\delta(\boldsymbol{s}|\boldsymbol{u})$ hampers estimation in nonlinear ICA, we establish the following theorem:

Theorem 1. *First, the following assumptions are made:*

(A1) Data $\boldsymbol{x}$ is generated from (1) *where $\boldsymbol{f}$ is invertible.*

(A2) The target density $p^\star(\boldsymbol{s}|\boldsymbol{u})$ is conditionally independent and belongs to the exponential family (6)*.*

(A3) For all $\boldsymbol{s}$ and $\boldsymbol{u}$, $\frac{\delta(\boldsymbol{s}|\boldsymbol{u})}{p^\star(\boldsymbol{s}|\boldsymbol{u})}$ is finite.

(A4) In the limit of infinite data, $p(\boldsymbol{x}|\boldsymbol{u})$ is universally approximated as

$$\log \frac{p(\boldsymbol{x}|\boldsymbol{u})}{c(\boldsymbol{x})e(\boldsymbol{u})} = \boldsymbol{w}(\boldsymbol{u})^\top \boldsymbol{h}(\boldsymbol{x}), \quad (7)$$

where $\boldsymbol{w}(\boldsymbol{u}) := (w_1(\boldsymbol{u}), \ldots, w_{d_{\mathrm{x}}}(\boldsymbol{u}))^\top$ is a vector-valued function, and c and e some scalar functions.

(A5) There exist $m+1$ points $\boldsymbol{u}_0, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ such that the following matrices are full rank: $\bar{\boldsymbol{\Lambda}} := \sum_{i=1}^m \bar{\boldsymbol{\lambda}}(\boldsymbol{u}_i)\bar{\boldsymbol{\lambda}}(\boldsymbol{u}_i)^\top$ and $\sum_{i=1}^m \bar{\boldsymbol{w}}(\boldsymbol{u}_i)\bar{\boldsymbol{\lambda}}(\boldsymbol{u}_i)^\top$, where $\bar{\boldsymbol{\lambda}}(\boldsymbol{u}) := \boldsymbol{\lambda}(\boldsymbol{u}) - \boldsymbol{\lambda}(\boldsymbol{u}_0)$ and $\bar{\boldsymbol{w}}(\boldsymbol{u}) := \boldsymbol{w}(\boldsymbol{u}) - \boldsymbol{w}(\boldsymbol{u}_0)$.

Then, regarding sufficiently small $\epsilon(\boldsymbol{u})$ for all $\boldsymbol{u}$, in the limit of infinite data,

$$\boldsymbol{q}^\star(\boldsymbol{s}) + \boldsymbol{Q}(\boldsymbol{s}) = \boldsymbol{A}\boldsymbol{h}(\boldsymbol{x}) + \boldsymbol{\alpha}, \quad (8)$$

where $\boldsymbol{A}$ is a d_{x} by d_{x} invertible matrix, $\boldsymbol{\alpha}$ is a d_{x}-dimensional vector, and with $\bar{\boldsymbol{\omega}}(\boldsymbol{u}) := \bar{\boldsymbol{\Lambda}}^{-1}\bar{\boldsymbol{\lambda}}(\boldsymbol{u})$, $\mathbf{1}_{d_{\mathrm{x}}} = (1,1,\ldots,1)^\top$ and $\epsilon_{\max} := \max_{i=0,1,\ldots,m} \epsilon(\boldsymbol{u}_i)$,

$$\boldsymbol{Q}(\boldsymbol{s}) := \sum_{i=1}^m \left\{ \epsilon(\boldsymbol{u}_i)\frac{\delta(\boldsymbol{s}|\boldsymbol{u}_i)}{p^\star(\boldsymbol{s}|\boldsymbol{u}_i)} - \epsilon(\boldsymbol{u}_0)\frac{\delta(\boldsymbol{s}|\boldsymbol{u}_0)}{p^\star(\boldsymbol{s}|\boldsymbol{u}_0)} \right\} \bar{\boldsymbol{\omega}}(\boldsymbol{u}_i) + O(\epsilon_{\max}^2)\mathbf{1}_{d_{\mathrm{x}}}. \quad (9)$$

The proof is deferred to Section A in the supplementary material. Assumption (A5) implies that the conditional density (6) and $\boldsymbol{w}(\boldsymbol{u})^\top \boldsymbol{h}(\boldsymbol{x})$ in (7) are sufficiently diverse with respect to the auxiliary variables $\boldsymbol{u}$. For instance, if $\boldsymbol{\lambda}(\boldsymbol{u})$ and and $\boldsymbol{w}(\boldsymbol{u})$ are constant vectors (i.e., $\bar{\boldsymbol{\lambda}}(\boldsymbol{u}) = \mathbf{0}$ and $\bar{\boldsymbol{w}}(\boldsymbol{u}) = \mathbf{0}$), then Assumption (A5) never holds. Such a full-rank assumption is found in the previous theory of nonlinear ICA as well [Hyvärinen et al., 2019] even if in slightly different forms.

[1] The complexity and diversity of $p^\star(\boldsymbol{s}|\boldsymbol{u})$ is expressed by the *Assumption of Variability* in Hyvärinen et al. [2019].

In the case of no outliers, (8) is essentially the same identifiability result as Theorem 3 in Hyvärinen et al. [2019] as well as Theorem 1 in Hyvärinen and Morioka [2016] for TCL: $\epsilon(\boldsymbol{u}) = 0$ leads to $\boldsymbol{Q}(\boldsymbol{s}) = \boldsymbol{0}$, and therefore $\boldsymbol{h}(\boldsymbol{x})$ equals to $\boldsymbol{q}^\star(\boldsymbol{s})$ up to a linear transformation. This linear indeterminacy could be removed by applying some linear ICA method in postprocessing.

On the other hand, when there are outliers, i.e. $\epsilon(\boldsymbol{u}) \neq 0$, Theorem 1 indicates that estimation for the exponential family might be hampered by $\boldsymbol{Q}(\boldsymbol{s})$. In particular, the elements in $\boldsymbol{Q}(\boldsymbol{s})$, which gives the estimation error induced by the outliers, can be significantly nonzero. This can be the case if the density ratio $\frac{\delta(\boldsymbol{s}|\boldsymbol{u})}{p^\star(\boldsymbol{s}|\boldsymbol{u})}$ in (9) is very large. That can happen when $\delta(\boldsymbol{s}|\boldsymbol{u})$ lies on the tails of $p^\star(\boldsymbol{s}|\boldsymbol{u})$ (i.e., very small $p^\star(\boldsymbol{s}|\boldsymbol{u})$, yet large $\delta(\boldsymbol{s}|\boldsymbol{u})$). This shows the need for the development of robust nonlinear ICA methods.

We also note that in the no-outliers case (i.e., $p(\boldsymbol{x}|\boldsymbol{u}) = p^\star(\boldsymbol{x}|\boldsymbol{u})$), Assumption (A4) can be written as

$$\log \frac{p^\star(\boldsymbol{x}|\boldsymbol{u})}{c(\boldsymbol{x})e(\boldsymbol{u})} = \boldsymbol{w}(\boldsymbol{u})^\top \boldsymbol{h}(\boldsymbol{x}). \tag{10}$$

To satisfy (10), Hyvärinen et al. [2019] performs binary logistic regression where the log-odds ratio, $\log \frac{p^\star(\boldsymbol{x},\boldsymbol{u})}{p^\star(\boldsymbol{x})p(\boldsymbol{u})}$, is approximated by $\boldsymbol{w}(\boldsymbol{u})^\top \boldsymbol{h}(\boldsymbol{x})$ where $p^\star(\boldsymbol{x}) = \int p^\star(\boldsymbol{x}, \boldsymbol{u}) \mathrm{d}\boldsymbol{u}$. However, Eq.(10) (or Assumption (A4)) is a more general expression than the odds ratio: We do not necessarily need to accurately estimate the noncontaminated log-odds ratio as it is sufficient, in order to perform nonlinear ICA, that the numerator is the conditional density $p^\star(\boldsymbol{x}|\boldsymbol{u})$ or joint density $p^\star(\boldsymbol{x}, \boldsymbol{u})$ up to the product of nonzero scalar functions of $\boldsymbol{x}$ and $\boldsymbol{u}$. This is the key property used in our robust method proposed in Section 4.1.

3.3 Influence of outliers in non-exponential case

We performed a similar contamination analysis as Theorem 1 under the general (non-exponential) conditional independence condition (4) as well. We present the details in Section B of the supplementary material because of space constraints. The conclusion is slightly more complicated yet fundamentally similar as Theorem 1: Estimation in nonlinear ICA can be hampered when either of the four ratios, $\frac{\delta^m(\boldsymbol{s}|\boldsymbol{u})}{p^\star(\boldsymbol{s}|\boldsymbol{u})}$, $\frac{\delta^l(\boldsymbol{s}|\boldsymbol{u})}{p^\star(\boldsymbol{s}|\boldsymbol{u})}$, $\frac{\delta(\boldsymbol{s}|\boldsymbol{u})}{p^\star(\boldsymbol{s}|\boldsymbol{u})}$ and $\frac{\delta^{l,m}(\boldsymbol{s}|\boldsymbol{u})}{p^\star(\boldsymbol{s}|\boldsymbol{u})}$, are very large where $\delta^l(\boldsymbol{s}|\boldsymbol{u}) := \frac{\partial \delta(\boldsymbol{s}|\boldsymbol{u})}{\partial s_l}$ and $\delta^{l,m}(\boldsymbol{s}|\boldsymbol{u}) := \frac{\partial^2 \delta(\boldsymbol{s}|\boldsymbol{u})}{\partial s_l \partial s_m}$ for $l \neq m$. In addition to the basic outlier case above, these ratios might be large when smooth $\delta(\boldsymbol{s}|\boldsymbol{u})$ exists on the tails of $p^\star(\boldsymbol{s}|\boldsymbol{u})$. Therefore, again, it would be useful to develop a robust method in the general non-exponential case as well.

4 Robust contrastive learning

Our goal is to robustify nonlinear ICA methods. In light of the results above, the key is to estimate $\frac{p^\star(\boldsymbol{x}|\boldsymbol{u})}{c(\boldsymbol{x})e(\boldsymbol{u})}$ in (10) in spite of the contamination. (This is the case for the non-exponential family case as well, see Section B in the supplementary material for more details). To this end, this section proposes two robust methods for nonlinear ICA based on the γ-cross entropy [Fujisawa and Eguchi, 2008]. Then, we show that the desired robust estimation is possible by the proposed methods even under heavy contamination by outliers.

Before going to the details, let us clarify the notations as follows: $p(\boldsymbol{x}|\boldsymbol{u})$ denotes the contaminated conditional density of $\boldsymbol{x}$ given $\boldsymbol{u}$ from (5), while $p^\star(\boldsymbol{x}|\boldsymbol{u})$ and $\delta(\boldsymbol{x}|\boldsymbol{u})$ are the (noncontaminated) target and outlier conditional densities, respectively. We can further obtain two marginal densities from $p(\boldsymbol{x}|\boldsymbol{u})$ and $p^\star(\boldsymbol{x}|\boldsymbol{u})$ as $p(\boldsymbol{x}) := \int p(\boldsymbol{x}|\boldsymbol{u})p(\boldsymbol{u})\mathrm{d}\boldsymbol{u}$ and $p^\star(\boldsymbol{x}) := \int p^\star(\boldsymbol{x}|\boldsymbol{u})p(\boldsymbol{u})\mathrm{d}\boldsymbol{u}$. In the rest of this paper except for the influence function analysis, we suppose that $p(\boldsymbol{u})$ is contaminated by an outlier density but the contaminated model is not explicitly defined because a specific form is not required in the analysis of this paper.

4.1 Nonlinear ICA with robust binary classification

Our first method performs nonlinear ICA by solving a binary classification problem under the γ-cross entropy [Fujisawa and Eguchi, 2008, Hung et al., 2018]. Let us express a class label by y, and $y = 1$ and $y = 0$ correspond to datasets $\mathcal{D}$ and $\mathcal{D}_\mathrm{p}$ in (3) which are drawn from $p(\boldsymbol{x}, \boldsymbol{u}|y = 1) = p(\boldsymbol{x}, \boldsymbol{u})$ and $p(\boldsymbol{x}, \boldsymbol{u}|y = 0) = p(\boldsymbol{x})p(\boldsymbol{u})$, respectively. Moreover, symmetric class probabilities are assumed (i.e., $p(y = 0) = p(y = 1) = \frac{1}{2}$). Then, the γ-cross entropy for binary classification is defined as

$$d_\gamma(p(y|\boldsymbol{x}, \boldsymbol{u}), r(\boldsymbol{x}, \boldsymbol{u}); p(\boldsymbol{x}, \boldsymbol{u})) := \\ -\frac{1}{\gamma} \log \iint \sum_{y=0}^{1} \left\{ \frac{r(\boldsymbol{x}, \boldsymbol{u})^{y(\gamma+1)}}{1 + r(\boldsymbol{x}, \boldsymbol{u})^{\gamma+1}} \right\}^{\frac{\gamma}{\gamma+1}} p(y, \boldsymbol{x}, \boldsymbol{u}) \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u}, \tag{11}$$

where $r(\boldsymbol{x}, \boldsymbol{u})$ denotes a model (e.g., a neural network) and positive function. As proven in Fujisawa and Eguchi [2008], the γ-cross entropy has a number of remarkable properties. For instance, $d_\gamma(p(y|\boldsymbol{x}, \boldsymbol{u}), r(\boldsymbol{x}, \boldsymbol{u}); p(\boldsymbol{x}, \boldsymbol{u}))$ approaches to the cross entropy in logistic regression as $\gamma \to 0$. Notably, the γ-cross entropy has a robustness property on parameter estimation in the presence of outliers [Fujisawa and Eguchi, 2008, Kawashima and Fujisawa, 2018]. Next, we show that the robustness property holds in the context of nonlinear ICA.

Robustness in nonlinear ICA: First, we establish the following theorem to understand under what conditions a good estimation is possible for nonlinear ICA even under heavy contamination of outliers:

Theorem 2. *Assume that*

$$\nu := \iint \left\{ \frac{r(\boldsymbol{x},\boldsymbol{u})^{\gamma+1}}{1+r(\boldsymbol{x},\boldsymbol{u})^{\gamma+1}} \right\}^{\frac{\gamma}{\gamma+1}} \epsilon(\boldsymbol{u})\delta(\boldsymbol{x},\boldsymbol{u})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u} \quad (12)$$

is sufficiently small. Then, it holds that

$$\begin{aligned} &d_\gamma(p(y|\boldsymbol{x},\boldsymbol{u}), r(\boldsymbol{x},\boldsymbol{u}); p(\boldsymbol{x},\boldsymbol{u})) + O(\nu) \\ &= J[r(\boldsymbol{x},\boldsymbol{u}); (1-\epsilon(\boldsymbol{u}))p^\star(\boldsymbol{x},\boldsymbol{u}), p(\boldsymbol{x})p(\boldsymbol{u})], \quad (13) \end{aligned}$$

where

$$\begin{aligned} &J[r(\boldsymbol{x},\boldsymbol{u}); (1-\epsilon(\boldsymbol{u}))p^\star(\boldsymbol{x},\boldsymbol{u}), p(\boldsymbol{x})p(\boldsymbol{u})] \\ &:= -\frac{1}{\gamma}\log\left[\frac{1}{2}\iint \left\{ \frac{1}{1+r(\boldsymbol{x},\boldsymbol{u})^{\gamma+1}} \right\}^{\frac{\gamma}{\gamma+1}} p(\boldsymbol{x})p(\boldsymbol{u})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u} \right. \\ &\left. + \frac{1}{2}\iint \left\{ \frac{r(\boldsymbol{x},\boldsymbol{u})^{\gamma+1}}{1+r(\boldsymbol{x},\boldsymbol{u})^{\gamma+1}} \right\}^{\frac{\gamma}{\gamma+1}} (1-\epsilon(\boldsymbol{u}))p^\star(\boldsymbol{x},\boldsymbol{u})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u} \right]. \end{aligned}$$

Furthermore, under the assumption that $p(\boldsymbol{x})$, $p(\boldsymbol{u})$ and $r(\boldsymbol{x},\boldsymbol{u})$ are positive for all $\boldsymbol{x}$ and $\boldsymbol{u}$, $J[r(\boldsymbol{x},\boldsymbol{u}); (1-\epsilon(\boldsymbol{u}))p^\star(\boldsymbol{x},\boldsymbol{u}), p(\boldsymbol{x})p(\boldsymbol{u})]$ is minimized at

$$r^\star(\boldsymbol{x},\boldsymbol{u}) = \frac{(1-\epsilon(\boldsymbol{u}))p^\star(\boldsymbol{x}|\boldsymbol{u})}{p(\boldsymbol{x})}. \quad (14)$$

The proof is deferred to Section C in the supplementary material. Eq.(13) indicates that under the condition that ν is sufficiently small (this condition will be discussed below), minimization of the γ-cross entropy approximately equals to minimization of $J[r(\boldsymbol{x},\boldsymbol{u}); (1-\epsilon(\boldsymbol{u}))p^\star(\boldsymbol{x},\boldsymbol{u}), p(\boldsymbol{x})p(\boldsymbol{u})]$ whose minimizer is given by (14). Crucially, (14) is a special case of the ideal universal approximation condition (10) *without* outliers where $c(\boldsymbol{x}) = p(\boldsymbol{x})$ and $e(\boldsymbol{u}) = 1/(1-\epsilon(\boldsymbol{u}))$. Thus, Theorem 2 implies that we can obtain a consistent estimation result as in (14) almost as if outliers did not exist. Another notable point is that $\epsilon(\boldsymbol{u})$ is never assumed to be small in itself and therefore, heavy contamination of outliers is also within the scope of our method.

Next, we analyse the constant ν in Theorem 2 to understand when it can be considered to be sufficiently small. Let us define the supports of $p^\star(\boldsymbol{s}|\boldsymbol{u})$ and $\delta(\boldsymbol{s}|\boldsymbol{u})$ as $\mathcal{S}_{\boldsymbol{u}}^{p^\star} := \{\boldsymbol{s} \mid p^\star(\boldsymbol{s}|\boldsymbol{u}) > 0\}$ and $\mathcal{S}_{\boldsymbol{u}}^{\delta} := \{\boldsymbol{s} \mid \delta(\boldsymbol{s}|\boldsymbol{u}) > 0\}$, respectively. Then, the following proposition gives an important insight:

Proposition 1. *Let us denote the domains of $\boldsymbol{u}$ and $\boldsymbol{x}$ by $\mathcal{U}$ and $\mathcal{X}$, respectively. We assume that (i) the integrals in ν are defined over $\mathcal{U}$ and $\mathcal{X}$, (ii) data $\boldsymbol{x}$ is generated from* (1) *with an invertible nonlinear mixing function $\boldsymbol{f}$, (iii) $\mathcal{S}_{\boldsymbol{u}}^{p^\star} \cap \mathcal{S}_{\boldsymbol{u}}^{\delta} = \emptyset$, and (iv) $p(\boldsymbol{s}) > 0$ on $\mathcal{S}_{\boldsymbol{u}}^{\delta}$. For $\gamma > 0$,*

$$\nu \leq O\left(\sup_{\boldsymbol{x},\boldsymbol{u}} |r(\boldsymbol{x},\boldsymbol{u}) - r^\star(\boldsymbol{x},\boldsymbol{u})| \right).$$

The proof is given in the supplementary material. The most important condition is $\mathcal{S}_{\boldsymbol{u}}^{p^\star} \cap \mathcal{S}_{\boldsymbol{u}}^{\delta} = \emptyset$, which implies that $p^\star(\boldsymbol{s}|\boldsymbol{u})$ and $\delta(\boldsymbol{s}|\boldsymbol{u})$ are *separated* or nonoverlapping on $\mathcal{S}$. For instance, when $p^\star(\boldsymbol{s}|\boldsymbol{u})$ and $\delta(\boldsymbol{s}|\boldsymbol{u})$ are the uniform densities on $[0,1]^{d_{\mathrm{x}}}$ and $[2,3]^{d_{\mathrm{x}}}$ respectively, their supports are nonoverlapping in this sense.

Therefore, Proposition 1 implies that ν in Theorem 2 can be sufficiently small in the neighborhood of $r^\star(\boldsymbol{x},\boldsymbol{u})$ when $\delta(\boldsymbol{s}|\boldsymbol{u})$ and $p^\star(\boldsymbol{s}|\boldsymbol{u})$ are clearly separated. Such a density separation happens approximately when $\delta(\boldsymbol{s}|\boldsymbol{u})$ is non-zero only in the tails of $p^\star(\boldsymbol{s}|\boldsymbol{u})$, which is indeed typical contamination by outliers. Thus, our condition on ν should be realistic in many practical situations, and Theorem 2 can be expected to hold.

Influence function analysis: Next, we investigate the robustness of our nonlinear ICA method based on the alternative approach of influence functions (IF), which is an established measure in robust statistics [Hampel et al., 2011]. To this end, let us define the following two contaminated density models: With the marginal densities, $\bar{p}(\boldsymbol{x}) = \int \bar{p}(\boldsymbol{x},\boldsymbol{u})\mathrm{d}\boldsymbol{u}$ and $\bar{p}(\boldsymbol{u}) = \int \bar{p}(\boldsymbol{x},\boldsymbol{u})\mathrm{d}\boldsymbol{x}$,

- Contamination model (A):
$$\bar{p}(\boldsymbol{x},\boldsymbol{u}) = (1-\epsilon)p^\star(\boldsymbol{x},\boldsymbol{u}) + \epsilon\bar{\delta}_{\bar{\boldsymbol{x}}}(\boldsymbol{x})p^\star(\boldsymbol{u}),$$
where $p^\star(\boldsymbol{u}) = \int p^\star(\boldsymbol{x},\boldsymbol{u})\mathrm{d}\boldsymbol{u}$, $\epsilon \in [0,1)$ is a contamination ratio and $\bar{\delta}_{\boldsymbol{z}}$ is the Dirac delta function having a point mass at $\boldsymbol{z}$.
- Contamination model (B):
$$\bar{p}(\boldsymbol{x},\boldsymbol{u}) = (1-\epsilon)p^\star(\boldsymbol{x},\boldsymbol{u}) + \epsilon\bar{\delta}_{\bar{\boldsymbol{x}}}(\boldsymbol{x})\bar{\delta}_{\bar{\boldsymbol{u}}}(\boldsymbol{u}).$$

Contamination model (A) indicates only input data $\boldsymbol{x}$ is contaminated by outliers $\bar{\boldsymbol{x}}$, while both input and auxiliary data are contaminated by $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{u}}$ in Contamination model (B).

We suppose that a model $r_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})$ is parameterized by $\boldsymbol{\theta}$, and define $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_\epsilon$ as solutions of the following estimating functions over the (uncontaminated) target and contaminated densities, respectively:

$$\frac{\partial}{\partial\boldsymbol{\theta}} d_\gamma(p^\star(y|\boldsymbol{x},\boldsymbol{u}), r_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u}); p^\star(\boldsymbol{x},\boldsymbol{u}))\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \boldsymbol{0} \quad (15)$$

$$\frac{\partial}{\partial\boldsymbol{\theta}} d_\gamma(\bar{p}(y|\boldsymbol{x},\boldsymbol{u}), r_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u}); \bar{p}(\boldsymbol{x},\boldsymbol{u}))\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_\epsilon} = \boldsymbol{0}, \quad (16)$$

where $p^\star(\boldsymbol{x},\boldsymbol{u}|y=1)=p^\star(\boldsymbol{x},\boldsymbol{u})$, $p^\star(\boldsymbol{x},\boldsymbol{u}|y=0)=p^\star(\boldsymbol{x})p^\star(\boldsymbol{u})$, $\bar{p}(\boldsymbol{x},\boldsymbol{u}|y=1)=\bar{p}(\boldsymbol{x},\boldsymbol{u})$ and $\bar{p}(\boldsymbol{x},\boldsymbol{u}|y=0)=\bar{p}(\boldsymbol{x})\bar{p}(\boldsymbol{u})$. Then, IF is defined by

$$\mathrm{IF}(\bar{\boldsymbol{x}},\bar{\boldsymbol{u}})=\lim_{\epsilon\to 0}\frac{\hat{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}_\epsilon}{\epsilon}. \tag{17}$$

Eq.(17) means that IF measures how $\hat{\boldsymbol{\theta}}$ is influenced by outliers $(\bar{\boldsymbol{x}},\bar{\boldsymbol{u}})$ under the small contamination, and a larger IF implies that $\hat{\boldsymbol{\theta}}$ is more sensitive to outliers.

A desirable property of $\hat{\boldsymbol{\theta}}$ in terms of IF is the *B-robustness*: $\hat{\boldsymbol{\theta}}$ is said to be B-robust when $\sup_{\bar{\boldsymbol{x}},\bar{\boldsymbol{u}}}\|\mathrm{IF}(\bar{\boldsymbol{x}},\bar{\boldsymbol{u}})\|<\infty$ [Hampel et al., 2011]. Another useful property is the *redescending property*, which defined as $\lim_{\|\bar{\boldsymbol{x}}\|,\|\bar{\boldsymbol{u}}\|\to\infty}\|\mathrm{IF}(\bar{\boldsymbol{x}},\bar{\boldsymbol{u}})\|=0$. The redescending property ensures that $\hat{\boldsymbol{\theta}}$ has no influence from even strongly deviated data $\bar{\boldsymbol{x}}$ (and/or $\bar{\boldsymbol{u}}$).

The following proposition implies that our method based on the γ-cross entropy can have the redescending property and be B-robust under certain conditions:

Proposition 2. *Assume that the Hessian matrix of the γ-cross entropy over the contaminated densities (i.e., $d_\gamma(\bar{p}(y|\boldsymbol{x},\boldsymbol{u}),r_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u});\bar{p}(\boldsymbol{x},\boldsymbol{u}))$ in* (16)*) is invertible, and $r_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})$ satisfies*

$$\sup_{\boldsymbol{x},\boldsymbol{u}}\left\|S_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})\frac{\partial\log r_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})^{(\gamma+1)}}{\partial\boldsymbol{\theta}}\right\|<\infty \tag{18}$$

$$\lim_{\|\boldsymbol{x}\|\to\infty}\left[S_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})\frac{\partial\log r_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})^{(\gamma+1)}}{\partial\boldsymbol{\theta}}\right]=\boldsymbol{0}, \tag{19}$$

where $S_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u}):=\{L_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})(1-L_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u}))\}^{\frac{\gamma}{1+\gamma}}$ with $L_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u}):=\frac{1}{1+r_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})^{(\gamma+1)}}$. Then, under Contamination model (A), both the B-robustness and redescending property hold for $\hat{\boldsymbol{\theta}}$. On the other hand, under Contamination model (B), $\hat{\boldsymbol{\theta}}$ is B-robust.

The proof is given in Section E of the supplementary material. Assumptions (18) and (19) are mild in practice because $S_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})$ exponentially and quickly approaches 0 even when $|r_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})|$ diverges as $\|\boldsymbol{x}\|,\|\boldsymbol{u}\|\to\infty$ as in neural networks. Thus, Proposition 2 indicates that our method could have the redescending and B-robustness properties under the contamination model (A) even when $r_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})$ is modelled by a neural network with an unbounded activation function. Furthermore, the B-robustness still holds to the contamination model (B) whose contamination is more complicated than the contamination model (A). Thus, our influence function analysis also supports that the γ-cross entropy is promising for nonlinear ICA in the presence of outliers.

Robust permutation contrastive learning (RPCL): As a special case, we propose a robust variant of permutation contrastive learning (PCL) [Hyvärinen and Morioka, 2017] which we call *robust permutation contrastive learning* (RPCL). The original PCL supposes that sources are temporally dependent (e.g., $\boldsymbol{s}(t)$ and $\boldsymbol{s}(t-1)$ are statistically dependent), and then makes use of the temporal dependencies for nonlinear ICA by regarding past information as the auxiliary variable $\boldsymbol{u}(t)=\boldsymbol{x}(t-1)$.

RPCL estimates a model $r(\boldsymbol{x},\boldsymbol{u})$ based on the following empirical γ-cross entropy for binary classification:

$$\begin{aligned}&\widehat{d}_\gamma(p(y|\boldsymbol{x},\boldsymbol{u}),r(\boldsymbol{x},\boldsymbol{u});p(\boldsymbol{x},\boldsymbol{u}))\\&:=-\frac{1}{\gamma}\log\left[\frac{1}{2T}\sum_{t=1}^{n}\left\{\left(\frac{r(\boldsymbol{x}(t),\boldsymbol{u}(t))^{\gamma+1}}{1+r(\boldsymbol{x}(t),\boldsymbol{u}(t))^{\gamma+1}}\right)^{\frac{\gamma}{\gamma+1}}\right.\right.\\&\qquad\left.\left.+\left(\frac{1}{1+r(\boldsymbol{x}(t),\boldsymbol{u}_{\mathrm{p}}(t))^{\gamma+1}}\right)^{\frac{\gamma}{\gamma+1}}\right\}\right],\end{aligned}$$

where $\boldsymbol{u}_{\mathrm{p}}(t)$ denotes a random permutation of $\boldsymbol{u}(t)$ with respect to t. Based on the universal approximation assumption in Hyvärinen and Morioka [2017, Theorem 1 and Eq.(12)] or Section B in the supplementary material, we restrict a model r as $r(\boldsymbol{x}(t),\boldsymbol{u}(t))=\exp(\sum_{i=1}^{d_{\boldsymbol{x}}}\psi_i(h_i(\boldsymbol{x}(t)),h_i(\boldsymbol{u}(t)))$ $\boldsymbol{h}(\boldsymbol{x})=(h_1(\boldsymbol{x}),\ldots,h_{d_{\boldsymbol{x}}}(\boldsymbol{x}))^\top$ is a neural network. Following Hyvärinen and Morioka [2017], $\psi_i(h_i(\boldsymbol{x}),h_i(\boldsymbol{u}))$ was also modelled by $|a_{i,1}h_i(\boldsymbol{x})+a_{i,2}h_i(\boldsymbol{u})+b_i|-(\bar{a}_ih_i(\boldsymbol{x})+\bar{b}_i)^2+c$, where $a_{i,1},a_{i,2},b_i,\bar{a}_i,\bar{b}_i,c$ are parameters to be estimated from data. A minibatch stochastic gradient method is employed to optimize all parameters.

4.2 Nonlinear ICA with robust multiclass classification

We have considered so far an approach based on binary logistic regression. However, it is also possible to use multinomial logistic regression as done in TCL. From the viewpoint of an auxiliary variable $\boldsymbol{u}$ this can be seen to correspond to a case where the auxiliary variable $u\in\{1,\ldots,K\}$ is one-dimensional and discrete, e.g., class label or time segment label.

Thus, we next propose a second robust method, intended for this special case. To this end, we solve a multiclass classification problem based on the γ-cross entropy:

$$\begin{aligned}&d_\gamma(p(u|\boldsymbol{x}),r(u,\boldsymbol{x});p(\boldsymbol{x}))\\&:=-\frac{1}{\gamma}\log\int\left\{\frac{\sum_{u=1}^{K}r(u,\boldsymbol{x})^\gamma p(u|\boldsymbol{x})}{\left(\sum_{u'=1}^{K}r(u',\boldsymbol{x})^{\gamma+1}\right)^{\frac{\gamma}{\gamma+1}}}\right\}p(\boldsymbol{x})\mathrm{d}\boldsymbol{x},\end{aligned} \tag{20}$$

where we supposed that $p(u=1)=p(u=2)=\cdots=p(u=K)=\frac{1}{K}$. Regarding multiclass classi-

fication, a robustness property similar to what we had in Theorem 2 holds by modifying the above discussion on binary classification (11) or following Kawashima and Fujisawa [2018]. The result is that again when $p^\star(\boldsymbol{s}|u)$ and $\delta(\boldsymbol{s}|u)$ are clearly *separated*, minimization of $d_\gamma(p(u|\boldsymbol{x}), r(u, \boldsymbol{x}); p(\boldsymbol{x}))$ would enable us to estimate $p^\star(\boldsymbol{x}|u)$, which is an ideal estimation result and a special case of $\frac{p^\star(\boldsymbol{x}|u)}{c(\boldsymbol{x})e(u)}$ in (10). Details are given in Section F of the supplementary material.

Robust time contrastive learning (RTCL): As a practical method for such multinomial classification, we propose *robust time contrastive learning* (RTCL) which is a robust variant of TCL [Hyvärinen and Morioka, 2016] based on the γ-cross entropy (20). Both TCL and RTCL are intended for the conditional independent exponential family case (2), and suppose time series data (artificially or manually) divided into K time segments, and the auxiliary variable $u \in \{1, \ldots, K\}$ is the time segment label. RTCL employs the following empirical γ-cross entropy:

$$\widehat{d}_\gamma(p(u|\boldsymbol{x}, \boldsymbol{u}), r(\boldsymbol{x}, u); p(\boldsymbol{x})) := -\frac{1}{\gamma} \log \left[\frac{1}{T} \sum_{t=1}^{T} \frac{\sum_{k=1}^{K} \delta_{u(t),k}\, r(u(t), \boldsymbol{x}(t))^\gamma}{\left(\sum_{u'=1}^{K} r(u', \boldsymbol{x}(t))^{\gamma+1} \right)^{\frac{\gamma}{\gamma+1}}} \right],$$

where $u(t) \in \{1, \ldots, K\}$ are the observations of time-segment labels, and $\delta_{u(t),k}$ denotes the Kronecker delta. Based on the universal approximation assumption (A4) in Theorem 1, for $u = 1, \ldots, K$, we restrict r as $r(\boldsymbol{x}, u) = \exp(\boldsymbol{w}_u^\top \boldsymbol{h}(\boldsymbol{x}) + b_u)$ where $\boldsymbol{h}(\boldsymbol{x})$ denotes nonlinear ICA features modelled by a neural network, and $\boldsymbol{w}_u$ and b_u are parameters for weights and bias, respectively. In practice, all parameters are optimized by a minibatch stochastic gradient method.

4.3 Relation with Hyvärinen et al. [2019]

In order to clarify the relations between the existing nonlinear ICA methods, we make the following remarks. Our main theory (Theorem 2) provides a robustified version of the method in Hyvärinen et al. [2019]. Since PCL [Hyvärinen and Morioka, 2017] can be seen as a special case of Hyvärinen et al. [2019], our theory also leads to a special case called RPCL which robustifies PCL. On the other hand, TCL [Hyvärinen and Morioka, 2016] uses a different framework, multi-class classification, and thus we proposed another method called RTCL that robustifies TCL. It should be noted that while the generative model in Hyvärinen et al. [2019] contains TCL as a special case, the estimation method proposed in Hyvärinen and Morioka [2016] is different and not a special case of the estimation method by Hyvärinen et al. [2019]; that is why our robustified versions are also distinct for TCL and the auxiliary variables method.

In addition to Theorem 1 and Section 3.3, non-robustness of the previous methods in Hyvärinen et al. [2019] can be understood in terms of Theorem 2 analysing the γ-cross entropy as well. Hyvärinen et al. [2019] employ binary logistic regression for nonlinear ICA whose cross entropy is obtained as the limit of $\gamma = 0$ in the γ-cross entropy. When $\gamma = 0$, the robustness condition in Theorem 2 is never satisfied: It can be easily confirmed from the definition (12) that ν is a nonzero constant and cannot be sufficiently small in $\gamma = 0$. Thus, the nonlinear methods in Hyvärinen et al. [2019] can be more sensitive to outliers. Section F in supplementary material includes a similar discussion in the case of multiclass classification: Non-robustness of TCL can be also shown in terms of the γ-cross entropy.

Influence function analysis in Proposition 2 also reveals the outlier weakness of the previous methods. In the limit of $\gamma = 0$ (i.e., logistic regression), a class of models for $r_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{u})$ satisfying (18) and (19) is very limited because necessarily $S_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{u}) = 1$. For instance, when $r_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{u})$ is a neural network with an unbounded activation function, Assumptions (18) and (19) would not hold. This implies that estimation of previous methods with neural networks can be hampered by outliers.

5 Numerical experiments on artificial data

This section numerically investigates the robustness of RTCL and RPCL with comparison to existing nonlinear ICA methods on artificial data.

5.1 Robust time contrastive learning

Data generation, nonlinear ICA methods, evaluation: We slightly modified the experimental setting of TCL[2] in Hyvärinen and Morioka [2016] and experimental details are given in Section G of the supplementary material. Source vectors with time segment length 512 was first generated from (5): Following (2), given a time segment label, the target density $p^\star(\boldsymbol{s}|u)$ was conditionally independent Laplace distributions with means 0 and different scales across time segments. Regarding the outlier density $\delta(\boldsymbol{s}|u)$, two types of densities were used: An independent Laplace distribution, and a mixture of two mean-modulated Gaussians. We set $\epsilon(u) = \epsilon$ for all time segments u. The total numbers of segments and of data samples were $K = 256$ and $T = 512 \times 256$, respectively. The dimensionality of data is $d_{\mathrm{x}} = 10$ in Table 1, while $d_{\mathrm{x}} = 5$ in Table 2. Finally, data $\boldsymbol{x}$ was generated as a nonlinear mixing of the (contaminated) sources by three-layer (Table 1) or two-layer (Table 2) neural networks.

[2] https://github.com/hirosm/TCL

Table 1: RTCL and TCL on artificial data. Averages of the absolute correlations are computed over 10 runs. The outlier densities are the independent Laplace density and the modulated mixture of two Gaussians in the top and bottom panels, respectively. A larger value indicates a better result. The best and comparable methods judged by the t-test at the significance level 5% are described in boldface.

Laplace	TCL	RTCL ($\gamma = 0.1$)	RTCL ($\gamma = 0.3$)	RTCL ($\gamma = 0.5$)	RTCL ($\gamma = 1$)
$\epsilon = 0.01$	0.891(0.009)	0.929(0.022)	0.974(0.009)	**0.981(0.022)**	**0.988(0.023)**
$\epsilon = 0.03$	0.822(0.036)	0.860(0.030)	0.920(0.015)	0.948(0.015)	**0.976(0.015)**
$\epsilon = 0.05$	0.793(0.025)	0.814(0.023)	0.867(0.018)	0.898(0.024)	**0.946(0.024)**
$\epsilon = 0.1$	0.738(0.034)	0.768(0.026)	0.814(0.009)	0.848(0.008)	**0.895(0.013)**
Gaussian	TCL	RTCL ($\gamma = 0.1$)	RTCL ($\gamma = 0.3$)	RTCL ($\gamma = 0.5$)	RTCL ($\gamma = 1$)
$\epsilon = 0.01$	0.952(0.007)	0.981(0.008)	**0.990(0.007)**	**0.992(0.007)**	**0.993(0.008)**
$\epsilon = 0.03$	0.872(0.009)	0.904(0.008)	0.946(0.006)	0.962(0.007)	**0.978(0.008)**
$\epsilon = 0.05$	0.852(0.010)	0.855(0.010)	0.905(0.007)	0.932(0.006)	**0.957(0.007)**
$\epsilon = 0.1$	0.824(0.012)	0.815(0.015)	0.836(0.014)	0.871(0.017)	**0.909(0.025)**

Table 2: RTCL and iVAE on artificial data. Averages of the absolute correlations are computed over 10 runs.

Laplace	iVAE	RTCL ($\gamma = 1$)
$\epsilon = 0$	**0.931(0.055)**	**0.969(0.040)**
$\epsilon = 0.05$	0.844(0.081)	**0.965(0.020)**
$\epsilon = 0.1$	0.800(0.108)	**0.924(0.035)**

ICA features $\boldsymbol{h}(\boldsymbol{x})$ both in RTCL and TCL were modelled by a three layer neural network where the number of hidden units was $4d_{\mathrm{x}}$, but the final layer was d_{x}. ℓ_2 regularization was employed with the regularization parameter 10^{-4}. All parameters were optimized by the Adam optimizer [Kingma and Ba, 2015]. We also applied iVAE [Khemakhem et al., 2019] to the artificial data, which is a nonlinear ICA method based on the variational encoder. The performance was measured by averages of the absolute value of the Pearson correlation coefficient to the test sources without outliers.

Results: The top panel in Table 1 quantitatively indicates that RTCL is more robust against outliers than TCL. As the contamination ratio ϵ increases, the performance of TCL deteriorates. On the other hand, RTCL keeps high-correlation values even for larger γ. When the outlier density $\delta(\boldsymbol{x}|u)$ is the mixture of two mean-modulated Gaussians, RTCL still performs well (bottom panel in Table 1). Furthermore, Table 2 shows that iVAE is also sensitive to outliers, while RTCL reliably recovers the sources. This is presumably because iVAE is also related to MLE, which is sensitive to outliers.

5.2 Robust permutation contrastive learning

Data generation, nonlinear ICA methods, evaluation: We followed the experimental setting of PCL in Hyvärinen and Morioka [2017] and details can be seen in Section G of the supplementary material. First, the temporally dependent T sources were generated from $\log p^{\star}(\boldsymbol{s}(t)|\boldsymbol{s}(t-1)) = -\sum_{i=1}^{d_{\mathrm{x}}} |s_i(t) - \rho s_i(t-1)| + C$ where C denotes a constant and the auto-regressive coefficient ρ was fixed at 0.7. The total number of sources was $T = 65536$. Then, we randomly replaced the sources by outliers based on a constant contamination ratio ϵ, which were generated from the independent Laplace density. Data $\boldsymbol{x}$ was generated as the nonlinear mixing of the sources with outliers by a three-layer neural networks.

ICA features $\boldsymbol{h}(\boldsymbol{x})$ both in RPCL and PCL were modelled by a three-layer neural network as in the experiments of RTCL. We optimized the parameters in RPCL and PCL using the Adam optimizer. We used the same performance was measured as previous experiments.

Results: Table 3 clearly shows that the correlation for PCL quickly decreases as the contaminating ratio ϵ increases. On the other hand, RPCL works significantly better than PCL even for large ϵ. Thus, our methods based on the γ-cross entropy are promising.

6 Application to causal discovery of Hippocampal fMRI data

To demonstrate its applicability on a realworld dataset, we apply RTCL to causal discovery [Pearl, 2000] on resting-state fMRI data, which often contains outliers due to measurement issues such as head movement and variability in vascular health across a cohort of subjects [Poldrack et al., 2011]. Our dataset corresponds to resting state fMRI data collected from a single subject (caucasian male, 45 years old) over 84 successive days [Poldrack et al., 2015]. Here, each day is treated as

Table 3: RPCL and PCL on artificial data. Averages of the absolute correlations are computed over 10 runs.

	PCL	RPCL ($\gamma = 0.5$)	RPCL ($\gamma = 1$)	RPCL ($\gamma = 5$)	RPCL ($\gamma = 10$)
$\epsilon = 0.01$	0.917(0.028)	**0.935(0.010)**	**0.942(0.023)**	**0.934(0.026)**	0.911(0.027)
$\epsilon = 0.05$	0.904(0.022)	**0.917(0.015)**	**0.926(0.008)**	**0.932(0.024)**	0.899(0.030)
$\epsilon = 0.1$	0.854(0.053)	0.884(0.034)	**0.888(0.029)**	**0.912(0.026)**	0.886(0.023)
$\epsilon = 0.15$	0.803(0.058)	0.819(0.056)	**0.838(0.048)**	**0.851(0.056)**	**0.866(0.031)**

a distinct experimental condition. Section H in the supplementary material visualizes the presence of outliers in the time series data for the Parahippocampal brain region.

We follow a nonlinear-ICA-based method for causal discovery [Monti et al., 2019]. Let us consider the problem of causal discovery for bivariate data $\boldsymbol{x} = (x_1, x_2)^\top$. The goal of causal discovery is to determine whether x_1 causes x_2 or x_2 causes x_1 (i.e., $x_1 \to x_2$ or $x_2 \to x_1$), or to conclude that no acyclic causal relation exists. If the true causal direction is $x_1 \to x_2$, the nonlinear structural equation model (SEM) [Pearl, 2000] can be written as $x_1 = f_1(n_1)$ and $x_2 = f_2(x_1, n_2)$, where n_1 and n_2 are latent disturbances and assumed to be statistically independent each other. As discussed in Monti et al. [2019], the nonlinear SEM above has a clear connection to the data generative model (1) in nonlinear ICA. Roughly, the disturbance variables $(n_1, n_2)^\top$ in SEM correspond to the latent sources $(s_1, s_2)^\top$ in ICA up to their permutation. Thus, regarding the recovered sources by nonlinear ICA as estimates of $(n_1, n_2)^\top$, we could determine the causal direction by performing a series of independence tests with the observations of $\boldsymbol{x} = (x_1, x_2)^\top$. For instance, under the assumption that the true causal direction is $x_1 \to x_2$, we need to verify that $x_1 \perp\!\!\!\perp n_2$ while $x_1 \not\perp\!\!\!\perp n_1$, $x_2 \not\perp\!\!\!\perp n_1$ and $x_2 \not\perp\!\!\!\perp n_2$ [Monti et al., 2019, Property 1] by applying some independent test where $\perp\!\!\!\perp$ (or $\not\perp\!\!\!\perp$) denotes statistical independence (or dependence). Here, we employed Hilbert-Schmidt independence criteria [Gretton et al., 2005] for independence test. This approach for bivariate data can be extended to multivariate causal discovery by using a traditional constraint-based method such as the PC algorithm.

Fig. 1 shows the causal structures obtained via TCL by Monti et al. [2019], and RTCL. Both methods used a five layer neural network. Blue arrows denote edges which are plausible given the anatomical connectivity, while red arrows are not compatible with the known anatomical structure. We note that in the case of RTCL, the erroneous edges (highlighted in red) actually correspond to indirect causal effects. For example, see the edge between the Cornu Ammonis 1 (CA_1) node and the entorhinal cortex (ERc) node. While a direct connection between these nodes is anatomically implausible, there

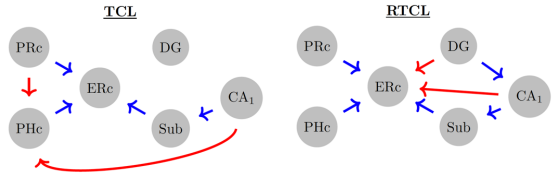

Figure 1: Estimated directed acyclic graphs based on TCL (left panel) and RTCL (right panel, $\gamma = 2.5$). For RTCL, the γ value was selected based on classification accuracy for validation data.

is an indirect effect which is mediated by the subiculum (Sub) node. This is in stark contrast with the results provided by TCL, where erroneous edges (highlighted in red) are not compatible with the anatomical structure (e.g., the TCL edge between CA_1 and PHc cannot be explained as an indirect causal effect).

A similar experiment was performed by using iVAE, which also led to improvements in the recovery of the associated DAG [Khemakhem et al., 2019]. To compare with those results, we note that for RTCL, both erroneous edges are directed to the entorhinal cortex (ERc) region, which serves as the main interface between neocortex (PRc, PHc, ERc) and hippocampal (CA_1, DG, Sub, ERc) regions. Thus, ERc region might cause these erroneous edges in RTCL to connect the two subfields, implying the presence of possible cyclic associations. In contrast, iVAE incorrectly inferred a causal effect from CA_1 to DG which is not anatomically plausible [Khemakhem et al., 2019, Fig.4].

7 Conclusion[3]

We first analyzed the behaviour of the nonlinear ICA estimators given by Hyvärinen et al. [2019] in the presence of outliers, and then proposed two robust methods for nonlinear ICA. We showed by theoretical analysis that our methods have robustness properties in the context of nonlinear ICA. The robustness was further empirically shown in simulations, and applicability to real-data was also demonstrated through causal discovery.

[3]The acknowledgements are included in the supplementary material.

References

S. Amari. *Information Geometry and Its Applications*. Springer, 2016.

A. Basu, I. Harris, N. Hjort, and M. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, 1998.

A. Cichocki and S. Amari. Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, 2010.

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

H. Fujisawa and S. Eguchi. Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99(9):2053–2081, 2008.

A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77, 2005.

F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust statistics: the approach based on influence functions*. John Wiley & Sons, 2011.

S. Harmeling, A. Ziehe, M. Kawanabe, and K.-R. Müller. Kernel-based nonlinear blind source separation. *Neural Computation*, 15(5):1089–1124, 2003.

H. Hung, Z.-Y. Jou, and S.-Y. Huang. Robust mislabel logistic regression without modeling mislabel probabilities. *Biometrics*, 74(1):145–154, 2018.

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

A. Hyvärinen and H. Morioka. Nonlinear ICA of temporally dependent stationary sources. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54, pages 460–469. PMLR, 2017.

A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4–5):411–430, 2000.

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. E. Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *Proceedings of the 22th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89, pages 859–868, 2019.

T. Kawashima and H. Fujisawa. On difference between two types of γ-divergence for regression. *arXiv:1805.06144*, 2018.

I. Khemakhem, D. P. Kingma, R. P. Monti, and A. Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. *Arxiv*, 2019.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, pages 1–15, 2015.

G. Larsson, M. Maire, and G. Shakhnarovich. Colorization as a proxy task for visual understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6874–6883, 2017.

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 *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 4114–4124. PMLR, 2019.

R. P. Monti, K. Zhang, and A. Hyvarinen. Causal discovery with general non-linear relationships using nonlinear ICA. *35th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2019.

M. Noroozi and P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision (ECCV)*, pages 69–84. Springer, 2016.

A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018.

J. Pearl. *Causality: models, reasoning and inference*, volume 29. Springer, 2000.

R. A. Poldrack, J. A. Mumford, and T. E. Nichols. *Handbook of functional MRI data analysis*. Cambridge University Press, 2011.

R. A. Poldrack et al. Long-term neural and physiological phenotyping of a single human. *Nature communications*, 6:8885, 2015.

H. Sprekeler, T. Zito, and L. Wiskott. An extension of slow feature analysis for nonlinear blind source separation. *Journal of machine learning research*, 15:921–947, 2014.

L. Wasserman. *All of nonparametric statistics*. Springer, 2006.

L. Wiskott and T. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14(4):715–770, 2002.