Representation Learning for Maximization of MI, Nonlinear ICA and Nonlinear Subspaces with Robust Density Ratio Estimation

Hiroaki Sasaki   HSASAKI@FUN.AC.JP
Department of Complexity and Intelligent Systems
Future University Hakodate
Hokkaido, Japan

Takashi Takenouchi   T-TAKENOUCHI@GRIPS.AC.JP
National Graduate Institute for Policy Studies
RIKEN AIP
Tokyo, Japan

Editor: Stefan Harmeling

Abstract

Unsupervised representation learning is one of the most important problems in machine learning. A recent promising approach is contrastive learning: A feature representation of data is learned by solving a *pseudo* classification problem where class labels are automatically generated from unlabelled data. However, it is not straightforward to understand what representation contrastive learning yields through the classification problem. In addition, most of practical methods for contrastive learning are based on the maximum likelihood estimation, which is often vulnerable to the contamination by outliers. In order to promote the understanding to contrastive learning, this paper first theoretically shows a connection to maximization of mutual information (MI). Our result indicates that density ratio estimation is necessary and sufficient for maximization of MI under some conditions. Since popular objective functions for classification can be regarded as estimating density ratios, contrastive learning related to density ratio estimation can be interpreted as maximizing MI. Next, in terms of density ratio estimation, we establish new recovery conditions for the latent source components in nonlinear independent component analysis (ICA). In contrast with existing work, the established conditions include a novel insight for the dimensionality of data, which is clearly supported by numerical experiments. Furthermore, inspired by nonlinear ICA, we propose a novel framework to estimate a nonlinear subspace for lower-dimensional latent source components, and some theoretical conditions for the subspace estimation are established with density ratio estimation. Motivated by the theoretical results, we propose a practical method through outlier-robust density ratio estimation, which can be seen as performing maximization of MI, nonlinear ICA or nonlinear subspace estimation. Moreover, a sample-efficient nonlinear ICA method is also proposed based on a variational lower-bound of MI. Then, we theoretically investigate outlier-robustness of the proposed methods. Finally, we numerically demonstrate usefulness of the proposed methods in nonlinear ICA and through application to a downstream task for linear classification.

Keywords: representation learning, contrastive learning, density ratio estimation, maximization of MI, nonlinear ICA, nonlinear subspace estimation, outlier-robustness

1. Introduction

Unsupervised representation learning has been a long-term issue in machine learning (Raina et al., 2007; Bengio et al., 2013). In contrast with supervised learning, a feature representation of data is learned only from unlabelled data. Since the acquisition cost of unlabelled data is often low, the learned representation would be useful particularly when the number of data is very limited in a downstream task (e.g., supervised tasks). Successful examples include video classification (Wang and Gupta, 2015; Arandjelovic and Zisserman, 2017; Sun et al., 2019), motion capture (Tung et al., 2017), and natural language processing (Peters et al., 2018; Devlin et al., 2019). More examples can be found in recent review papers for representation learning (Jing and Tian, 2020; Liu et al., 2020).

Recently, a number of methods for unsupervised representation learning have been proposed, and most of them can be roughly divided into three categories based on variational autoencoder (VAE) (Kingma and Welling, 2014; Higgins et al., 2017; Khemakhem et al., 2020), generative adversarial network (GAN) (Goodfellow et al., 2014; Chen et al., 2016) and contrastive learning (Dosovitskiy et al., 2014; Doersch et al., 2015; Noroozi and Favaro, 2016). VAE learns a useful representation of data by maximizing a tractable lower-bound of the likelihood function. However, VAE makes a restrictive prior assumption on densities such as the Gaussian assumption. The GAN approach learns both a representation and generator of data, but it requires to solve a max-min optimization problem, which could be numerically unstable. On the other hand, contrastive learning does not make strong assumptions on densities, and simply solves a straightforward optimization problem.

In order to learn a feature representation of data, contrastive learning solves a *pseudo* classification problem where class labels are automatically generated from unlabelled data. For example, Arandjelovic and Zisserman (2017) make a dataset for binary classification from unlabelled video data: The positive labels are assigned to pairs of video frames and audio clips taken from the same video, while the negative labels are given to pairs from different videos. Then, representations of both video frames and audio clips are learned by solving the binary classification problem. Misra et al. (2016) also solve a binary classification problem for unsupervised representation learning where the positive data is temporally consecutive video frames in a video, but the negative one is temporally shuffled frames in the same video. This learning scheme based on classification problems is also called *self-supervised learning* (SSL). See a recent review article for SSL (Liu et al., 2020) and references therein. However, it is not straightforward to understand what representation contrastive learning yields by solving classification problems.

Contrastive learning has been justified as being closely related to a classical framework so called *maximization of mutual information* (MI) (Linsker, 1989; Bell and Sejnowski, 1995). This comes from an intuitive idea that MI must be high if a classifier can accurately distinguish positive samples drawn from the joint density (e.g., for image frames and audio clips) and negative ones drawn from the product of the marginal densities (Belghazi et al., 2018; Hjelm et al., 2019; Tschannen et al., 2019). However, this intuitive idea is rather superficial because it is not necessarily obvious why solving the classification problem leads to maximization of MI. A recent promising approach is based on maximization of tractable variational lower-bounds of MI and also related to contrastive learning (Nguyen et al., 2008; Belghazi et al., 2018; van den Oord et al., 2018), but the original purpose of maximizing the variational lower-bounds is *estimation* of MI, not necessarily maximization of MI. More rigorous justification would be preferable to understand how and when contrastive learning can be regarded as performing maximization of MI.

Contrastive learning has been employed in recent algorithms of *nonlinear independent component analysis* (ICA) (Hyvärinen and Morioka, 2016, 2017; Hyvärinen et al., 2019), which is a solid framework for unsupervised representation learning. Nonlinear ICA rigorously defines a generative model where the input data is assumed to be observed as a general nonlinear mixing of the latent source components. Then, the problem is to recover the source components from the observations of input data. Regarding the linear ICA where the mixing function is linear, the recovery conditions are well-established and the important condition is mutual independence of the source components (Comon, 1994). On the other hand, the problem of nonlinear ICA has been proved to be fundamentally ill-posed under the same independence condition because there exist an infinite number of decompositions of a random vector into mutually independent variables (Hyvärinen and Pajunen, 1999; Locatello et al., 2019). Very recently, novel recovery conditions for nonlinear ICA have been established (Sprekeler et al., 2014; Hyvärinen and Morioka, 2016, 2017; Hyvärinen et al., 2019). The main idea is to introduce additional data called the *complementary data* in this paper, and alternative condition to mutual independence is conditional independence of the source components given complementary data. An example of the complementary data is time segment labels obtained by dividing time series data into a number of time segments (Hyvärinen and Morioka, 2016), while Hyvärinen and Morioka (2017) employed the history of time-series data as complementary data. With the complementary data, contrastive learning based on the logistic regression has been performed for nonlinear ICA.

Most of practical methods in contrastive learning have been based on maximum likelihood estimation (MLE): In logistic regression, the conditional probability (i.e., posterior probability) is estimated by MLE, while maximizing a variational lower-bound of MI can be interpreted as performing MLE as we show later. MLE has a number of excellent properties such the asymptotic efficiency (Wasserman, 2006), but it is known not to be robust against outliers: If data is contaminated by outliers, MLE often suffers from a severe bias. Thus, estimation of existing methods for maximization of MI and nonlinear ICA might be magnified by outliers. This is a serious problem because outliers are observed in many practical situations. For instance, the contamination of outliers has been an important issue in functional MRI data (Poldrack, 2012) to which ICA methods have been applied (Monti et al., 2020).

This paper first shows that unsupervised representation learning on three frameworks can be performed through density ratio estimation. Previous work has already discussed that contrastive learning can be seen as density ratio estimation when popular objective functions are used for classification such as the cross entropy (Nguyen et al., 2008; Belghazi et al., 2018; van den Oord et al., 2018). Our primary contributions are to perform theoretical analysis on the three frameworks in terms of density ratio estimation. The first framework to be tackled is maximization of MI. We show that density ratio estimation is necessary and sufficient for maximization of MI under some conditions. This result supports the intuitive belief above between contrastive learning and maximization of MI, and clarifies when and how contrastive learning can be considered as performing maximization of MI. The second framework is nonlinear ICA, where we provide two new proofs for source recovery in terms of density ratio estimation. The key point is that the recovery conditions in both proofs include a novel insight, which has not been seen in previous work of nonlinear ICA (Hyvärinen et al., 2019): The dimensionality of complementary data is an important factor for source recovery. This insight is clearly supported by experimental results that the latent sources are more accurately recovered as the dimensionality of complementary data increases. Furthermore, one of the proofs can be seen as a generalization of Hyvärinen and Morioka (2017). The third

framework is *nonlinear subspace estimation* proposed in this paper. This framework is inspired by nonlinear ICA, and aimed at estimating a nonlinear subspace of lower-dimensional latent source components. To this end, we propose a novel generative model where data is generated as a nonlinear mixing of lower-dimensional source components and nuisance variables. Unlike nonlinear ICA, the source components are not necessarily assumed to be conditionally independent, and thus the proposed generative model is more general than nonlinear ICA in terms of the conditional independence. Moreover, as in nonlinear ICA, with density ratio estimation, we establish theoretical conditions that a nonlinear subspace of the lower-dimensional source components, which is separated from the nuisance variables, can be estimated.

Motivated by the density-ratio view, we propose an outlier-robust method for unsupervised representation learning. The proposed method performs density ratio estimation based on a robust alternative to the KL-divergence called the γ-*divergence* (Fujisawa and Eguchi, 2008), which has a favorable robustness property expressed as the *strong robustness* (Cichocki and Amari, 2010; Amari, 2016): The latent bias caused from outliers can be small even in the case of heavy contamination of outliers. Furthermore, we develop a new method of nonlinear ICA by applying an existing variational lower-bound of MI (Belghazi et al., 2018). Then, we theoretically investigate outlier-robustness of these methods. Finally, we numerically demonstrate that the proposed method based on the γ-divergence is very robust against outliers both in nonlinear ICA and a downstream task for linear classification, while the nonlinear ICA method based on the variational lower-bound is experimentally shown to perform better than existing ICA methods when the number of data is small.

This paper is organized as follows[1]: Section 2 formulates the problem of density ratio estimation for unsupervised representation learning, and reviews existing works for contrastive learning, variational MI estimation and nonlinear ICA. Section 3 theoretically shows that unsupervised representation learning on the three frameworks can be performed through density ratio estimation, and discusses theoretical contributions in each of the frameworks. Section 4 proposes two practical methods for unsupervised representation learning and theoretically analyzes them in terms of outlier-robustness. Section 5 numerically demonstrates usefulness of the proposed methods in nonlinear ICA and a downstream task for linear classification. Section 6 concludes this paper.

2. Problem formulation and background

This section first formulates the problem of unsupervised representation learning based on density ratio estimation, and then reviews existing works for contrastive learning, variational estimation of mutual information and nonlinear independent component analysis.

2.1 Problem formulation of density ratio estimation for representation learning

Suppose that we are given T pairs of two data samples drawn from the joint distribution with density $p(\boldsymbol{x}, \boldsymbol{u})$:

$$\mathcal{D} := \left\{ (\boldsymbol{x}(t)^\top, \boldsymbol{u}(t)^\top)^\top \mid \boldsymbol{x}(t) = (x_1(t), \ldots, x_{D_\mathrm{x}}(t))^\top, \boldsymbol{u}(t) = (u_1(t), \ldots, u_{D_\mathrm{u}}(t))^\top \right\}_{t=1}^{T}, \quad (1)$$

where $\boldsymbol{x}(t)$ and $\boldsymbol{u}(t)$ denote the t-th observations of $\boldsymbol{x}$ and $\boldsymbol{u}$, respectively. Throughout the paper, we call $\boldsymbol{x}$ and $\boldsymbol{u}$ as *input* and *complementary* data, respectively. Our primary goal is to estimate

---

1. A very preliminary version of this paper was published in Sasaki et al. (2020).

Table 1: List of notations used in Section 2.

$\boldsymbol{x} \in \mathbb{R}^{Dx}$	Input data
$\boldsymbol{u} \in \mathbb{R}^{D_{\mathrm{u}}}$	Complementary data
D_{x}	Dimensionality of input data
D_{u}	Dimensionality of complementary data
$p(\boldsymbol{x}, \boldsymbol{u})$	Joint probability density function of $\boldsymbol{x}$ and $\boldsymbol{u}$
$p(\boldsymbol{x})$	Marginal probability density function of $\boldsymbol{x}$
$p(\boldsymbol{u})$	Marginal probability density function of $\boldsymbol{u}$
$\perp$	Statistical independence
$\boldsymbol{u}^* \in \mathbb{R}^{D_{\mathrm{u}}}$	$\boldsymbol{u}^* \sim p(\boldsymbol{u})$ and $\boldsymbol{u}^* \perp \boldsymbol{x}$
$\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}) : \mathbb{R}^{D_{\mathrm{x}}} \to \mathbb{R}^{d_{\mathrm{x}}}$	Representation function of $\boldsymbol{x}$
$\boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u}) : \mathbb{R}^{D_{\mathrm{u}}} \to \mathbb{R}^{d_{\mathrm{u}}}$	Representation function of $\boldsymbol{u}$
d_{x}	Dimensionality of representation function $\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x})$
d_{u}	Dimensionality of representation function $\boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u})$
$r(\boldsymbol{x}, \boldsymbol{u})$	Model to approximate $\log \frac{p(\boldsymbol{x}, \boldsymbol{u})}{p(\boldsymbol{x})p(\boldsymbol{u})}$
$\psi : \mathbb{R}^{d_{\mathrm{x}}} \times \mathbb{R}^{d_{\mathrm{u}}} \to \mathbb{R}$	Function in $r(\boldsymbol{x}, \boldsymbol{u})$
$a : \mathbb{R}^{d_{\mathrm{x}}} \to \mathbb{R}$	Function in $r(\boldsymbol{x}, \boldsymbol{u})$
$b : \mathbb{R}^{d_{\mathrm{u}}} \to \mathbb{R}$	Function in $r(\boldsymbol{x}, \boldsymbol{u})$
J_{LR}	Cross entropy in logistic regression

$\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}) = (h_{\mathrm{x},1}(\boldsymbol{x}), \ldots, h_{\mathrm{x},d_{\mathrm{x}}}(\boldsymbol{x}))^{\top}$ called a *representation function* of $\boldsymbol{x}$ such that the logarithmic ratio of the joint density to the product of the marginal densities $p(\boldsymbol{x})$ and $p(\boldsymbol{u})$,

$$\log \frac{p(\boldsymbol{x}, \boldsymbol{u})}{p(\boldsymbol{x})p(\boldsymbol{u})}, \tag{2}$$

is accurately approximated up to a constant under the following basic form of a model $r(\boldsymbol{x}, \boldsymbol{u})$:

$$r(\boldsymbol{x}, \boldsymbol{u}) := \psi(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}), \boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u})) + a(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x})) + b(\boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u})), \tag{3}$$

where $\boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u}) = (h_{\mathrm{u},1}(\boldsymbol{u}), \ldots, h_{\mathrm{u},d_{\mathrm{u}}}(\boldsymbol{u}))^{\top}$ is a representation function of $\boldsymbol{u}$, and $\psi(\cdot, \cdot)$, and $a(\cdot)$ and $b(\cdot)$ are scalar functions. For nonlinear ICA, $\psi(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}), \boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u}))$ is slightly modified: In Theorem 2, $\psi(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}), \boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u}))$ is expressed by the sum of elementwise functions with respect to $\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x})$ (not $\boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u})$) as $\psi(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}), \boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u})) = \sum_{i=1}^{d_{\mathrm{x}}} \psi_i(h_{\mathrm{x},i}(\boldsymbol{x}), \boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u}))$. All of $\boldsymbol{h}_{\mathrm{u}}$, ψ, a and b are also estimated from $\mathcal{D}$. Here, we assume that the dimensionalities of the representation functions are smaller than or equal to ones of input and complementary data, i.e., $d_{\mathrm{u}} \leq D_{\mathrm{u}}$ and $d_{\mathrm{x}} \leq D_{\mathrm{x}}$. Furthermore, in order to simplify the review of existing works and our theoretical results in Section 3, it is supposed that data samples is *not* contaminated by outliers, while we propose a practical method and theoretically investigate its robustness under the contamination of outliers in Section 4.

Function $\psi(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}), \boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u}))$ in (3) takes a role of capturing the statistical dependencies between $\boldsymbol{x}$ and $\boldsymbol{u}$ with dimensionality reduction, and has been modeled previously such as $\psi(\boldsymbol{h}_{\mathrm{x}}, \boldsymbol{h}_{\mathrm{u}}) = \boldsymbol{h}_{\mathrm{x}}^{\top}\boldsymbol{h}_{\mathrm{u}}$ (Bachman et al., 2019) and $\psi(\boldsymbol{h}_{\mathrm{x}}, \boldsymbol{h}_{\mathrm{u}}) = \boldsymbol{h}_{\mathrm{x}}^{\top}\boldsymbol{W}\boldsymbol{h}_{\mathrm{u}}$ (van den Oord et al., 2018; Tian et al., 2019) where $\boldsymbol{W}$ is a d_{x} by d_{u} matrix. More generally, $\psi(\boldsymbol{h}_{\mathrm{x}}, \boldsymbol{h}_{\mathrm{u}})$ can be modeled by a feedforward neural network (Arandjelovic and Zisserman, 2017). Scalar functions of $a(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}))$ and $b(\boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u}))$ express any functions in the log-density ratio (2) that depend only on either $\boldsymbol{x}$ or $\boldsymbol{u}$. Section 3 often

5

assumes that there exist functions $\psi^\star$, $\boldsymbol{h}_\mathrm{x}^\star$, $\boldsymbol{h}_\mathrm{u}^\star$, $a^\star$ and $b^\star$ such that the logarithmic density ratio is universally approximated at $\psi = \psi^\star$, $\boldsymbol{h}_\mathrm{x} = \boldsymbol{h}_\mathrm{x}^\star$, $\boldsymbol{h}_\mathrm{u} = \boldsymbol{h}_\mathrm{u}^\star$, $a = a^\star$ and $b = b^\star$ in (3) as follows:

$$\log \frac{p(\boldsymbol{x}, \boldsymbol{u})}{p(\boldsymbol{x})p(\boldsymbol{u})} = \psi^\star(\boldsymbol{h}_\mathrm{x}^\star(\boldsymbol{x}), \boldsymbol{h}_\mathrm{u}^\star(\boldsymbol{u})) + a^\star(\boldsymbol{h}_\mathrm{x}^\star(\boldsymbol{x})) + b^\star(\boldsymbol{h}_\mathrm{u}^\star(\boldsymbol{u})).$$

This universal approximation assumption would be realistic when the log-density ratio is a continuous function and neural networks are employed for modelling ψ, $\boldsymbol{h}_\mathrm{x}$, $\boldsymbol{h}_\mathrm{u}$, a and b (Hornik, 1991). Section 4 proposes practical methods to approximate $\psi^\star$, $\boldsymbol{h}_\mathrm{x}^\star$, $\boldsymbol{h}_\mathrm{u}^\star$, $a^\star$ and $b^\star$ from data samples.

The approach of using the complementary data is recently becoming more popular as in multi-view learning (Li et al., 2018). Furthermore, it might not be so expensive to obtain complementary data samples, but they are rather generated from a single unlabeled dataset in many practical situations. Let us list the following examples:

- Denoising autoencoder makes complementary data samples $\boldsymbol{u}(t)$ by injecting noises to input data samples $\boldsymbol{x}(t)$ (Vincent et al., 2010).

- $\boldsymbol{x}(t)$ and $\boldsymbol{u}(t)$ can be drawn by dividing a single video data into image (i.e., a video frame) and sound data samples at each time index t, respectively (Arandjelovic and Zisserman, 2017).

- Regarding time series data $\boldsymbol{x}(t)$ including video data, past input data is an example of the complementary data (e.g., $\boldsymbol{u}(t) = \boldsymbol{x}(t-1)$) (Misra et al., 2016; Hyvärinen and Morioka, 2017).

- Suppose that $\boldsymbol{x}(t)$ are image patches extracted from a large single image where the index t conveys some positional information of patches. Then, pairs of image patches at t and $t'(\neq t)$ can be used as input and complementary data samples, that is, $\boldsymbol{u}(t) = \boldsymbol{x}(t')$ (Noroozi and Favaro, 2016).

- Clustering labels to input data samples $\boldsymbol{x}(t)$ can be used as the complementary data samples $\boldsymbol{u}(t)$ (Caron et al., 2018).

- In Hjelm et al. (2019), $\boldsymbol{x}(t)$ is an image, while $\boldsymbol{u}(t)$ is a smaller image patch extracted from the single image $\boldsymbol{x}(t)$.

Density ratio estimation is useful particularly when neural networks are used for $r(\boldsymbol{x}, \boldsymbol{u})$ because the density ratio is invariant under any invertible transformations or reparametrizations of $\boldsymbol{x}$ and/or $\boldsymbol{u}$. This invariant property has been exploited by noise contrastive estimation as well (Gutmann and Hyvärinen, 2012). In addition to the practical usefulness, in this paper, we theoretically show that three frameworks for unsupervised representation learning can be performed by density ratio estimation.

2.2 Contrastive learning and variational estimation of mutual information

In order to estimate the representation functions, contrastive learning solves a classification problem where class labels are automatically generated from unlabelled data. A common setting is based on the following two datasets:

$$\mathcal{D}_+ := \{(\boldsymbol{x}(t), \boldsymbol{u}(t)\}_{t=1}^T \sim p(\boldsymbol{x}, \boldsymbol{u}) \quad \text{vs.} \quad \mathcal{D}_- := \{(\boldsymbol{x}(t), \boldsymbol{u}^*(t))\}_{t=1}^T \sim p(\boldsymbol{x})p(\boldsymbol{u}), \qquad (4)$$

where $\boldsymbol{u}^*(t)$ is a D_{u}-dimensional data vector sampled from the marginal density of $\boldsymbol{u}$, and thus independent to $\boldsymbol{x}(t)$. In practice, $\boldsymbol{u}^*(t)$ can be generated by randomly shuffling $\boldsymbol{u}(t)$ with respect to t under the i.i.d. assumption. Interestingly, the random shuffling has been heuristically used in a number of previous works (Misra et al., 2016; Lee et al., 2017). One of the most popular objective functions in contrastive learning is the following cross entropy for binary classification used in logistic regression:

$$J_{\mathrm{LR}}(r) := -E_{\mathrm{xu}}\left[\log \frac{e^{r(\boldsymbol{X},\boldsymbol{U})}}{1 + e^{r(\boldsymbol{X},\boldsymbol{U})}}\right] - E_{\mathrm{x \times u}}\left[\log \frac{1}{1 + e^{r(\boldsymbol{X},\boldsymbol{U})}}\right], \tag{5}$$

where E_{xu} and $E_{\mathrm{x \times u}}$ denote the expectations over $p(\boldsymbol{x}, \boldsymbol{u})$ and $p(\boldsymbol{x})p(\boldsymbol{u})$, respectively. The representation functions $\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x})$ and $\boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u})$ in $r(\boldsymbol{x}, \boldsymbol{u})$ can be estimated by minimizing the empirical version of $J_{\mathrm{LR}}(r)$. Logistic regression has been previously used to estimate a density ratio (Sugiyama et al., 2012), and the minimizer of $J_{\mathrm{LR}}(r)$ with respect to r is equal to

$$\log \frac{p(\boldsymbol{x}, \boldsymbol{u})}{p(\boldsymbol{x})p(\boldsymbol{u})}$$

up to a constant. Intuitively, minimizing $J_{\mathrm{LR}}(r)$ enables to well-capture statistical dependencies between $\boldsymbol{x}$ and $\boldsymbol{u}$ by contrasting $\mathcal{D}_+$ with $\mathcal{D}_-$, and thus to estimate data representations $\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x})$ and $\boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u})$ having high mutual information. We theoretically justify this intuition, and clarifies when contrastive learning based on $J_{\mathrm{LR}}(r)$ can be considered to maximize mutual information.

Infomax is a classical framework for unsupervised representation learning, and advocates maximizing mutual information (MI) between input data and its representation (Linsker, 1989; Bell and Sejnowski, 1995). However, in practice, MI usually requires density estimation, which makes it hard to apply neural networks because of the notorious partition function problem. Recent promising approach employs complementary data and alternatively maximizes variational lower-bounds of MI between input and complementary data. For instance, the following lower bounds are often employed:

$$I(\mathcal{X},\mathcal{U}) \geq E_{\mathrm{xu}}[r(\boldsymbol{X},\boldsymbol{U})] - E_{\mathrm{x \times u}}[e^{r(\boldsymbol{X},\boldsymbol{U})-1}] \quad \text{(Nguyen et al., 2008; Sugiyama et al., 2008)} \tag{6}$$

$$I(\mathcal{X},\mathcal{U}) \geq E_{\mathrm{xu}}[r(\boldsymbol{X},\boldsymbol{U})] - \log E_{\mathrm{x \times u}}[e^{r(\boldsymbol{X},\boldsymbol{U})}] \quad \text{(Ruderman et al., 2012; Belghazi et al., 2018),} \tag{7}$$

where $I(\mathcal{X},\mathcal{U})$ denotes mutual information between $\boldsymbol{x}$ and $\boldsymbol{u}$, and is defined by

$$I(\mathcal{X},\mathcal{U}) := \int p(\boldsymbol{x}, \boldsymbol{u}) \log \frac{p(\boldsymbol{x}, \boldsymbol{u})}{p(\boldsymbol{x})p(\boldsymbol{u})}\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u}.$$

For other lower-bounds of MI, we refer to Poole et al. (2019). The key advantage is that these lower bounds enable us to employ neural networks without any special efforts, while it comes at the price for a statistical limitation that these lower-bounds may require an exponentially number of samples to accurately estimate MI (McAllester and Stratos, 2020, Theorem 3.1). As in logistic regression (5), the maximizers of both lower bounds in (6) and (7) with respect to r have been shown to be equal to

$$\log \frac{p(\boldsymbol{x}, \boldsymbol{u})}{p(\boldsymbol{x})p(\boldsymbol{u})}$$

up to constants (Nguyen et al., 2008; Ruderman et al., 2012). Thus, these lower bounds can be used for density ratio estimation. The primary purpose of maximizing these lower-bounds is to estimate mutual information between x and u, yet it has been believed that maximizing these lower bounds leads to maximization of mutual information between $h_{\mathrm{x}}(x)$ and $h_{\mathrm{u}}(u)$. Here, we provide a theoretically rigorous support to this belief.

2.3 Nonlinear independent component analysis

Nonlinear ICA is a solid framework for unsupervised representation learning, and assumes that data $x = (x_1, \ldots, x_{D_{\mathrm{x}}})^\top$ is generated as the following nonlinear mixing of the latent source $s = (s_1, \ldots, s_{D_{\mathrm{x}}})^\top$:

$$x = f(s), \tag{8}$$

where $f(s) = (f_1(s), \ldots, f_{D_{\mathrm{x}}}(s))^\top$, and f is an invertible function. The problem is to recover (or identify) the latent source components s_i. In the case of the linear mixing, i.e., $f(s) = As$ with an invertible matrix A, the latent source s can be recovered up to the permutation (i.e., ordering) and scales of $s_1, s_2, \ldots, s_{D_{\mathrm{x}}}$ when they are mutual independent and follow a nonGaussian density (Comon, 1994). However, the problem of nonlinear ICA has been proved to be seriously illposed under the same condition as the linear case because there exist an infinite number of decompositions of a random vector into mutually independent variables (Hyvärinen and Pajunen, 1999; Locatello et al., 2019). Nonetheless, a number of methods for nonlinear ICA have been previously proposed (Tan et al., 2001; Almeida, 2003; Blaschke et al., 2007), but most of them lack theoretical guarantees for source recovery and it is thus unclear to what extent these methods can recover the source components under the nonlinear mixing function f.

Recently, novel recovery conditions for nonlinear ICA have been established (Sprekeler et al., 2014; Hyvärinen and Morioka, 2016, 2017; Hyvärinen et al., 2019). The key condition alternative to mutual independence is conditional independence of $s_1, s_2, \ldots, s_{D_{\mathrm{x}}}$ given some complementary data u. Time contrastive learning (TCL) employs time segment labels as complementary data samples, and assumes that the conditional density of the source s given a time segment label is conditionally independent and belongs to an exponential family (Hyvärinen and Morioka, 2016). Then, the representation function $h_{\mathrm{x}}(x)$ learned by the multinomial logistic regression has been shown to asymptotically correspond to elementwise nonlinear functions of the latent source components $s_1, s_2, \ldots, s_{D_{\mathrm{x}}}$ up to a linear transformation. Hyvärinen et al. (2019) removed the exponential family assumption in TCL and has established recovery conditions in terms of more general conditional densities (nonexponential family). Then, it was proved that the representation function $h_{\mathrm{x}}(x)$ learned by contrastive learning based on $J_{\mathrm{LR}}(r)$ (i.e., logistic regression) is asymptotically equal to the latent source components $s_1, s_2, \ldots, s_{D_{\mathrm{x}}}$ up to their permutation and elementwise invertible functions. Details of the recovery conditions in Hyvärinen et al. (2019) are discussed in Section 3.2. Nonlinear ICA has been applied to causal analysis (Monti et al., 2020; Wu and Fukumizu, 2020) and transfer learning (Teshima et al., 2020).

This paper provides two new recovery proofs for nonlinear ICA, both of which include a novel insight: The dimensionality of complementary data is an important factor for source recovery. This insight is clearly supported by numerical experiments. Furthermore, we propose a novel generative model where data is generated as a nonlinear mixing of lower-dimensional latent source components and nuisance variables. The proposed generative model is more general than the one (8) in

Table 2: List of notations mainly used in Section 3. Wrt is an abbreviation of "with respect to".

$\boldsymbol{h}_{\mathrm{x}}^{\star}, \boldsymbol{h}_{\mathrm{u}}^{\star}$	Optimal representation functions of $\boldsymbol{h}_{\mathrm{x}}$ and $\boldsymbol{h}_{\mathrm{u}}$
$\psi^{\star}, a^{\star}, b^{\star}$	Optimal functions of ψ, a and b
$I(\mathcal{X}, \mathcal{U})$	Mutual information (MI) between $\boldsymbol{x}$ and $\boldsymbol{u}$
$I(\mathcal{Y}_{\mathrm{x}}, \mathcal{Y}_{\mathrm{u}})$	MI between $\boldsymbol{y}_{\mathrm{x}} = \boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x})$ and $\boldsymbol{y}_{\mathrm{u}} = \boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u})$
$\boldsymbol{h}_{\mathrm{x}}^{\perp} : \mathbb{R}^{D_{\mathrm{x}}} \to \mathbb{R}^{D_{\mathrm{x}}-d_{\mathrm{x}}}$	Function such that $(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}), \boldsymbol{h}_{\mathrm{x}}^{\perp}(\boldsymbol{x}))$ is invertible.
$\boldsymbol{h}_{\mathrm{u}}^{\perp} : \mathbb{R}^{D_{\mathrm{u}}} \to \mathbb{R}^{D_{\mathrm{u}}-d_{\mathrm{u}}}$	Function such that $(\boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u}), \boldsymbol{h}_{\mathrm{u}}^{\perp}(\boldsymbol{u}))$ is invertible.
$\boldsymbol{s} \in \mathbb{R}^{d_{\mathrm{x}}}$	Vector of latent source components
$\boldsymbol{f} : \mathbb{R}^{D_{\mathrm{x}}} \to \mathbb{R}^{D_{\mathrm{x}}}$	Nonlinear mixing function
$q_i : \mathbb{R} \times \mathbb{R}^{D_{\mathrm{u}}} \to \mathbb{R}$ (Proposition 2)	Exponent of the conditional density $p(s_i|\boldsymbol{u})$
q_i', q_i''	First- and second-order derivatives of $q_i(t, \boldsymbol{u})$ wrt t
$\nabla_{\boldsymbol{u}}$	Gradient wrt $\boldsymbol{u}$
$\boldsymbol{w} : \mathbb{R}^{D_{\mathrm{x}}} \times \mathbb{R}^{D_{\mathrm{u}}} \to \mathbb{R}^{2D_{\mathrm{x}}}$	$\boldsymbol{w}(\boldsymbol{v}, \boldsymbol{u}) := (q_1'(v_1, \boldsymbol{u}), \ldots, q_{D_{\mathrm{x}}}''(v_{D_{\mathrm{x}}}, \boldsymbol{u}))^{\top}$
$q_i : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ (Theorem 3)	Exponent of the conditional density $p(s_i|\boldsymbol{u})$
$\alpha_i^1, \alpha_i^2 : \mathbb{R} \to \mathbb{R}$	$\alpha_i^1(v) := \left. \frac{\partial^2 q_i(v,r)}{\partial v \partial r} \right|_{r=\lambda_i(\boldsymbol{u}_1)}$ and $\alpha_i^2(v) := \left. \frac{\partial^2 q_i(v,r)}{\partial v \partial r} \right|_{r=\lambda_i(\boldsymbol{u}_2)}$
$\boldsymbol{\alpha} : \mathbb{R}^{D_{\mathrm{x}}} \to \mathbb{R}^{D_{\mathrm{x}}}$	$\boldsymbol{\alpha}(\boldsymbol{v}) := \left(\frac{\alpha_1^2(v_1)}{\alpha_1^1(v_1)}, \ldots, \frac{\alpha_{D_{\mathrm{x}}}^2(v_{D_{\mathrm{x}}})}{\alpha_{D_{\mathrm{x}}}^1(v_{D_{\mathrm{x}}})} \right)^{\top}$
$\boldsymbol{n} \in \mathbb{R}^{D_{\mathrm{x}}-d_{\mathrm{x}}}$	Vector of nuisance variables

Table 3: Summary and comparison of theoretical conditions in maximization of MI, nonlinear ICA and nonlinear subspace estimation. Gen. model and source cond. are abbreviations of generative model and source condition, respectively. "-" means that the condition is not required. See the main text for more details.

	Maximization of MI	Nonlinear ICA	Nonlinear subspace
Form of $\psi^{\star}$	$\psi^{\star}(\boldsymbol{h}_{\mathrm{x}}^{\star}(\boldsymbol{x}), \boldsymbol{h}_{\mathrm{u}}^{\star}(\boldsymbol{u}))$	$\sum_{i=1}^{D_{\mathrm{x}}} \psi_i^{\star}(h_{\mathrm{x},i}^{\star}(\boldsymbol{x}), \boldsymbol{h}_{\mathrm{u}}^{\star}(\boldsymbol{u}))$ or $\sum_{i=1}^{D_{\mathrm{x}}} \psi_i^{\star}(h_{\mathrm{x},i}^{\star}(\boldsymbol{x}), h_{\mathrm{u},i}^{\star}(\boldsymbol{u}))$	$\psi^{\star}(\boldsymbol{h}_{\mathrm{x}}^{\star}(\boldsymbol{x}), \boldsymbol{h}_{\mathrm{u}}^{\star}(\boldsymbol{u}))$
Dim. assump.	$d_{\mathrm{x}} \leq D_{\mathrm{x}}$ and $d_{\mathrm{u}} \leq D_{\mathrm{u}}$	$d_{\mathrm{x}} = D_{\mathrm{x}}$ and $2D_{\mathrm{x}} \leq D_{\mathrm{u}}$ or $D_{\mathrm{x}} \leq D_{\mathrm{u}}$	$d_{\mathrm{x}} \leq D_{\mathrm{x}}$ and $d_{\mathrm{x}} \leq d_{\mathrm{u}}$
Gen. model	-	$\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{s})$	$\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{s}, \boldsymbol{n})$
Source cond.	-	$s_i \perp s_j|\boldsymbol{u}$ $(i \neq j)$	$\boldsymbol{s}, \boldsymbol{u} \perp \boldsymbol{n}$ and $\boldsymbol{s} \not\perp \boldsymbol{u}$

nonlinear ICA in the sense that the conditional independence of the latent source components is no longer assumed. Based on the proposed generative model, we establish theoretical conditions that complementary data enables us to automatically ignore the nuisance variables, and to estimate a nonlinear subspace related to only the lower-dimensional latent source components.

3. Three frameworks for unsupervised representation learning through density ratio estimation

This section shows that unsupervised representation learning on three frameworks can be performed by density ratio estimation, and discusses theoretical contributions on each of the frameworks. Ta-

bles 2 and 3 are a list of notations and summary of theoretical conditions in the three frameworks, respectively.

3.1 Maximization of mutual information

Maximization of mutual information (MI) (Barlow, 1961; Linsker, 1989; Bell and Sejnowski, 1995) is a classical yet recently retrieved framework for unsupervised representation learning combined with the recent development of deep neural networks (Hjelm et al., 2019; Tschannen et al., 2019). Density ratio estimation seems not to be strongly related, but our analysis implies that it is essential for maximization of MI.

Let us re-denote the representation functions of $\boldsymbol{x}$ and $\boldsymbol{u}$ by

$$\boldsymbol{y}_\mathrm{x} := \boldsymbol{h}_\mathrm{x}(\boldsymbol{x}) \quad \text{and} \quad \boldsymbol{y}_\mathrm{u} := \boldsymbol{h}_\mathrm{u}(\boldsymbol{u}),$$

respectively. The goal of maximization of MI is to find $\boldsymbol{h}_\mathrm{x}$ and $\boldsymbol{h}_\mathrm{u}$, which maximize MI between $\boldsymbol{y}_\mathrm{x}$ and $\boldsymbol{y}_\mathrm{u}$ defined by

$$I(\mathcal{Y}_\mathrm{x}, \mathcal{Y}_\mathrm{u}) := \int p(\boldsymbol{y}_\mathrm{x}, \boldsymbol{y}_\mathrm{u}) \log \frac{p(\boldsymbol{y}_\mathrm{x}, \boldsymbol{y}_\mathrm{u})}{p(\boldsymbol{y}_\mathrm{x})p(\boldsymbol{y}_\mathrm{u})} \mathrm{d}\boldsymbol{y}_\mathrm{x}\mathrm{d}\boldsymbol{y}_\mathrm{u}.$$

Data processing inequality shows that $I(\mathcal{Y}_\mathrm{x}, \mathcal{Y}_\mathrm{u})$ is a lower bound of MI between $\boldsymbol{x}$ and $\boldsymbol{u}$, i.e.,

$$I(\mathcal{X}, \mathcal{U}) \geq I(\mathcal{Y}_\mathrm{x}, \mathcal{Y}_\mathrm{u}). \tag{9}$$

Inequality (9) indicates that any reparametrizations of $\boldsymbol{x}$ and $\boldsymbol{u}$ never exceed $I(\mathcal{X}, \mathcal{U})$, and $I(\mathcal{Y}_\mathrm{x}, \mathcal{Y}_\mathrm{u})$ is maximized at $I(\mathcal{X}, \mathcal{U})$ if there exist such representation functions $\boldsymbol{h}_\mathrm{x}$ and $\boldsymbol{h}_\mathrm{u}$.

The following theorem proved in Appendix A clarifies how and when contrastive learning and variational MI estimation can be regarded as performing maximization of MI, and establishes conditions on which $I(\mathcal{Y}_\mathrm{x}, \mathcal{Y}_\mathrm{u}) = I(\mathcal{X}, \mathcal{U})$, i.e., $I(\mathcal{Y}_\mathrm{x}, \mathcal{Y}_\mathrm{u})$ is maximized:

Theorem 1. *We make the following assumptions:*

(A1) $p(\boldsymbol{x}, \boldsymbol{u}) > 0$, $p(\boldsymbol{x}) > 0$ *and* $p(\boldsymbol{u}) > 0$.

(A2) *There exist some functions* $\boldsymbol{h}_\mathrm{x}^\perp : \mathbb{R}^{D_\mathrm{x}} \to \mathbb{R}^{D_\mathrm{x}-d_\mathrm{x}}$ *and* $\boldsymbol{h}_\mathrm{u}^\perp : \mathbb{R}^{D_\mathrm{u}} \to \mathbb{R}^{D_\mathrm{u}-d_\mathrm{u}}$ *such that*

$$\begin{pmatrix} \boldsymbol{h}_\mathrm{x}(\boldsymbol{x}) \\ \boldsymbol{h}_\mathrm{x}^\perp(\boldsymbol{x}) \end{pmatrix} \in \mathbb{R}^{D_\mathrm{x}} \quad \text{and} \quad \begin{pmatrix} \boldsymbol{h}_\mathrm{u}(\boldsymbol{u}) \\ \boldsymbol{h}_\mathrm{u}^\perp(\boldsymbol{u}) \end{pmatrix} \in \mathbb{R}^{D_\mathrm{u}}$$

are both invertible[2].

(A3) *There exist functions* $\psi^\star$, $\boldsymbol{h}_\mathrm{x}^\star$, $\boldsymbol{h}_\mathrm{u}^\star$, $a^\star$ *and* $b^\star$ *such that the following equation holds:*

$$\log \frac{p(\boldsymbol{x}, \boldsymbol{u})}{p(\boldsymbol{x})p(\boldsymbol{u})} = \psi^\star(\boldsymbol{h}_\mathrm{x}^\star(\boldsymbol{x}), \boldsymbol{h}_\mathrm{u}^\star(\boldsymbol{u})) + a^\star(\boldsymbol{h}_\mathrm{x}^\star(\boldsymbol{x})) + b^\star(\boldsymbol{h}_\mathrm{u}^\star(\boldsymbol{u})). \tag{10}$$

Then, $I(\mathcal{Y}_\mathrm{x}, \mathcal{Y}_\mathrm{u}) = I(\mathcal{X}, \mathcal{U})$ *at* $\boldsymbol{h}_\mathrm{x} = \boldsymbol{h}_\mathrm{x}^\star$, *and* $\boldsymbol{h}_\mathrm{u} = \boldsymbol{h}_\mathrm{u}^\star$. *Conversely, suppose that* $I(\mathcal{Y}_\mathrm{x}, \mathcal{Y}_\mathrm{u}) = I(\mathcal{X}, \mathcal{U})$ *at* $\boldsymbol{h}_\mathrm{x} = \boldsymbol{h}_\mathrm{x}^\star$ *and* $\boldsymbol{h}_\mathrm{u} = \boldsymbol{h}_\mathrm{u}^\star$ *under Assumptions (A1-2). Then, there exist functions* $\psi^\star$, $a^\star$ *and* $b^\star$ *such that* (10) *holds.*

---

2. Invertibility means that there exist $\boldsymbol{g}_\mathrm{x}$ and $\boldsymbol{g}_\mathrm{u}$ such that $\boldsymbol{x} = \boldsymbol{g}_\mathrm{x}(\boldsymbol{y}_\mathrm{x}, \boldsymbol{y}_\mathrm{x}^\perp)$ and $\boldsymbol{u} = \boldsymbol{g}_\mathrm{u}(\boldsymbol{y}_\mathrm{u}, \boldsymbol{y}_\mathrm{u}^\perp)$ where $\boldsymbol{y}_\mathrm{x}^\perp := \boldsymbol{h}_\mathrm{x}^\perp(\boldsymbol{x})$ and $\boldsymbol{y}_\mathrm{u}^\perp := \boldsymbol{h}_\mathrm{u}^\perp(\boldsymbol{u})$.

Theorem 1 implies that $I(\mathcal{Y}_\mathrm{x}, \mathcal{Y}_\mathrm{u})$ can be maximized through density ratio estimation, and thus motivates us to develop practical methods for maximization of MI through density ratio estimation. Eq.(10) is inspired by *sufficient dimension reduction* (Li, 1991; Cook, 1998; Fukumizu et al., 2004), which is a solid framework for *supervised* dimensionality reduction and whose goal is to find an informative lower-dimensional subspace to the output variable based on the conditional independence condition. As shown in the proof of Theorem 1, (10) can be rewritten as the following conditional independence conditions:

$$\boldsymbol{u} \perp \boldsymbol{x} \mid \boldsymbol{y}_\mathrm{x} = \boldsymbol{h}_\mathrm{x}^\star(\boldsymbol{x}) \quad \text{and} \quad \boldsymbol{x} \perp \boldsymbol{u} \mid \boldsymbol{y}_\mathrm{u} = \boldsymbol{h}_\mathrm{u}^\star(\boldsymbol{u}).$$

The conditional independence between $\boldsymbol{u}$ and $\boldsymbol{x}$ given $\boldsymbol{y}_\mathrm{x} = \boldsymbol{h}_\mathrm{x}^\star(\boldsymbol{x})$ implies that the lower-dimensional representation $\boldsymbol{h}_\mathrm{x}^\star(\boldsymbol{x})$ has the same amount of information for $\boldsymbol{u}$ as the original input data $\boldsymbol{x}$. The same implication holds the conditional independence between $\boldsymbol{u}$ and $\boldsymbol{x}$ given $\boldsymbol{y}_\mathrm{u} = \boldsymbol{h}_\mathrm{u}^\star(\boldsymbol{u})$ as well. Thus, accurately estimating the density ratio would yield representations of $\boldsymbol{u}$ and $\boldsymbol{x}$ possibly with minimum information loss.

An interesting point of Theorem 1 is that $I(\mathcal{Y}_\mathrm{x}, \mathcal{Y}_\mathrm{u}) = I(\mathcal{X}, \mathcal{U})$ conversely implies the density ratio equation (10), and is useful for understanding when contrastive learning does not perform maximization of MI. In Section 2.2, we suppose that $\boldsymbol{u}^*(t)$ are drawn from the marginal density $p(\boldsymbol{u})$ in (4), but in practice, $\boldsymbol{u}^*(t)$ are often taken from another dataset (Arandjelovic and Zisserman, 2017)[3] whose probability density is different from the marginal density $p(\boldsymbol{u})$ and denoted by $\tilde{p}(\boldsymbol{u})$. Then, by assuming that the $\boldsymbol{u}^*$ are independent to $\boldsymbol{x}$, contrastive learning related to density ratio estimation yields an estimate of

$$\log \frac{p(\boldsymbol{x}, \boldsymbol{u})}{p(\boldsymbol{x})\tilde{p}(\boldsymbol{u})}$$

up to a constant, and (10) is never fulfilled. Thus, contrastive learning based on another marginal density $\tilde{p}(\boldsymbol{u})$ might not be regarded as maximizing $I(\mathcal{Y}_\mathrm{x}, \mathcal{Y}_\mathrm{u})$ in general. On the other hand, as long as $\boldsymbol{u}^*(t)$ are drawn form the marginal density $p(\boldsymbol{u})$, Theorem 1 would be a direct support to the belief that contrastive learning and variational MI estimation can be considered to perform maximization of MI because the popular objective functions are related to density ratio estimation as reviewed in Section 2.2.

Another simple practical point of Theorem 1 is the form of the right-hand side on (10): The right-hand side indicates that all terms have to be functions of $\boldsymbol{h}_\mathrm{u}^\star$ and/or $\boldsymbol{h}_\mathrm{x}^\star$. Thus, even when $a^\star(\boldsymbol{h}_\mathrm{x}^\star(\boldsymbol{x}))$ is simply replaced with $a(\boldsymbol{x})$ (i.e., a is not a function of $\boldsymbol{h}_\mathrm{x}^\star$), Theorem 1 does not hold.

3.2 Nonlinear ICA: A new insight for source recovery

Here, we perform two theoretical analyses for source recovery in nonlinear ICA with density ratio estimation (Proposition 2 and Theorem 3). These analyses shed light on a novel insight that the dimensionality of complementary data is an important factor for source recovery, which has not been revealed in previous work of nonlinear ICA. Furthermore, Theorem 3 can be regarded as a generalization of Theorem 1 in Hyvärinen and Morioka (2017).

---

3. For instance, in Arandjelovic and Zisserman (2017), the positive pairs of input and complimentary data samples are video frames and audio clips in the same video that overlap in time, while the negative pairs are randomly extracted from two different videos. Thus, the marginal density of the complimentary data (i.e., video frame or audio clip) can be different in the positive and negative data.

Importance of the dimensionality of complementary data: As reviewed in Section 2.3, nonlinear ICA assumes that input data $\boldsymbol{x} = (x_1, x_2, \ldots, x_{D_x})^\top$ is generated as a nonlinear mixing of the latent source components $s_1, s_2, \ldots, s_{D_x}$ as follows:

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

where $\boldsymbol{s} = (s_1, s_2, \ldots, s_{D_x})^\top$ and $\boldsymbol{f} : \mathbb{R}^{D_x} \to \mathbb{R}^{D_x}$ is assumed to be invertible. Then, the goal is to recover the latent source components s_i. To this end, recent work of nonlinear ICA employs complementary data $\boldsymbol{u}(t)$ in addition to input data $\boldsymbol{x}(t)$. For instance, by regarding $\boldsymbol{x}(t)$ as time series data at the time index t, Hyvärinen and Morioka (2017) use past input data as $\boldsymbol{u}(t)$ (e.g., $\boldsymbol{u}(t) = \boldsymbol{x}(t-1)$).

We first establish the following theorem showing that the latent source components can be recovered up to their permutation (i.e., ordering) and elementwise invertible functions, and that density ratio estimation plays an important role in nonlinear ICA as well:

Proposition 2. *Suppose that $d_x = D_x$. We further make the following assumptions:*

(B1) The latent source components s_i are conditionally independent given $\boldsymbol{u}$. More specifically, the conditional density of $\boldsymbol{s}$ given $\boldsymbol{u}$ takes the following form:

$$\log p(\boldsymbol{s}|\boldsymbol{u}) = \sum_{i=1}^{D_x} q_i(s_i, \boldsymbol{u}) - \log Z(\boldsymbol{u}),$$

where $Z(\boldsymbol{u})$ denotes the partition function and q_i are differentiable functions.

(B2) Input data $\boldsymbol{x}$ is generated according to (8) where the mixing function $\boldsymbol{f}$ is invertible.

(B3) Dimensionality of complementary data is twice larger than or twice as large as input data, i.e., $2D_x \le D_u$.

(B4) There exists a single point $\boldsymbol{u}_1$ such that the rank of $\nabla_{\boldsymbol{u}} \boldsymbol{w}(\boldsymbol{v}, \boldsymbol{u}) \in \mathbb{R}^{2D_x \times D_u}$ at $\boldsymbol{u} = \boldsymbol{u}_1$ is $2D_x$ for all $\boldsymbol{v} = (v_1, v_2, \ldots, v_{D_x})^\top$ where $\nabla_{\boldsymbol{u}}$ denotes the differential operator with respect to $\boldsymbol{u}$, and

$$\boldsymbol{w}(\boldsymbol{v}, \boldsymbol{u}) := \big(q_1'(v_1, \boldsymbol{u}), \ldots, q_{D_x}'(v_{D_x}, \boldsymbol{u}), q_1''(v_1, \boldsymbol{u}), \ldots, q_{D_x}''(v_{D_x}, \boldsymbol{u})\big)^\top \in \mathbb{R}^{2D_x},$$

with $q_i'(t, \boldsymbol{u}) := \frac{\partial}{\partial t} q_i(t, \boldsymbol{u})$ and $q_i''(t, \boldsymbol{u}) := \frac{\partial^2}{\partial t^2} q_i(t, \boldsymbol{u})$.

(B5) There exist functions $\psi_i^\star$, $\boldsymbol{h}_x^\star$, $\boldsymbol{h}_u^\star$, $a^\star$ and $b^\star$ such that the following equation holds:

$$\log \frac{p(\boldsymbol{x}, \boldsymbol{u})}{p(\boldsymbol{x})p(\boldsymbol{u})} = \sum_{i=1}^{D_x} [\psi_i^\star(h_{x,i}^\star(\boldsymbol{x}), \boldsymbol{h}_u^\star(\boldsymbol{u}))] + a^\star(\boldsymbol{h}_x^\star(\boldsymbol{x})) + b^\star(\boldsymbol{h}_u^\star(\boldsymbol{u})), \tag{11}$$

where $\psi_i^\star$ are differentiable functions and $\boldsymbol{h}_x^\star(\boldsymbol{x}) := (h_{x,1}^\star(\boldsymbol{x}), \ldots, h_{x,D_x}^\star(\boldsymbol{x}))^\top$ is invertible.

Then, under Assumptions (B1-5), the representation function $\boldsymbol{h}_x^\star(\boldsymbol{x})$ is equal to $\boldsymbol{s}$ up to a permutation and elementwise invertible functions.

The proof is given in Appendix B. We essentially followed the proof of Theorem 1 in Hyvärinen et al. (2019) and derived the same conclusion. Here, the main difference is Assumptions (B3-4), which give a new insight for source recovery. Hyvärinen et al. (2019) adopted an alternative assumption to Assumptions (B3-4) called the *assumption of variability*, which assumes that there exist $2D_\mathrm{x} + 1$ points, $\boldsymbol{u}_0, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_{2D_\mathrm{x}}$, such that the following $2D_\mathrm{x}$ vectors are linearly independent:

$$\boldsymbol{w}(\boldsymbol{v}, \boldsymbol{u}_j) - \boldsymbol{w}(\boldsymbol{v}, \boldsymbol{u}_0) \text{ for } j = 1, \ldots, 2D_\mathrm{x} \text{ are linearly independent for all } \boldsymbol{v}.$$

The dimensionality assumption $2D_\mathrm{x} \leq D_\mathrm{u}$ in Assumption (B3) would correspond to the $2D_\mathrm{x}$ vectors in the assumption of variability, but, in contrast, sheds light on a novel insight: The dimensionality of complementary data is an important factor for source recovery, which has not been revealed in previous work of nonlinear ICA. In fact, we numerically demonstrate that the accuracy of source recovery clearly depends on the dimensionality of complementary data in Section 5.1.2. Furthermore, Assumption (B3) is practically useful because it can be checked very easily. Regarding Assumption (B4), it implies that the complementary data $\boldsymbol{u}$ is clearly dependent to $\boldsymbol{s}$: When the i-th element in $\boldsymbol{u}$ is independent to $\boldsymbol{s}$ and $2D_\mathrm{x} = D_\mathrm{u}$ (i.e., $\nabla_{\boldsymbol{u}} \boldsymbol{w}(\boldsymbol{v}, \boldsymbol{u})$ is a square matrix), the i-th column in $\nabla_{\boldsymbol{u}} \boldsymbol{w}(\boldsymbol{v}, \boldsymbol{u})$ is the zero vector, and thus the rank assumption in Assumption (B4) is never satisfied.

According to Hyvärinen et al. (2019), the assumption of variability implies that the underlying conditional density $p(\boldsymbol{s}|\boldsymbol{u})$ has to be diverse and complex to recover the source components. As in Theorem 2 in Hyvärinen et al. (2019), we show that Assumption (B4) includes the same implication under the following exponential family:

$$\log p(\boldsymbol{s}|\boldsymbol{u}) = \sum_{i=1}^{D_\mathrm{x}} \sum_{k=1}^{K} \lambda_{ik}(\boldsymbol{u}) q_{ik}(s_i) - \log Z(\boldsymbol{u}), \tag{12}$$

where λ_{ik} and q_{ik} are some scalar functions and K is a positive integer. Appendix C proves that

- When $K = 1$, $\mathrm{rank}(\nabla_{\boldsymbol{u}} \boldsymbol{w}(\boldsymbol{v}, \boldsymbol{u})) \leq D_\mathrm{x}$ where $\mathrm{rank}(\cdot)$ denotes the rank of a matrix.

- When $K > 1$, $\mathrm{rank}(\nabla_{\boldsymbol{u}} \boldsymbol{w}(\boldsymbol{v}, \boldsymbol{u})) \leq 2D_\mathrm{x}$

Thus, when $K = 1$, the $2D_\mathrm{x}$ rank assumption in Assumption (B4) is never fulfilled. This implies that in order to recover the source components, the conditional density has to be diverse and complex such as the exponential family (12) with a relatively large mixture number K.

A milder dimensionality assumption: The dimensionality assumption $2D_\mathrm{x} \leq D_\mathrm{u}$ in Assumption (B3) might be strong because we need to have relatively high-dimensional complementary data. However, this assumption can be relaxed by restricting the underlying conditional density of $\boldsymbol{s}$ given $\boldsymbol{u}$. The following theorem is based on a milder dimensionality assumption than Assumption (B3), and can be seen as a generalization of Theorem 1 in Hyvärinen and Morioka (2017):

Theorem 3. *Suppose that $d_\mathrm{x} = d_\mathrm{u} = D_\mathrm{x}$. We make the following assumptions:*

(B′1) The latent source components s_i are conditionally independent given $\boldsymbol{u}$, and the conditional density of $\boldsymbol{s}$ given $\boldsymbol{u}$ takes the following form:

$$\log p(\boldsymbol{s}|\boldsymbol{u}) = \sum_{i=1}^{D_\mathrm{x}} q_i(s_i, \lambda_i(\boldsymbol{u})) - \log Z(\boldsymbol{u}), \tag{13}$$

where q_i and λ_i are differentiable functions for $i = 1, \ldots, D_\mathrm{x}$.

(B′2) Input Data $\boldsymbol{x}$ is generated according to (8) where the mixing function $\boldsymbol{f}$ is invertible.

(B′3) Dimensionality of complementary data is larger than or equal to input data, i.e., $D_{\mathrm{x}} \leq D_{\mathrm{u}}$.

(B′4) There exist two points, $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$, such that $\alpha_i^1(v) := \frac{\partial^2 q_i(v,r)}{\partial v \partial r}\Big|_{r=\lambda_i(\boldsymbol{u}_1)} \neq 0$ and $\alpha_i^2(v) := \frac{\partial^2 q_i(v,r)}{\partial v \partial r}\Big|_{r=\lambda_i(\boldsymbol{u}_2)} \neq 0$ for all i and v.

(B′5) There exist D_{x} points, $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{D_{\mathrm{x}}}$, such that the D_{x} vectors $\boldsymbol{\alpha}(\boldsymbol{v}_1)$, $\boldsymbol{\alpha}(\boldsymbol{v}_2)$, ..., $\boldsymbol{\alpha}(\boldsymbol{v}_{D_{\mathrm{x}}})$ are linearly independent where with $\boldsymbol{v} := (v_1, v_2, \ldots, v_{D_{\mathrm{x}}})^\top$,

$$\boldsymbol{\alpha}(\boldsymbol{v}) := \left(\frac{\alpha_1^2(v_1)}{\alpha_1^1(v_1)}, \frac{\alpha_2^2(v_2)}{\alpha_2^1(v_2)}, \ldots, \frac{\alpha_{D_{\mathrm{x}}}^2(v_{D_{\mathrm{x}}})}{\alpha_{D_{\mathrm{x}}}^1(v_{D_{\mathrm{x}}})} \right)^\top \in \mathbb{R}^{D_{\mathrm{x}}}.$$

(B′6) There exist functions $\psi_i^\star$, $\boldsymbol{h}_{\mathrm{x}}^\star$, $\boldsymbol{h}_{\mathrm{u}}^\star$, $a^\star$ and $b^\star$ such that the following equation holds:

$$\log \frac{p(\boldsymbol{x}, \boldsymbol{u})}{p(\boldsymbol{x}) p(\boldsymbol{u})} = \sum_{i=1}^{D_{\mathrm{x}}} [\psi_i^\star(h_{\mathrm{x},i}^\star(\boldsymbol{x}), h_{\mathrm{u},i}^\star(\boldsymbol{u}))] + a^\star(\boldsymbol{h}_{\mathrm{x}}^\star(\boldsymbol{x})) + b^\star(\boldsymbol{h}_{\mathrm{u}}^\star(\boldsymbol{u})), \qquad (14)$$

where $\psi_i^\star$ are differentiable functions, $\boldsymbol{h}_{\mathrm{x}}^\star(\boldsymbol{x}) := (h_{\mathrm{x},1}^\star(\boldsymbol{x}), \ldots, h_{\mathrm{x},D_{\mathrm{x}}}^\star(\boldsymbol{x}))^\top$ is invertible, and $\boldsymbol{h}_{\mathrm{u}}^\star(\boldsymbol{x}) := (h_{\mathrm{u},1}^\star(\boldsymbol{u}), \ldots, h_{\mathrm{u},D_{\mathrm{x}}}^\star(\boldsymbol{u}))^\top.$

Then, under Assumptions (B′1-6), the representation function $\boldsymbol{h}_{\mathrm{x}}^\star(\boldsymbol{x})$ is equal to $\boldsymbol{s}$ up to a permutation and elementwise invertible functions.

The proof is given in Appendix D. The key point in Theorem 3 is that the dimensionality assumption $D_{\mathrm{x}} \leq D_{\mathrm{u}}$ in Assumption (B′3) is milder than $2D_{\mathrm{x}} \leq D_{\mathrm{u}}$ in Assumption (B3). However, this milder assumption comes at a cost of making the conditional density $p(\boldsymbol{s}|\boldsymbol{u})$ less general than Proposition 2 and of restricting the form of the right-hand side on (14): The conditional density $p(\boldsymbol{s}|\boldsymbol{u})$ is restricted into a form of pairwise combinations of s_i and $\lambda_i(\boldsymbol{u})$ for $i = 1 \ldots, D_{\mathrm{x}}$ in (13), and the right-hand side on (14) also takes a pairwise form of $h_{\mathrm{x},i}^\star(\boldsymbol{x})$ and $h_{\mathrm{u},i}^\star(\boldsymbol{u})$.

Similarly as Assumption (B4) in Proposition 2, Assumption (B′5) also implies that the underlying conditional density (13) is diverse and complex. Indeed, when $p(\boldsymbol{s}|\boldsymbol{u})$ belongs to the exponential family (12) in $K = 1$, the ratio $\frac{\alpha_i^2(v_i)}{\alpha_i^1(v_i)}$ is equal to a constant $\frac{\lambda_{i1}(\boldsymbol{u}_2)}{\lambda_{i1}(\boldsymbol{u}_1)}$ for all i. Then, $\boldsymbol{\alpha}(\boldsymbol{v})$ is a constant vector, and cannot be linearly independent over the D_{x} points. Thus, Assumption (B′5) is never satisfied under the exponential family (12) in $K = 1$.

Theorem 3 generalizes Theorem 1 in Hyvärinen and Morioka (2017): By regarding $\boldsymbol{x}(t), t = 1, \ldots, T$ as time series data at the time index t, Theorem 1 in Hyvärinen and Morioka (2017) is a special case of Theorem 3 where $\boldsymbol{u}(t) = \boldsymbol{x}(t-1)$ (thus, implicitly suppose $D_{\mathrm{x}} = D_{\mathrm{u}}$ and Assumption (B′3) is satisfied), $\boldsymbol{\lambda}(\boldsymbol{u}) = \boldsymbol{u}$, and $\boldsymbol{h}_{\mathrm{u}} = \boldsymbol{h}_{\mathrm{x}}$ (e.g., the same neural architecture with weight sharing). This generalization is not straightforward because we had to derive a new lemma (Lemma 10), and thus the proof is substantially different. Furthermore, by this generalization, Theorem 3, again, reveals that the dimensionality of complementary data is an important factor, which has not been seen in Theorem 1 of Hyvärinen and Morioka (2017).

Finally, we note a subtle difference of $\psi(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}), \boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u}))$ in nonlinear ICA. Nonlinear ICA requires us to slightly modify $\psi(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}), \boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u}))$ as follows: For Proposition 2,

$$\psi(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}), \boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u})) = \sum_{i=1}^{D_{\mathrm{x}}} \psi_i(h_{\mathrm{x},i}(\boldsymbol{x}), \boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u})), \tag{15}$$

while in Theorem 3,

$$\psi(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}), \boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u})) = \sum_{i=1}^{D_{\mathrm{x}}} [\psi_i(h_{\mathrm{x},i}(\boldsymbol{x}), h_{\mathrm{u},i}(\boldsymbol{u}))]. \tag{16}$$

In contrast with maximization of mutual information (Theorem 1) as well as nonlinear subspace estimation (Theorem 4), $\psi(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}), \boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u}))$ in (15) (or in (16)) is expressed as the sum of element-wise functions with respect to $\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x})$ (or $\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x})$ and $\boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u})$). Thus, in practice, we need to slightly modify the form of $\psi(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}), \boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u}))$ to (15) or (16) when performing nonlinear ICA.

3.3 Nonlinear subspace estimation with complementary data

Here, we first propose a new generative model where input data is generated as a nonlinear mixing of lower-dimensional latent source components and nuisance variables. Then, we establish some theoretical conditions to estimate a lower-dimensional nonlinear subspace of the latent source components only.

A new generative model: Let us consider the following novel generative model for input data $\boldsymbol{x} \in \mathbb{R}^{D_{\mathrm{x}}}$:

$$\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{s}, \boldsymbol{n}), \tag{17}$$

where $\boldsymbol{f}$ is an invertible nonlinear mixing function, $\boldsymbol{s} \in \mathbb{R}^{d_{\mathrm{x}}}$ and $\boldsymbol{n} \in \mathbb{R}^{D_{\mathrm{x}}-d_{\mathrm{x}}}$. We further assume that

- Source components s_i are statistically independent to nuisance variables $\boldsymbol{n}$, i.e., $\boldsymbol{s} \perp \boldsymbol{n}$

- Complementary data is supposed to be available such that $\boldsymbol{u} \perp \boldsymbol{n}$, while $\boldsymbol{u}$ depends on $\boldsymbol{s}$, i.e., $\boldsymbol{s} \not\perp \boldsymbol{u}$ where $\not\perp$ denotes statistical dependence.

Based on the observations of $\boldsymbol{x}$ and $\boldsymbol{u}$, the goal is to estimate a lower-dimensional nonlinear subspace of only $\boldsymbol{s}$, which is separated from $\boldsymbol{n}$. In contrast with nonlinear ICA, we do not necessarily assume that s_i are conditionally independent given $\boldsymbol{u}$. Thus, (17) is more general than the generative model (8) in nonlinear ICA in the sense that the conditional independence of s_i given $\boldsymbol{u}$ is no longer assumed.

The new generative model (17) can be motivated by many practical situations. For example, time-series data such as brain signals (Dornhege et al., 2007) could be observed as a mixture of stationary and nonstationary components, and the nonstationarity could be important for a range of tasks such as change detection (Blythe et al., 2012). For this example, by using the past data in time series data as complementary data (e.g., $\boldsymbol{u}(t) = \boldsymbol{x}(t-1)$), we may estimate a subspace of the useful nonstationary components. Another example is image data. The pixel values around the center of an image usually depends on surrounding pixels, while the pixels around the corners of the image are often almost independent to the other pixels (e.g., MNIST images). By using surrounding pixels as the complementary data, it would be very informative to estimate some lower-dimensional subspace for image data, which constitutes the fundamental part of image data.

Estimating a lower-dimensional subspace of the source components: First of all, it is important to understand whether we can estimate a lower-dimensional subspace of the latent sources s, which is separated from nuisance variables n. The following theorem gives conditions to estimate such a subspace as a (vector-valued) function of s only:

Theorem 4. *Assume that*

(C1) Input data x is generated according to (17) where the mixing function f is invertible, $s \perp n$, $s \not\perp u$ and $u \perp n$.

(C2) There exist functions $\psi^\star$, $h_{\mathrm{x}}^\star$, $h_{\mathrm{u}}^\star$, $a^\star$ and $b^\star$ such that the following equation holds:

$$\log \frac{p(x, u)}{p(x)p(u)} = \psi^\star(h_{\mathrm{x}}^\star(x), h_{\mathrm{u}}^\star(u)) + a^\star(h_{\mathrm{x}}^\star(x)) + b^\star(h_{\mathrm{u}}^\star(u)), \tag{18}$$

where $h_{\mathrm{x}}^\star$ is surjective.

(C3) Dimensionality of $h_{\mathrm{x}}^\star(x)$ is smaller than or equal to $h_{\mathrm{u}}^\star(u)$, i.e., $d_{\mathrm{x}} \leq d_{\mathrm{u}}$.

(C4) There exists at least a single point u_1 such that the rank of $\nabla_v \nabla_u \psi^\star(v, h_{\mathrm{u}}^\star(u)) \in \mathbb{R}^{D_{\mathrm{u}} \times d_{\mathrm{x}}}$ is d_{x} at $u = u_1$ for all $v \in \mathbb{R}^{d_{\mathrm{x}}}$ where $v := h_{\mathrm{x}}^\star(x)$.

Under Assumptions (C1-4), the representation function $h_{\mathrm{x}}^\star(x)$ is a nonlinear function of only s.

The proof is given in Appendix E. The interesting point of Theorem 4 is that complementary data u enables us to automatically ignore the nuisance variables n, and Theorem 4 implies that a lower-dimensional nonlinear subspace only for s can be estimated through density ratio estimation. Assumption (C3) is a necessary condition that the rank of $\nabla_v \nabla_u \psi^\star(v, h_{\mathrm{u}}^\star(u))$ is d_{x}: By defining $r := h_{\mathrm{u}}^\star(u)$,

$$\mathrm{rank}(\nabla_v \nabla_u \psi^\star(v, h_{\mathrm{u}}^\star(u))) = \mathrm{rank}(J_r(u)^\top [\nabla_v \nabla_r \psi^\star(v, r)]) \leq \min(d_{\mathrm{x}}, d_{\mathrm{u}}),$$

where $J_r(u) := \nabla_u r \in \mathbb{R}^{d_{\mathrm{u}} \times D_{\mathrm{u}}}$ is the Jacobian of r at u and $\nabla_v \nabla_r \psi^\star(v, r) \in \mathbb{R}^{d_{\mathrm{u}} \times d_{\mathrm{x}}}$. Thus, it must be $d_{\mathrm{x}} \leq d_{\mathrm{u}}$ to satisfy Assumption (C4).

In order to understand the implication of Assumption (C4), we derive the following equation from (18) in Appendix F:

$$J_v(s)[\nabla_s \nabla_u \log p(s|u)\big|_{u=u_1}] = \nabla_v \nabla_u \psi^\star(v, u)\big|_{u=u_1}, \tag{19}$$

where $J_v(s) := \nabla_s v \in \mathbb{R}^{d_{\mathrm{x}} \times d_{\mathrm{x}}}$ is the Jacobian of $v = h_{\mathrm{x}}^\star(f(s, n))$ at s. According to Lemma 11 and Assumption (C4), (19) ensures that the rank of $\nabla_s \nabla_u \log p(s|u)\big|_{u=u_1}$ is d_{x}. Next, as discussed in nonlinear ICA (Section 3.2), we investigate whether or not the rank of $\nabla_s \nabla_u \log p(s|u)\big|_{u=u_1}$ is d_{x} under the following exponential family:

$$\log p(s|u) = \sum_{j=1}^{D} \lambda_j(u) q_j(s) - \log Z(u), \tag{20}$$

16

where λ_j and q_j are some scalar functions. Then, we have

$$\nabla_{\boldsymbol{s}} \nabla_{\boldsymbol{u}} \log p(\boldsymbol{s}|\boldsymbol{u}) = \sum_{j=1}^{D} \{\nabla_{\boldsymbol{u}} \lambda_j(\boldsymbol{u})\} \{\nabla_{\boldsymbol{s}} q_j(\boldsymbol{s})\}^{\top},$$

which indicates that by the rank factorization theorem,

$$\mathrm{rank}(\nabla_{\boldsymbol{s}} \nabla_{\boldsymbol{u}} \log p(\boldsymbol{s}|\boldsymbol{u})) \leq \min(d_{\mathrm{x}}, D)$$

where we used $d_{\mathrm{x}} \leq d_{\mathrm{u}} \leq D_{\mathrm{u}}$ from Assumption (C3). Thus, when $D < d_{\mathrm{x}}$, Assumption (C4) is not fulfilled. As in nonlinear ICA, Assumption (C4) possibly implies that in order to estimate a nonlinear subspace of $\boldsymbol{s}$, $p(\boldsymbol{s}|\boldsymbol{u})$ has to be sufficiently complex such as the exponential family (20) with relatively large D ($> d_{\mathrm{x}}$).

Compared with nonlinear ICA in Section 3.2, Theorem 4 stands on a more general setting in the sense that the elements in $\boldsymbol{s}$ are not necessarily assumed to be conditionally independent. Furthermore, the dimensionality condition $d_{\mathrm{x}} \leq d_{\mathrm{u}}$ in Assumption (C3) is milder than Assumption (B3) in Proposition 2 and Assumption (B′3) in Theorem 3 because d_{x} and d_{u} are the dimensionalities of the representation functions $\boldsymbol{h}_{\mathrm{x}}^{\star}$ and $\boldsymbol{h}_{\mathrm{u}}^{\star}$ and assumed to be smaller than D_{x} and D_{u}, respectively. However, these milder conditions come at a price for losing the recover of each latent source component s_i: Nonlinear ICA recovers each component s_i up to a permutation and elementwise invertible function, while Theorem 4 only guarantees that a nonlinear function of $\boldsymbol{s}$ can be estimated.

A similar generative model as (17) was proposed in multi-view learning (Gresele et al., 2019), which considers that two data, $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$, are generated as

$$\boldsymbol{x}^{(1)} = \boldsymbol{f}^{(1)}(\boldsymbol{s}) \quad \text{and} \quad \boldsymbol{x}^{(2)} = \boldsymbol{f}^{(2)}(\boldsymbol{g}(\boldsymbol{s}, \boldsymbol{n})), \tag{21}$$

where $\boldsymbol{f}^{(1)}$ and $\boldsymbol{f}^{(2)}$ are the mixing functions, and $\boldsymbol{g}$ denotes an elementwise vector-valued function. The generative model (21) assumes that $\boldsymbol{s} \perp \boldsymbol{n}$, $s_i \perp s_j$ and $n_i \perp n_j$ for $i \neq j$ where n_i denotes the i-th element in $\boldsymbol{n}$. Then, Gresele et al. (2019) proved that each component s_i can be recovered up to elementwise invertible functions. In contrast, the assumptions in our generative model (17) can be more general because it is not assumed that the elements both in $\boldsymbol{s}$ and $\boldsymbol{n}$ are independent, i.e., $s_i \not\perp s_j$ and $n_i \not\perp n_j$ for $i \neq j$ in our generative model.

4. Estimation methods

Section 3 theoretically showed that density ratio estimation plays a key role in the three frameworks, and motivates us to develop practical methods for unsupervised representation learning through density ratio estimation. This section proposes two practical methods to estimate the log-density ratio based on neural networks. The first method employs the γ-cross entropy (Fujisawa and Eguchi, 2008), which is a robust variant of the standard cross entropy against outliers. For nonlinear ICA, the second one employs the Donsker-Varadhan variational estimation for mutual information (Ruderman et al., 2012; Belghazi et al., 2018). After describing the proposed methods, we investigate the outlier-robustness of the proposed methods. A list of notations mainly used in this section is given in Table 4.

Table 4: List of notations used in Section 4.

E_{xu}	Expectation over $p(\boldsymbol{x}, \boldsymbol{u})$
$E_{\mathrm{x} \times \mathrm{u}}$	Expectation over $p(\boldsymbol{x})p(\boldsymbol{u})$
E_{x}	Expectation over $p(\boldsymbol{x})$
E_{u}	Expectation over $p(\boldsymbol{u})$
J_γ	γ-cross entropy
$\widehat{J}_\gamma,\ \widetilde{J}_\gamma$	Empirical approximations of J_γ
J_{DV}	Donsker-Varadhan variational lower-bound of MI
$\widehat{J}_{\mathrm{DV}},\ \widetilde{J}_{\mathrm{DV}}$	Empirical approximations of J_{DV}
$\boldsymbol{u}_{\mathrm{p}}(t)$	Random permutation of $\boldsymbol{u}(t)$ with respect to t
$\bar{p}(\boldsymbol{x}, \boldsymbol{u})$	Contaminated joint probability density function of $\boldsymbol{x}$ and $\boldsymbol{u}$
$\bar{p}(\boldsymbol{x})$	Contaminated marginal probability density function of $\boldsymbol{x}$
$\bar{p}(\boldsymbol{u})$	Contaminated marginal probability density function of $\boldsymbol{u}$
$\delta(\boldsymbol{x}, \boldsymbol{u})$	Joint probability density function for outliers
$\delta(\boldsymbol{x}),\ \delta(\boldsymbol{u})$	Marginal probability density functions for outliers
ϵ	Contamination ratio ($0 \le \epsilon \le 1$)
$\bar{\boldsymbol{x}},\ \bar{\boldsymbol{u}}$	Outlier points
$\delta_{\bar{\boldsymbol{x}}}(\boldsymbol{x})$	Dirac delta function with a point mass at $\bar{\boldsymbol{x}}$
$\delta_{\bar{\boldsymbol{u}}}(\boldsymbol{u})$	Dirac delta function with a point mass at $\bar{\boldsymbol{u}}$
$r_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{u})$	Model parametrized by $\boldsymbol{\theta}$
$\boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{u})$	Gradient of $r_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{u})$ with respect to $\boldsymbol{\theta}$
$\boldsymbol{\theta}^\star$	Parameters optimized over the (noncontaminated) densities
$\boldsymbol{\theta}_\epsilon$	Parameters optimized over the contaminated densities
$\mathrm{IF}_{\mathrm{DV}}$	Influence function over J_{DV}
$\boldsymbol{V}_{\mathrm{DV}}$	$\boldsymbol{V}_{\mathrm{DV}} := E_{\mathrm{xu}}[\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X}, \boldsymbol{U})\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X}, \boldsymbol{U})^\top] - E_{\mathrm{xu}}[\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X}, \boldsymbol{U})]E_{\mathrm{xu}}[\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X}, \boldsymbol{U})]^\top$
IF_γ	Influence function over J_γ
$\boldsymbol{V}_\gamma$	$\boldsymbol{V}_\gamma := E_{\mathrm{xu}}[\eta(\boldsymbol{X}, \boldsymbol{U})^{\frac{\gamma}{1+\gamma}} S(\boldsymbol{X}, \boldsymbol{U})^{\frac{1}{1+\gamma}} \boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X}, \boldsymbol{U})\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X}, \boldsymbol{U})^\top]$
$S(\boldsymbol{x}, \boldsymbol{u})$	$S(\boldsymbol{x}, \boldsymbol{u}) := \frac{1}{1+e^{(1+\gamma)r_{\boldsymbol{\theta}^\star}(\boldsymbol{x}, \boldsymbol{u})}}$
$\eta(\boldsymbol{x}, \boldsymbol{u})$	$\eta(\boldsymbol{x}, \boldsymbol{u}) := S(\boldsymbol{x}, \boldsymbol{u})(1 - S(\boldsymbol{x}, \boldsymbol{u}))$
$\bar{E}_{\mathrm{xu}}$	Expectation over $\bar{p}(\boldsymbol{x}, \boldsymbol{u})$
$\bar{E}_{\mathrm{x} \times \mathrm{u}}$	Expectation over $\bar{p}(\boldsymbol{x})\bar{p}(\boldsymbol{u})$
$\bar{J}_\gamma$	γ-cross entropy over $\bar{p}(\boldsymbol{x}, \boldsymbol{u})$ and $\bar{p}(\boldsymbol{x})\bar{p}(\boldsymbol{u})$

4.1 Robust representation learning based on the γ-cross entropy

The first method is based on the following γ-cross entropy for binary classification (Fujisawa and Eguchi, 2008; Hung et al., 2018). Let us recall the following two datasets:

$$\mathcal{D}_+ := \{(\boldsymbol{x}(t), \boldsymbol{u}(t)\}_{t=1}^T \sim p(\boldsymbol{x}, \boldsymbol{u}) \quad \text{vs.} \quad \mathcal{D}_- := \{(\boldsymbol{x}(t), \boldsymbol{u}^*(t))\}_{t=1}^T \sim p(\boldsymbol{x})p(\boldsymbol{u}).$$

By assigning class labels $y = 0$ and $y = 1$ to $\mathcal{D}_+$ and $\mathcal{D}_-$ respectively, the γ-cross entropy for posterior probability estimation can be formulated as

$$J_\gamma(f_+, f_-) := -\frac{1}{\gamma} \log \left[\iint \sum_{y=0}^1 p(y, \boldsymbol{x}, \boldsymbol{u}) \left(\frac{\{f_+(\boldsymbol{x}, \boldsymbol{u})^{\gamma+1}\}^{1-y} \{f_-(\boldsymbol{x}, \boldsymbol{u})^{\gamma+1}\}^y}{f_+(\boldsymbol{x}, \boldsymbol{u})^{\gamma+1} + f_-(\boldsymbol{x}, \boldsymbol{u})^{\gamma+1}} \right)^{\frac{\gamma}{\gamma+1}} d\boldsymbol{x}d\boldsymbol{u} \right],$$

$$= -\frac{1}{\gamma} \log \left[p(y=0)E_{\text{xu}} \left[\left(\frac{f_+(\boldsymbol{X}, \boldsymbol{U})^{\gamma+1}}{f_+(\boldsymbol{X}, \boldsymbol{U})^{\gamma+1} + f_-(\boldsymbol{X}, \boldsymbol{U})^{\gamma+1}} \right)^{\frac{\gamma}{\gamma+1}} \right] \right.$$

$$\left. + p(y=1)E_{\text{x}\times\text{u}} \left[\left(\frac{f_-(\boldsymbol{X}, \boldsymbol{U})^{\gamma+1}}{f_+(\boldsymbol{X}, \boldsymbol{U})^{\gamma+1} + f_-(\boldsymbol{X}, \boldsymbol{U})^{\gamma+1}} \right)^{\frac{\gamma}{\gamma+1}} \right] \right],$$

where f_+ and f_- are models for posterior probabilities, $p(y=0)$ and $p(y=1)$ are class probabilities, and we used the following relation based on the datasets $\mathcal{D}_+$ and $\mathcal{D}_-$:

$$p(\boldsymbol{x}, \boldsymbol{u}|y=0) = p(\boldsymbol{x}, \boldsymbol{u}) \quad \text{and} \quad p(\boldsymbol{x}, \boldsymbol{u}|y=1) = p(\boldsymbol{x})p(\boldsymbol{u}).$$

By denoting $\log \frac{f_+(\boldsymbol{x}, \boldsymbol{u})}{f_-(\boldsymbol{x}, \boldsymbol{u})}$ by $r(\boldsymbol{x}, \boldsymbol{u})$ and assuming symmetric class probabilities (i.e., $p(y=0) = p(y=1) = \frac{1}{2}$), the γ-cross entropy can be written as

$$J_\gamma(r) := -\frac{1}{\gamma} \log \left[E_{\text{xu}} \left[\left(\frac{e^{(\gamma+1)r(\boldsymbol{X}, \boldsymbol{U})}}{1 + e^{(\gamma+1)r(\boldsymbol{X}, \boldsymbol{U})}} \right)^{\frac{\gamma}{\gamma+1}} \right] + E_{\text{x}\times\text{u}} \left[\left(\frac{1}{1 + e^{(\gamma+1)r(\boldsymbol{X}, \boldsymbol{U})}} \right)^{\frac{\gamma}{\gamma+1}} \right] \right], \tag{22}$$

where the term related to $p(y=0)$ and $p(y=1)$ is omitted because it is irrelevant to estimation of a model r. Since $r(\boldsymbol{x}, \boldsymbol{u})$ is a model for the log-ratio of posterior probabilities, $J_\gamma(r)$ is minimized at

$$\log \frac{p(y=0|\boldsymbol{x}, \boldsymbol{u})}{p(y=1|\boldsymbol{x}, \boldsymbol{u})} = \log \frac{p(\boldsymbol{x}, \boldsymbol{u}|y=0)p(y=0)}{p(\boldsymbol{x}, \boldsymbol{u}|y=1)p(y=1)} = \log \frac{p(\boldsymbol{x}, \boldsymbol{u})}{p(\boldsymbol{x})p(\boldsymbol{u})}.$$

Thus, the γ-cross entropy (22) can be used for density ratio estimation. As proven in Fujisawa and Eguchi (2008), the cross entropy in logistic regression can be obtained in the limit of the γ-cross entropy as follows:

$$\lim_{\gamma \to 0} J_\gamma(r) = J_{\text{LR}}(r),$$

indicating $J_\gamma(r)$ can be regarded as a generalization of the cross entropy in logistic regression. The remarkable property of $J_\gamma(r)$ is robustness against outliers. The positive parameter γ controls the

robustness, and a larger value of γ tends to be more robust to outliers. We theoretically characterize the robustness of the γ-cross entropy in the context of density ratio estimation later. On the other hand, as discussed in Fujisawa and Eguchi (2008), there seems to exist a trade-off between robustness and efficiency, indicating that outlier-robustness for the γ-cross entropy may come at a price of making estimation sample-inefficient. In fact, the experimental results in Section 5.1.1 partially demonstrate this trade-off: The proposed method based on the γ-cross entropy is the most outlier-robust method, but it is not the best in terms of sample-efficiency.

In practice, we empirically approximate $J_\gamma(r)$ as

$$\widehat{J}_\gamma(r) := -\frac{1}{\gamma} \log \left[\frac{1}{T} \sum_{t=1}^{T} \left\{ \left(\frac{e^{(\gamma+1)r(\boldsymbol{x}(t),\boldsymbol{u}(t))}}{1 + e^{(\gamma+1)r(\boldsymbol{x}(t),\boldsymbol{u}(t))}} \right)^{\frac{\gamma}{\gamma+1}} + \left(\frac{1}{1 + e^{(\gamma+1)r(\boldsymbol{x}(t),\boldsymbol{u}_{\mathrm{p}}(t))}} \right)^{\frac{\gamma}{\gamma+1}} \right\} \right],$$

where $\boldsymbol{u}_{\mathrm{p}}(t)$ denotes a random permutation of $\boldsymbol{u}(t)$ with respect to t. Another empirical approximation is also possible as

$$\widetilde{J}_\gamma(r) := -\frac{1}{\gamma} \log \left[\frac{1}{T} \sum_{t=1}^{T} \left(\frac{e^{(\gamma+1)r(\boldsymbol{x}(t),\boldsymbol{u}(t))}}{1 + e^{(\gamma+1)r(\boldsymbol{x}(t),\boldsymbol{u}(t))}} \right)^{\frac{\gamma}{\gamma+1}} + \frac{1}{T^2} \sum_{t=1}^{T} \sum_{t'=1}^{T} \left(\frac{1}{1 + e^{(\gamma+1)r(\boldsymbol{x}(t),\boldsymbol{u}(t'))}} \right)^{\frac{\gamma}{\gamma+1}} \right].$$

These empirical objective functions are minimized with a minibatch stochastic gradient method in Section 5.

4.2 Nonlinear ICA with the Donsker-Varadhan variational estimation

Our second method is based on variational estimation of mutual information (Ruderman et al., 2012; Belghazi et al., 2018), and employs the negative of the lower-bound in (7) as the objective function:

$$J_{\mathrm{DV}}(r) := -E_{\mathrm{xu}}[r(\boldsymbol{X},\boldsymbol{U})] + \log(E_{\mathrm{x\times u}}[e^{r(\boldsymbol{X},\boldsymbol{U})}]). \tag{23}$$

As proved in Banerjee (2006) and Belghazi et al. (2018), $J_{\mathrm{DV}}(r)$ is minimized at $\log \frac{p(\boldsymbol{x},\boldsymbol{u})}{p(\boldsymbol{x})p(\boldsymbol{u})}$ up to some constant, and thus can be used for density ratio estimation. $J_{\mathrm{DV}}(r)$ has been already employed in the context of maximization of mutual information (Hjelm et al., 2019). Here, our contribution is to apply $J_{\mathrm{DV}}(r)$ to nonlinear ICA, and we numerically demonstrate its usefulness.

The objective function $J_{\mathrm{DV}}(r)$ has been independently derived in terms of density ratio estimation based on the KL-divergence (Sugiyama et al., 2008; Tsuboi et al., 2009). Since $r(\boldsymbol{x},\boldsymbol{u})$ is a model for $\log \frac{p(\boldsymbol{x},\boldsymbol{u})}{p(\boldsymbol{x})p(\boldsymbol{u})}$, the following model $p_m(\boldsymbol{x},\boldsymbol{u})$ can be regarded as a density model for $p(\boldsymbol{x},\boldsymbol{u})$:

$$p_m(\boldsymbol{x},\boldsymbol{u}) := \frac{p(\boldsymbol{x})p(\boldsymbol{u})e^{r(\boldsymbol{x},\boldsymbol{u})}}{E_{\mathrm{x\times u}}[e^{r(\boldsymbol{X},\boldsymbol{U})}]},$$

where the denominator ensures that $p_m(\boldsymbol{x},\boldsymbol{u})$ is normalized to be one. Then, the KL-divergence between $p(\boldsymbol{x},\boldsymbol{u})$ and $p_m(\boldsymbol{x},\boldsymbol{u})$ is given by

$$\begin{aligned}
\mathrm{KL}[p(\boldsymbol{x},\boldsymbol{u})||p_m(\boldsymbol{x},\boldsymbol{u})] &= \iint p(\boldsymbol{x},\boldsymbol{u}) \log \frac{p(\boldsymbol{x},\boldsymbol{u})}{p_m(\boldsymbol{x},\boldsymbol{u})} \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u} \\
&= I(\mathcal{X},\mathcal{U}) \underbrace{-E_{\mathrm{xu}}[r(\boldsymbol{X},\boldsymbol{U})] + \log\left(E_{\mathrm{x\times u}}[e^r(\boldsymbol{X},\boldsymbol{U})]\right)}_{=J_{\mathrm{DV}}(r)}.
\end{aligned} \tag{24}$$

Eq.(24) shows that minimizing $\mathrm{KL}[p(\boldsymbol{x}, \boldsymbol{u}) || p_m(\boldsymbol{x}, \boldsymbol{u})]$ is equal to minimizing $J_{\mathrm{DV}}(r)$ because the mutual information $I(\mathcal{X}, \mathcal{U})$ on the right-hand side is a constant with respect to r. As shown in Kanamori et al. (2010), the density ratio estimator based on the KL-divergence could be more accurate than the estimator based on the cross entropy J_{LR} (i.e., logistic regression) under the miss-specified setting where the true density ratio is not necessarily included in the function class of a model r. In fact, we experimentally demonstrate that the nonlinear ICA method based on J_{DV} performs better than an existing method based on J_{LR} when the number of data samples is small.

In practice, we empirically approximate $J_{\mathrm{DV}}(r)$ as

$$\widehat{J}_{\mathrm{DV}}(r) := -\frac{1}{T}\sum_{t=1}^{T} r(\boldsymbol{x}(t), \boldsymbol{u}(t)) + \log\left(\frac{1}{T}\sum_{t=1}^{T} e^{r(\boldsymbol{x}(t), \boldsymbol{u}_{\mathrm{p}}(t))}\right),$$

or

$$\widetilde{J}_{\mathrm{DV}}(r) := -\frac{1}{T}\sum_{t=1}^{T} r(\boldsymbol{x}(t), \boldsymbol{u}(t)) + \log\left(\frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T} e^{r(\boldsymbol{x}(t), \boldsymbol{u}(t'))}\right).$$

Section 5 uses both approximations in numerical experiments with a mini-batch stochastic gradient method.

4.3 Theoretical analysis for outlier-robustness

Here, we investigate outlier-robustness of estimators based on J_γ and J_{DV}. To this end, we consider the following two scenarios of contamination by outliers:

- *Contamination model 1*: Given input data $\boldsymbol{x}$, complementary data $\boldsymbol{u}$ is conditionally contaminated by outliers as follows:

$$\bar{p}(\boldsymbol{u}|\boldsymbol{x}) := (1 - \epsilon)p(\boldsymbol{u}|\boldsymbol{x}) + \epsilon\delta(\boldsymbol{u}|\boldsymbol{x}), \tag{25}$$

  where $\bar{p}(\boldsymbol{u}|\boldsymbol{x})$ is the contaminated conditional density of $\boldsymbol{u}$ given $\boldsymbol{x}$, $\delta(\boldsymbol{u}|\boldsymbol{x})$ denotes the conditional density for outliers, and $0 \leq \epsilon < 1$ denotes the contamination ratio. By assuming that input data $\boldsymbol{x}$ is *non*contaminated, (25) leads to the following contaminated joint and marginal densities:

$$\bar{p}(\boldsymbol{x}, \boldsymbol{u}) := (1 - \epsilon)p(\boldsymbol{x}, \boldsymbol{u}) + \epsilon\delta(\boldsymbol{u}|\boldsymbol{x})p(\boldsymbol{x}),$$

$$\bar{p}(\boldsymbol{x}) := p(\boldsymbol{x}) \quad \text{and} \quad \bar{p}(\boldsymbol{u}) := (1 - \epsilon)p(\boldsymbol{u}) + \epsilon\int \delta(\boldsymbol{u}|\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x},$$

  where $\bar{p}(\boldsymbol{x}, \boldsymbol{u})$ is the contaminated joint density, and $\bar{p}(\boldsymbol{x})$ and $\bar{p}(\boldsymbol{u})$ denote the contaminated marginal densities of $\boldsymbol{x}$ and $\boldsymbol{u}$, respectively.

- *Contamination model 2*: Input data $\boldsymbol{x}$ and complementary data $\boldsymbol{u}$ are jointly contaminated by outliers:

$$\bar{p}(\boldsymbol{x}, \boldsymbol{u}) := (1 - \epsilon)p(\boldsymbol{x}, \boldsymbol{u}) + \epsilon\delta(\boldsymbol{x}, \boldsymbol{u}), \tag{26}$$

Table 5: Comparison of contamination models in influence function analysis (Propositions 5 and 6). $\delta_{\bar{x}}(x)$ and $\delta_{\bar{u}}(u)$ denote the Delta functions having point masses at $\bar{x}$ and $\bar{u}$, respectively.

Contamination model 1	Contamination model 2
$\bar{p}(x, u) = (1 - \epsilon)p(x, u) + \epsilon\delta_{\bar{u}}(u)p(x)$	$\bar{p}(x, u) = (1 - \epsilon)p(x, u) + \epsilon\delta_{\bar{x}}(x)\delta_{\bar{u}}(u)$
$\bar{p}(x) = p(x)$	$\bar{p}(x) = (1 - \epsilon)p(x) + \epsilon\delta_{\bar{x}}(x)$
$\bar{p}(u) = (1 - \epsilon)p(u) + \epsilon\delta_{\bar{u}}(u)$	$\bar{p}(u) = (1 - \epsilon)p(u) + \epsilon\delta_{\bar{u}}(u)$

where $\delta(x, u)$ denotes the joint density for outliers. Eq.(26) leads to the following contaminated marginal densities:

$$\bar{p}(x) := (1 - \epsilon)p(x) + \epsilon\delta(x) \quad \text{and} \quad \bar{p}(u) := (1 - \epsilon)p(u) + \epsilon\delta(u),$$

where

$$\delta(x) := \int \delta(x, u)\mathrm{d}u \quad \text{and} \quad \delta(u) := \int \delta(x, u)\mathrm{d}x.$$

The fundamental difference between these contamination scenarios is whether or not input data x is contaminated by outliers. Next, we perform two analyses for outlier-robustness: One analysis is performed under the condition that the contamination ratio ϵ is small, while another one does *not* necessarily assume that ϵ is small and thus heavy contamination of outliers is also within the scope of the analysis.

4.3.1 INFLUENCE FUNCTION ANALYSIS

Here, we perform influence function analysis (Hampel et al., 2011), which is an established tool in robust statistics and assumes that the densities for outliers are given as follows:

$$\delta(x, u) = \delta_{\bar{x}}(x)\delta_{\bar{u}}(u), \quad \delta(x) = \delta_{\bar{x}}(x) \quad \text{and} \quad \delta(u) = \delta_{\bar{u}}(u), \tag{27}$$

where $\delta_{\bar{x}}(x)$ and $\delta_{\bar{u}}(u)$ are the Dirac delta functions having point masses at $\bar{x}$ and $\bar{u}$, respectively. Eq.(27) implies that $\delta(u|x) = \delta_{\bar{u}}(u)$ and $\int \delta(u|x)p(x)\mathrm{d}x = \delta_{\bar{u}}(u)$ in the contamination model 1. Table 5 is a comparison of contamination models 1 and 2 in this analysis.

In this analysis, we suppose that a model $r_\theta(x, u)$ is parametrized by θ and has the following properties:

- $r_\theta(x, u)$ and its gradient $g_\theta(x, u)$ are continuous where $g_\theta(x, u) := \nabla_\theta r_\theta(x, u)$.

- $|r_\theta(x, u)| \to \infty$ and $\|g_\theta(x, u)\| \to \infty$ as $\|x\| \to \infty$ and/or $\|u\| \to \infty$.

These properties are simply used to discuss the robustness of the estimators based on the influence function, and satisfied by standard models such neural networks with an unbounded activation function (e.g., the softplus function). Based on $r_\theta(x, u)$, we define two estimators as

$$\theta^\star := \underset{\theta}{\mathrm{argmin}}\, J(r_\theta) \quad \text{and} \quad \theta_\epsilon := \underset{\theta}{\mathrm{argmin}}\, \bar{J}(r_\theta),$$

where J is some objective function over the *non*contaminated densities (e.g., J_γ), while $\bar{J}$ is the contaminated version of J computed over the contaminated densities $\bar{p}(\boldsymbol{x}, \boldsymbol{u})$, $\bar{p}(\boldsymbol{x})$ and $\bar{p}(\boldsymbol{u})$ instead of $p(\boldsymbol{x}, \boldsymbol{u})$, $p(\boldsymbol{x})$ and $p(\boldsymbol{u})$. Then, we define the *influence function* as

$$\mathrm{IF}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}) = \lim_{\epsilon \to 0} \frac{\boldsymbol{\theta}^\star - \boldsymbol{\theta}_\epsilon}{\epsilon}. \tag{28}$$

Eq.(28) indicates that $\|\mathrm{IF}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}})\|$ measures how the estimator $\boldsymbol{\theta}^\star$ is influenced by small contamination of outliers $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{u}}$, and a larger $\|\mathrm{IF}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}})\|$ implies stronger influence by outliers. Based on the influence function, some robust properties of estimators can be characterized: Estimator $\boldsymbol{\theta}^\star$ is called *B-robust* if $\sup_{\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}} \|\mathrm{IF}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}})\| < \infty$ (Hampel et al., 2011). B-robustness guarantees that the influence from outliers is limited. A more favorable property is the *redescending property* (Hampel et al., 2011), 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 $\boldsymbol{\theta}^\star$ has almost no influence from even strongly deviated data $\bar{\boldsymbol{x}}$ and/or $\bar{\boldsymbol{u}}$.

First, the following proposition shows the forms of the influence functions for J_{DV}:

Proposition 5. *Assume that $r_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{u}) = \log \frac{p(\boldsymbol{x}, \boldsymbol{u})}{p(\boldsymbol{x})p(\boldsymbol{u})}$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^\star$. For the contamination model 1, the influence function of the estimator $\boldsymbol{\theta}^\star$ based on J_{DV} is given by*

$$\mathrm{IF}_{\mathrm{DV}}(\bar{\boldsymbol{u}}) = \boldsymbol{V}_{\mathrm{DV}}^{-1} E_{\mathrm{x}}[\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X}, \bar{\boldsymbol{u}}) - e^{r_{\boldsymbol{\theta}^\star}(\boldsymbol{X}, \bar{\boldsymbol{u}})} \boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X}, \bar{\boldsymbol{u}})], \tag{29}$$

where $\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{x}, \boldsymbol{u}) = \nabla_{\boldsymbol{\theta}} r_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{u})|_{\boldsymbol{\theta} = \boldsymbol{\theta}^\star}$ and

$$\boldsymbol{V}_{\mathrm{DV}} := -E_{\mathrm{xu}}[\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X}, \boldsymbol{U})\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X}, \boldsymbol{U})^\top] + E_{\mathrm{xu}}[\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X}, \boldsymbol{U})]E_{\mathrm{xu}}[\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X}, \boldsymbol{U})]^\top. \tag{30}$$

Regarding the contamination model 2, the influence function is given by

$$\begin{aligned}
\mathrm{IF}_{\mathrm{DV}}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}) = \boldsymbol{V}_{\mathrm{DV}}^{-1} \Big[&\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}) + E_{\mathrm{xu}}[\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X}, \boldsymbol{U})] \\
&- E_{\mathrm{x}}[e^{r_{\boldsymbol{\theta}^\star}(\boldsymbol{X}, \bar{\boldsymbol{u}})} \boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X}, \bar{\boldsymbol{u}})] - E_{\mathrm{u}}[e^{r_{\boldsymbol{\theta}^\star}(\bar{\boldsymbol{x}}, \boldsymbol{U})} \boldsymbol{g}_{\boldsymbol{\theta}^\star}(\bar{\boldsymbol{x}}, \boldsymbol{U})] \Big],
\end{aligned} \tag{31}$$

where E_{x} and E_{u} denote the expectations over $p(\boldsymbol{x})$ and $p(\boldsymbol{u})$, respectively.

The proof is deferred in Appendix G. The key to interpret the influence functions in Proposition 5 is unboundedness of $r_{\boldsymbol{\theta}^\star}(\boldsymbol{x}, \boldsymbol{u})$ and its gradient $\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{x}, \boldsymbol{u})$. For example, the right-hand side of (29) indicates that $\|\mathrm{IF}_{\mathrm{DV}}(\bar{\boldsymbol{u}})\|$ can be unbounded when $r_{\boldsymbol{\theta}^\star}(\boldsymbol{x}, \bar{\boldsymbol{u}}) \to \infty$ and $\|\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{x}, \bar{\boldsymbol{u}})\| \to \infty$ as $\|\bar{\boldsymbol{u}}\| \to \infty$. This means that the estimator based on J_{DV} is not B-robust, and can be strongly influenced by outliers $\bar{\boldsymbol{u}}$. The same discussion is applicable to the influence function (31) of the contamination model 2. Thus, these results imply that estimation based on J_{DV} can be sensitive to outliers.

Next, we show the influence functions for the γ-cross entropy J_γ in the following proposition:

Proposition 6. *Suppose the same assumption holds as Proposition 5. Regarding the contamination model 1, the influence function of the estimator $\boldsymbol{\theta}^\star$ based on J_γ is obtained as follows:*

$$\mathrm{IF}_\gamma(\bar{\boldsymbol{u}}) = V_\gamma^{-1} E_{\mathrm{x}}[\{S(\boldsymbol{X}, \bar{\boldsymbol{u}})^{\frac{1}{1+\gamma}} - (1 - S(\boldsymbol{X}, \bar{\boldsymbol{u}}))^{\frac{1}{1+\gamma}}\} \eta(\boldsymbol{X}, \bar{\boldsymbol{u}})^{\frac{\gamma}{1+\gamma}} \boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X}, \bar{\boldsymbol{u}})], \tag{32}$$

where

$$V_\gamma := E_{\mathrm{xu}}[\eta(\boldsymbol{X},\boldsymbol{U})^{\frac{\gamma}{1+\gamma}} S(\boldsymbol{X},\boldsymbol{U})^{\frac{1}{1+\gamma}} \boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})^\top], \tag{33}$$

$$S(\boldsymbol{x},\boldsymbol{u}) := \frac{1}{1+e^{(1+\gamma)r_{\boldsymbol{\theta}^\star}(\boldsymbol{x},\boldsymbol{u})}} \quad \text{and} \quad \eta(\boldsymbol{x},\boldsymbol{u}) := S(\boldsymbol{x},\boldsymbol{u})(1-S(\boldsymbol{x},\boldsymbol{u})). \tag{34}$$

Regarding the contamination model 2, the influence function is given by

$$\begin{aligned}
\mathrm{IF}_\gamma(\bar{\boldsymbol{x}},\bar{\boldsymbol{u}}) = V_\gamma^{-1}\Big[& \eta(\bar{\boldsymbol{x}},\bar{\boldsymbol{u}})^{\frac{\gamma}{1+\gamma}} S(\bar{\boldsymbol{x}},\bar{\boldsymbol{u}})^{\frac{1}{1+\gamma}} \boldsymbol{g}_{\boldsymbol{\theta}^\star}(\bar{\boldsymbol{x}},\bar{\boldsymbol{u}}) \\
& - E_{\mathrm{x}}[\eta(\boldsymbol{X},\bar{\boldsymbol{u}})^{\frac{\gamma}{1+\gamma}} (1-S(\boldsymbol{X},\bar{\boldsymbol{u}}))^{\frac{1}{1+\gamma}} \boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\bar{\boldsymbol{u}})] \\
& - E_{\mathrm{u}}[\eta(\bar{\boldsymbol{x}},\boldsymbol{U})^{\frac{\gamma}{1+\gamma}} (1-S(\bar{\boldsymbol{x}},\boldsymbol{U}))^{\frac{1}{1+\gamma}} \boldsymbol{g}_{\boldsymbol{\theta}^\star}(\bar{\boldsymbol{x}},\boldsymbol{U})] \\
& + E_{\mathrm{x}\times\mathrm{u}}[\eta(\boldsymbol{X},\boldsymbol{U})^{\frac{\gamma}{1+\gamma}} (1-S(\boldsymbol{X},\boldsymbol{U}))^{\frac{1}{1+\gamma}} \boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})]\Big].
\end{aligned} \tag{35}$$

The proof of Proposition 6 is not given because it is almost the same as Proposition 5. In contrast to J_{DV}, the influence functions based on the γ-cross entropy J_γ include the weight functions $\eta(\boldsymbol{x},\boldsymbol{u})$ and $S(\boldsymbol{x},\boldsymbol{u})$, and $\|\mathrm{IF}_\gamma(\bar{\boldsymbol{u}})\|$ and $\|\mathrm{IF}_\gamma(\bar{\boldsymbol{x}},\bar{\boldsymbol{u}})\|$ can be bounded even when $\|\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\bar{\boldsymbol{x}},\bar{\boldsymbol{u}})\| \to \infty$ as $\|\bar{\boldsymbol{x}}\| \to \infty$ and/or $\|\bar{\boldsymbol{u}}\| \to \infty$. The following corollary characterizes the robustness of the γ-cross entropy and shows that in stark contrast with J_{DV}, the estimator based on J_γ is B-robust for both contamination models, and has the redescending property for the contamination model 1 under some conditions:

Corollary 7. *Assume that*

$$\sup_{\boldsymbol{x},\boldsymbol{u}} \|\eta(\boldsymbol{x},\boldsymbol{u})^{\frac{\gamma}{1+\gamma}} \boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{x},\boldsymbol{u})\| < \infty. \tag{36}$$

Then, the estimator $\boldsymbol{\theta}^\star$ based on J_γ is B-robust both for the contamination models 1 and 2. Another assumption is made as follows:

$$\lim_{\|\bar{\boldsymbol{u}}\|\to\infty} \|\eta(\boldsymbol{x},\bar{\boldsymbol{u}})^{\frac{\gamma}{1+\gamma}} \boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{x},\bar{\boldsymbol{u}})\| = 0 \quad \text{for all } \boldsymbol{x}. \tag{37}$$

Then, the estimator $\boldsymbol{\theta}^\star$ based on J_γ has the redescending property for the contamination model 1.

The results are an almost direct consequence of Proposition 6, and thus the detailed calculation is omitted. Assumptions (36) and (37) would be mild: By definition (34), the function $\eta(\boldsymbol{x},\boldsymbol{u})$ exponentially convergences to zero when $r_{\boldsymbol{\theta}^\star}(\boldsymbol{x},\boldsymbol{u}) \to \infty$ as $\|\boldsymbol{x}\| \to \infty$ and/or $\|\boldsymbol{u}\| \to \infty$. Thus, B-robustness would hold under both contamination models for standard models (e.g., neural networks). On the other hand, the redescending property is shown only for the contamination model 1. This comes from the fact that the contamination of outliers is more complicated in the contamination model 2: Both $\boldsymbol{x}$ and $\boldsymbol{u}$ are contaminated by outliers in the contamination model 2, while $\boldsymbol{u}$ is only contaminated in the contamination model 1. However, we numerically demonstrate that the proposed method based on the γ-cross entropy is robust against outliers even for the contamination model 2.

Proposition 5 also reveals the outlier weakness of the cross entropy J_{LR} used in logistic regression. We recall that the γ-cross entropy J_γ approaches J_{LR} as $\gamma \to 0$. Then, it is obvious that Assumptions (36) and (37) in Corollary 7 are never satisfied in the limit of $\gamma \to 0$ when $\|\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{x},\boldsymbol{u})\|$

is an unbounded function (e.g., neural networks). This implies that the estimator based on J_{LR} is neither B-robust nor redescending, and estimation based on J_{LR} can be hampered by outliers.

In short, the analysis based on the influence function implies that it is a promising approach to use the γ-cross entropy for density ratio estimation in the presence of outliers even with neural networks, while estimation based on J_{DV} and J_{LR} can be strongly influenced by outliers particularly when $r_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{u})$ is modelled by using unbounded functions such as neural networks.

4.3.2 STRONG ROBUSTNESS OF THE γ-CROSS ENTROPY

Here, we investigate the robustness of the γ-cross entropy even when the contamination ratio ϵ is *not* necessarily assumed to be small. Unlike influence function analysis, the contaminated densities, $\delta(\boldsymbol{x}, \boldsymbol{u})$, $\delta(\boldsymbol{x})$ and $\delta(\boldsymbol{u})$, are not Dirac delta functions, but rather general probability density functions. To clarify the notations, we denote the γ-cross entropy over the contaminated densities as

$$
\bar{J}_{\gamma}(r) := -\frac{1}{\gamma} \log \left[\bar{E}_{\mathrm{xu}} \left[\left(\frac{e^{(\gamma+1)r(\boldsymbol{X},\boldsymbol{U})}}{1 + e^{(\gamma+1)r(\boldsymbol{X},\boldsymbol{U})}} \right)^{\frac{\gamma}{\gamma+1}} \right] + \bar{E}_{\mathrm{x}\times\mathrm{u}} \left[\left(\frac{1}{1 + e^{(\gamma+1)r(\boldsymbol{X},\boldsymbol{U})}} \right)^{\frac{\gamma}{\gamma+1}} \right] \right],
$$

where $\bar{E}_{\mathrm{xu}}$ and $\bar{E}_{\mathrm{x}\times\mathrm{u}}$ denote the expectations over $\bar{p}(\boldsymbol{x}, \boldsymbol{u})$ and $\bar{p}(\boldsymbol{x})\bar{p}(\boldsymbol{u})$ in contamination models, respectively.

The remarkable property of the γ-cross entropy is *strong robustness* (Fujisawa and Eguchi, 2008): The latent bias caused from outliers can be small even in the case of heavy contamination (i.e., nonsmall ϵ). To prove the strong robustness in the context of density ratio estimation, we recall that $r(\boldsymbol{x}, \boldsymbol{u})$ is defined by

$$
r(\boldsymbol{x}, \boldsymbol{u}) = \psi(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}), \boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u})) + a(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x})) + b(\boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u})).
$$

The following proposition proved in Appendix H focuses on the contamination model 1 and establishes certain conditions for strong robustness:

Proposition 8. *Let us define a constant ν_1 as*

$$
\nu_1 := \iint p(\boldsymbol{x})\delta(\boldsymbol{u}|\boldsymbol{x}) \left(\frac{e^{(\gamma+1)r(\boldsymbol{x},\boldsymbol{u})}}{1 + e^{(\gamma+1)r(\boldsymbol{x},\boldsymbol{u})}} \right)^{\frac{\gamma}{\gamma+1}} \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u} + \iint p(\boldsymbol{x})\delta(\boldsymbol{u}) \left(\frac{1}{1 + e^{(\gamma+1)r(\boldsymbol{x},\boldsymbol{u})}} \right)^{\frac{\gamma}{\gamma+1}} \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u}.
$$

We suppose the contamination model 1 and make the following assumptions:

(D1) Assume that

$$
\iint p(\boldsymbol{x})\delta(\boldsymbol{u}|\boldsymbol{x}) \left(\frac{e^{(\gamma+1)\{\psi(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}),\boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u}))+a(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}))\}/2}}{e^{-(\gamma+1)b(\boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u}))} + e^{(\gamma+1)\{\psi(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}),\boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u}))+a(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}))\}}} \right)^{\frac{2\gamma}{\gamma+1}} \mathrm{d}\boldsymbol{u}\mathrm{d}\boldsymbol{x} < \infty
$$

$$
\text{and } \iint p(\boldsymbol{x})\delta(\boldsymbol{u}) \left(\frac{e^{-(\gamma+1)b(\boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u}))/2}}{e^{-(\gamma+1)b(\boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u}))} + e^{(\gamma+1)\{\psi(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}),\boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u}))+a(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}))\}}} \right)^{\frac{2\gamma}{\gamma+1}} \mathrm{d}\boldsymbol{u}\mathrm{d}\boldsymbol{x} < \infty.
$$

(D2) The following integrals are sufficiently small:

$$
\iint p(\boldsymbol{x})\delta(\boldsymbol{u}|\boldsymbol{x})e^{\gamma\{\psi(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}),\boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u}))+a(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}))\}} \mathrm{d}\boldsymbol{u}\mathrm{d}\boldsymbol{x} \text{ and } \int \delta(\boldsymbol{u})e^{-\gamma b(\boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u}))} \mathrm{d}\boldsymbol{u}.
$$

Then, it holds that ν_1 is sufficient small and

$$\bar{J}_\gamma(r) = J_\gamma(r) - \frac{1}{\gamma}\log(1-\epsilon) + O\left(\frac{\epsilon}{1-\epsilon}\nu_1\right). \tag{38}$$

The proof can be found in Appendix H. Eq.(38) in Proposition 8 indicates that for the contamination model 1, minimization of $\bar{J}_\gamma(r)$ is approximately equal to minimization of $J_\gamma(r)$ over *non*contaminated densities under Assumptions (D1-2) even when the contamination ratio ϵ is not necessarily small (e.g., $\epsilon = 0.2$ as in the numerical experiments of Fujisawa and Eguchi (2008, Section 8)). This means that we could perform density ratio estimation in many practical situations almost as if outliers did not exist.

Assumption (D1) implies that $\delta(\boldsymbol{u}|\boldsymbol{x})$ and $\delta(\boldsymbol{u})$ are localized on the domain of $\boldsymbol{u}$ and have a fast decay on their tails as often seen in densities for outliers: For example, when both $\delta(\boldsymbol{u})$ and $\delta(\boldsymbol{u}|\boldsymbol{x})$ have finite supports[4], Assumption (D1) is fulfilled. Assumption (D2) means that outliers drawn from $\delta(\boldsymbol{u}|\boldsymbol{x})$ and $\delta(\boldsymbol{u})$ exist on the respective tails of $e^{\psi(\boldsymbol{h}_\mathrm{x}(\boldsymbol{x}),\boldsymbol{h}_\mathrm{u}(\boldsymbol{u}))+a(\boldsymbol{h}_\mathrm{x}(\boldsymbol{x}))}$ and $e^{-b(\boldsymbol{h}_\mathrm{u}(\boldsymbol{u}))}$, which can be roughly regarded as models of $p(\boldsymbol{u}|\boldsymbol{x})$ and $p(\boldsymbol{u})$ respectively because $r(\boldsymbol{x},\boldsymbol{u})$ is modelling $\log\frac{p(\boldsymbol{x},\boldsymbol{u})}{p(\boldsymbol{x})p(\boldsymbol{u})} = \log p(\boldsymbol{u}|\boldsymbol{x}) - \log p(\boldsymbol{u})$. Thus, when $\boldsymbol{r}(\boldsymbol{x},\boldsymbol{u})$ is near to $\log\frac{p(\boldsymbol{x},\boldsymbol{u})}{p(\boldsymbol{x})p(\boldsymbol{u})}$, Assumption (D2) implies that the outliers exist on the tails of $p(\boldsymbol{x},\boldsymbol{u})$ and $p(\boldsymbol{u})$, and indeed reflects typical contamination by outliers. Overall, Assumptions (D1-2) would be practically reasonable.

For the contamination model 2, the strong robustness does not hold, but the γ-cross entropy still has a favorable property which we call *semistrong robustness*. The detail of the semistrong robustness is given in Appendix I. In addition, we experimentally demonstrate that the proposed method based on the γ-cross entropy is robust against outliers under the contamination model 2.

5. Numerical Experiments

This section numerically investigates how the proposed methods work in nonlinear ICA and a downstream task for linear classification, and empirically supports the implication of the theoretical analysis for nonlinear ICA in Section 3.2.

5.1 Nonlinear ICA on artificial data

Here, we demonstrate the numerical performance of the proposed methods in nonlinear ICA and investigate how the dimensionality of complementary data affects the source recovery as implied in Section 3.2. All experiments suppose $d_\mathrm{x} = D_\mathrm{x}$ according to Proposition 2 and Theorem 3.

5.1.1 SAMPLE EFFICIENCY AND OUTLIER ROBUSTNESS

We followed the experimental setting of a nonlinear ICA method called *permutation contrastive learning* (PCL) in Hyvärinen and Morioka (2017), which is intended for temporally dependent data $\boldsymbol{x}(t)$, $t = 1,\ldots,T$ and supposes the complementary data samples are $\boldsymbol{u}(t) = \boldsymbol{x}(t-1)$ (i.e., past input data). First, as in the autoregressive process, the ten-dimensional temporally dependent T sources were generated from the contaminated density model as

$$\bar{p}(\boldsymbol{s}(t)|\boldsymbol{s}(t-1)) = (1-\epsilon)p(\boldsymbol{s}(t)|\boldsymbol{s}(t-1)) + \epsilon\delta(\boldsymbol{s}(t)|\boldsymbol{s}(t-1)), \tag{39}$$

---

4. The support of $\delta(\boldsymbol{u})$ is defined by $\{\boldsymbol{u} \mid \delta(\boldsymbol{u}) \neq 0\}$.

where ϵ denotes the outlier ratio, and with $\rho = 0.7$ and $D_{\mathrm{x}} = 10$,

$$p(\boldsymbol{s}(t)|\boldsymbol{s}(t-1)) \propto \prod_{i=1}^{D_{\mathrm{x}}} \exp\left(-\sqrt{2}|s_i(t) - \rho s_i(t-1)|/(1-\rho^2)\right) \tag{40}$$

$$\delta(\boldsymbol{s}(t)|\boldsymbol{s}(t-1)) \propto \prod_{i=1}^{D_{\mathrm{x}}} \exp\left(-\frac{\{s_i(t) + \rho s_i(t-1)\}^2}{2}\right). \tag{41}$$

Then, input data $\boldsymbol{x}$ was generated according to $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{s})$ in (8) where $\boldsymbol{f}$ was modelled by a multi-layer feedforward neural network with the leaky ReLU activation function and random weights. The numbers of all hidden and output units were the same as the dimensionality of data (i.e., D_{x}). Since complementary data $\boldsymbol{u}(t)$ is past input data $\boldsymbol{x}(t-1)$, the contamination process corresponds to the contamination model 2 in Section 4.3 where both $\boldsymbol{x}(t)$ and $\boldsymbol{u}(t)$ are contaminated by outliers. As preprocessing, we performed whitening based on the γ-cross entropy (Chen et al., 2013).

We applied the following methods to data samples:

- Permutation contrastive learning (PCL) (Hyvärinen and Morioka, 2017): A nonlinear ICA method for temporally dependent sources based on logistic regression whose objective function is an empirical version of J_{LR}.

- DV-PCL: A proposed nonlinear ICA method for temporally dependent sources based on the Donsker-Varadhan variational estimation $\widehat{J}_{\mathrm{DV}}$ (Section 4.2).

- Robust PCL (RPCL): A proposed nonlinear ICA method for temporally dependent sources based on the γ-cross entropy $\widehat{J}_{\gamma}$ (Section 4.1). The value of γ was increased from 0.0 to 5.0 during minibatch stochastic gradient at every 100 epoch where $\gamma = 0.0$ means to perform the standard logistic regression.

We used a same model $r(\boldsymbol{x}, \boldsymbol{u})$ for all three methods: The representation function $\boldsymbol{h}_{\mathrm{x}}$ was modelled by a multi-layer feedforward neural network where the number of hidden units was $4D_{\mathrm{x}}$, but the final layer was D_{x}. The number of layers was the same as $\boldsymbol{f}(\boldsymbol{s})$ in the data generative model (8). The activation functions in hidden layers were the max-out function (Goodfellow et al., 2013) with two groups, while the final layer has no activation function. Regarding the representation function $\boldsymbol{h}_{\mathrm{u}}$, the same architecture as $\boldsymbol{h}_{\mathrm{x}}$ was employed with parameter sharing (i.e., $\boldsymbol{h}_{\mathrm{u}} = \boldsymbol{h}_{\mathrm{x}}$). Following Hyvärinen and Morioka (2017), $r(\boldsymbol{x}, \boldsymbol{u})$ was modelled by

$$r(\boldsymbol{x}, \boldsymbol{u}) = \sum_{i=1}^{D_{\mathrm{x}}} |a_{i,1}h_{\mathrm{x},i}(\boldsymbol{x}) + a_{i,2}h_{\mathrm{u},i}(\boldsymbol{u}) + b_i| - (\bar{a}_i h_{\mathrm{x},i}(\boldsymbol{x}) + \bar{b}_i)^2 + c, \tag{42}$$

where $a_{i,1}, a_{i,2}, b_i, \bar{a}_i, \bar{b}_i, c$ are parameters to be estimated from data. All parameters were estimated by the Adam optimizer for $1{,}600$ epochs with mini-batch size 256 and learning rate 0.001. We also applied the ℓ_2 regularization to the weight parameters with the regularization parameter 10^{-4}.

With the test sources $s_i^{\mathrm{te}}(t)$ without outliers drawn from $p(\boldsymbol{s}(t)|\boldsymbol{s}(t-1))$, the performance was measured by the mean absolute correlation as

$$\frac{1}{T_{\mathrm{te}}D_{\mathrm{x}}} \sum_{i=1}^{D_{\mathrm{x}}} \left| \sum_{t=1}^{T_{\mathrm{te}}} (\hat{h}_{\mathrm{x},i}(\boldsymbol{x}^{\mathrm{te}}(t)) - \mu_i^{\hat{\mathrm{h}}})(s_{\pi(i)}^{\mathrm{te}}(t) - \mu_{\pi(i)}^{\mathrm{te}}) \right|,$$

where $\hat{h}_{\mathrm{x},i}$ denotes the learned representation function, $T_{\mathrm{te}} = 50,000$ denotes the number of test samples, and $\mu_i^{\hat{\mathrm{h}}}$ and μ_i^{te} are the sample means of $\hat{h}_{\mathrm{x},i}(\boldsymbol{x}^{\mathrm{te}}(t))$ and $s_i^{\mathrm{te}}(t)$, respectively. Since estimation of nonlinear ICA is indeterminant with respect to permutation (i.e., ordering) of the source components s_i, the permutation indices $\pi(i) \in \{1, 2, \ldots, D_{\mathrm{x}}\}$ were determined so that the absolute mean correlation was maximized.

Fig.1(a-c) investigate the sample efficiency of each method when $\epsilon = 0$ (i.e., no outliers) and the numbers of layers are one, two and three, respectively. When the number of samples is large, all methods well-recover the source components and work similarly. However, when the number of samples gets smaller, DV-PCL tends to perform better than PCL and RPCL. This presumably because the objective function J_{DV} in DV-PCL is based on the KL-divergence, and the density ratio estimator based on the KL-divergence would be accurate as implied in Kanamori et al. (2010). The same tendency can be seen when the numbers of layers are four (Fig.1(d)) and five (Fig.1(e)).

Fig.2 shows outlier-robustness of each nonlinear ICA method in $T = 100,000$. When the number of layers is one, all methods perform well (Fig.2(a)). This would be because we perform an outlier-robust whitening based on the γ-cross entropy (Chen et al., 2013) in preprocessing. However, when the number of layer is larger than one, the performance of DV-PCL and PCL quickly decreases as the contamination ratio ϵ is increased (Fig.2(b-e)). On the other hand, RPCL keeps high correlation values on a wide range of ϵ, and thus is very robust against outliers. These results are consistent with the theoretical results in Section 4.3 that estimators based on J_{DV} (DV-PCL) and J_{LR} (PCL) can be seriously hampered by outliers, while the γ-cross entropy (RPCL) is promising in the presence of outliers.

Overall, the proposed methods, DV-PCL and RPCL, have respective advantages over PCL. DV-PCL seems to be more useful in the limited number of samples, while RPCL is very robust against outliers particularly when the contamination ratio ϵ is large.

5.1.2 IMPORTANCE OF THE DIMENSIONALITY OF COMPLEMENTARY DATA

Next, we investigate how the dimensionality of complementary data $\boldsymbol{u}$ affects the source recovery in nonlinear ICA as implied in Proposition 2 and Theorem 3.

We followed the recovery conditions in Theorem 3. In order for the conditional density $p(\boldsymbol{s}|\boldsymbol{u})$ to be differentiable, the conditionally independent sources were first generated from

$$p(\boldsymbol{s}|\boldsymbol{u}) \propto \prod_{i=1}^{D_{\mathrm{x}}} \exp\left(-\log(\cosh(s_i - \boldsymbol{w}_i^{\mathrm{s}\top}\boldsymbol{u}))\right), \tag{43}$$

where $\boldsymbol{w}_i^{\mathrm{s}}$, $i = 1 \ldots, D_{\mathrm{x}}$ are D_{u}-dimensional vectors randomly determined from the independent uniform density on $[-1, 1]^{D_{\mathrm{u}}}$. The function $\log(\cosh(\cdot))$ is a smooth approximation of the absolute function $|\cdot|$, and thus $p(\boldsymbol{s}|\boldsymbol{u})$ can be regarded as a smoothed version of the Laplace density (40) in Section 5.1.1. This smooth approximation has been previously used in linear ICA as well (Hyvärinen, 1999). In this experiment, complementary data samples $\boldsymbol{u}(t)$ were simply drawn from the independent uniform density on $[0, 1]^{D_{\mathrm{u}}}$. The total number of samples is $T = 100,000$. Input data $\boldsymbol{x}$ was generated according to (8) where the mixing function $\boldsymbol{f}$ is modelled by a feedforward neural network with random connections.

Here, we used the nonlinear ICA methods based on the γ-cross entropy $\widehat{J}_{\gamma}$ (Section 4.1) and Donsker-Varadhan variational estimation $\widehat{J}_{\mathrm{DV}}$ (Section 4.2). As in Section 5.1.1, the representation function $\boldsymbol{h}_{\mathrm{x}}$ was modelled by a feedforward neural network where the number of hidden units

(a) $L = 1$ (b) $L = 2$ (c) $L = 3$

(d) $L = 4$ (e) $L = 5$
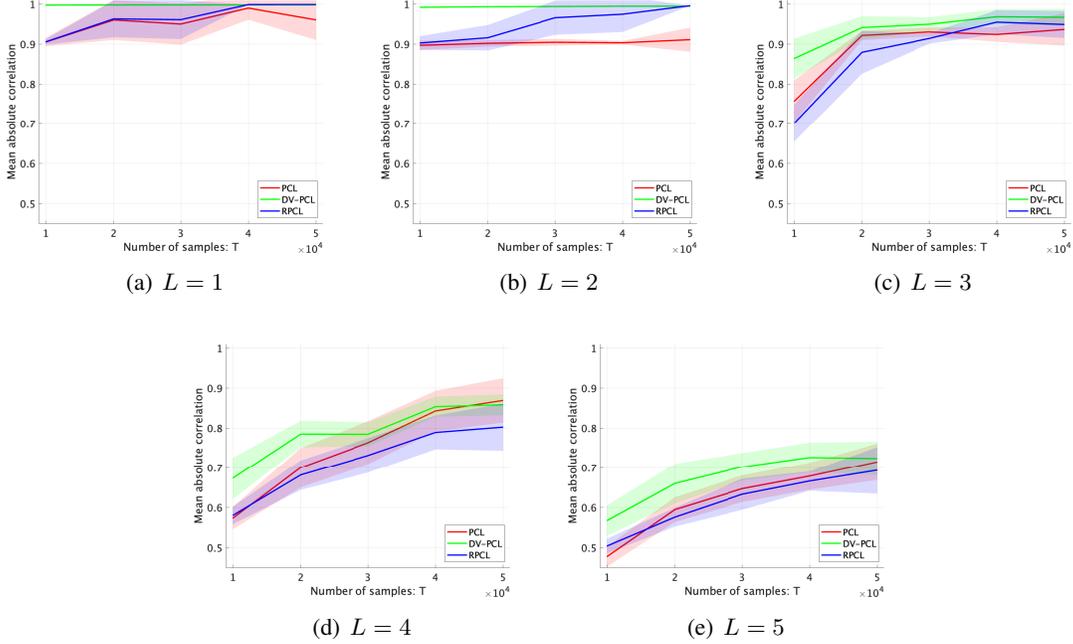
Figure 1: Sample efficiency of nonlinear ICA methods in the case of no outliers (i.e., $\epsilon = 0$). The averages of the mean absolute correlation were computed over 10 runs. L denotes the number of layers both in the nonlinear mixing function $\boldsymbol{f}$ in (8) and representation function $\boldsymbol{h}_\mathrm{x}$.

was $4D_\mathrm{x}$, but the final layer was D_x. The number of layers was the same as $\boldsymbol{f}$ in the data generative model. $\boldsymbol{h}_\mathrm{u}(\boldsymbol{u})$ was modelled by a one-layer neural network without the activation function as $\boldsymbol{h}_\mathrm{u}(\boldsymbol{u}) = \boldsymbol{W}_\mathrm{u}\boldsymbol{u} + \boldsymbol{b}_\mathrm{u}$ where $\boldsymbol{W}_\mathrm{u} \in \mathbb{R}^{D_\mathrm{x} \times D_\mathrm{u}}$ and $\boldsymbol{b}_\mathrm{u} \in \mathbb{R}^{D_\mathrm{x}}$. Since $p(\boldsymbol{s}|\boldsymbol{u})$ in (43) is a smoother density, $r(\boldsymbol{x}, \boldsymbol{u})$ was also modelled by a smother function than (42) in Section 5.1.1 as follows:

$$r(\boldsymbol{x}, \boldsymbol{u}) = \sum_{i=1}^{D_\mathrm{x}} \log(\cosh(a_{i,1} h_{\mathrm{x},i}(\boldsymbol{x}) + a_{i,2} h_{\mathrm{u},i}(\boldsymbol{u}) + b_i)) - (\bar{a}_i h_{\mathrm{x},i}(\boldsymbol{x}) + \bar{b}_i)^2 + c.$$

All parameters were optimized by Adam for $1,500$ epochs with mini-batch size 256 and learning rate 0.001. The ℓ_2 regularization was also applied as done in Section 5.1.1. For the γ-cross entropy, the value of γ was increased from 0.1 to 3.0 during minibatch stochastic gradient at every 100 epoch. The performance was evaluated by the mean absolute correlation as in Section 5.1.1.

The two plots in the first row of Fig.3 are results for Donsker-Varadhan variational estimation, and clearly show that the performance for source recovery depends on the dimensionality of complementary data. When $D_\mathrm{u} < D_\mathrm{x}$, the mean correlation is small in all layers. However, when D_u is larger than or equal to D_x, the mean correlation gets significantly larger. This is consistent with implication of Theorem 3: In order to recover the source components, the dimensionality of complementary data is larger than or equal to input data (i.e., $D_\mathrm{x} \leq D_\mathrm{u}$ in Assumption (B$'$3)), which has not been revealed in previous work of nonlinear ICA. These empirical results clearly support our theoretical implications, and suggest to use fairly high-dimensional complementary data in

29
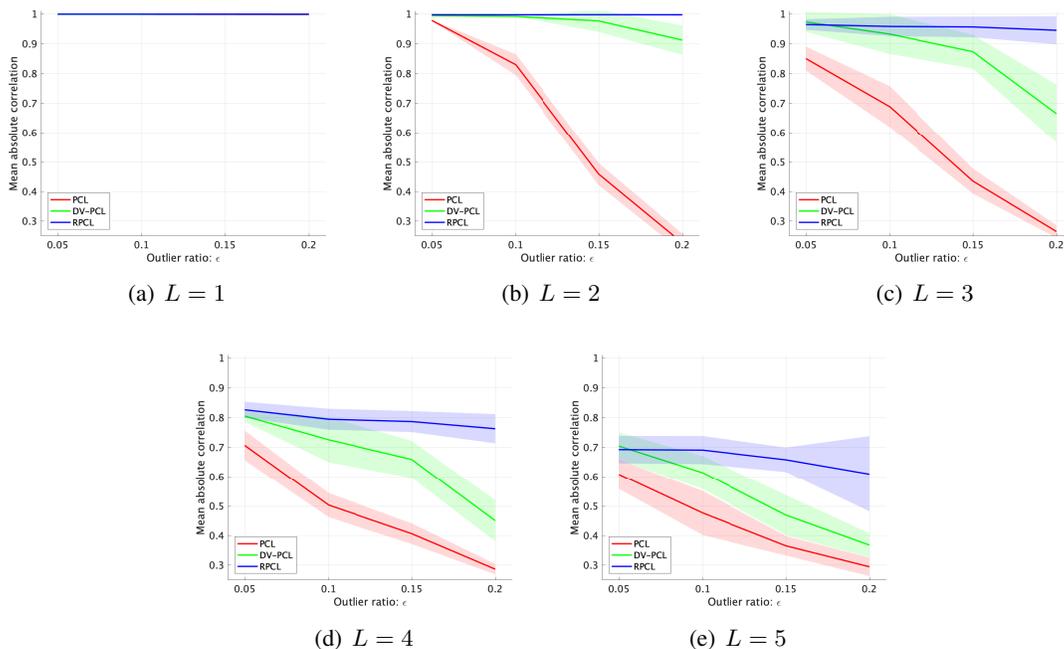
Figure 2: Outlier robustness of nonlinear ICA methods. The averages of the mean absolute corre-
lation were computed over 10 runs. L denotes the number of layers both in the nonlinear
mixing function $\boldsymbol{f}$ in (8) and representation function $\boldsymbol{h}_x$.

practice. Similar results were observed for the γ-cross entropy as well (the second row of Fig.3).
Furthermore, especially for Donsker-Varadhan variational estimation, another interesting point is
that higher-dimensional complementary data often decreases the variance of the mean absolute cor-
relation, and this implies that higher-dimensional complementary data takes a role of stabilizing
estimation as well.

5.2 Evaluation on downstream linear classification

We finally demonstrate how the proposed method based on the γ-cross entropy works on bench-
mark datasets as done in the context of maximization of mutual information (Tschannen et al.,
2019), and implicitly investigate Theorem 4 as well because the representation function $\boldsymbol{h}_x(\boldsymbol{x})$ out-
puts a lower-dimensional feature than input data. In order to evaluate methods for unsupervised
representation learning, a number of protocols have been previously proposed: Multi-scale struc-
tural similarity (Wang et al., 2003), mutual information estimation and see more protocols in Hjelm
et al. (2019). Here, we employ a linear classification protocol (van den Oord et al., 2018; Tian et al.,
2019; Tschannen et al., 2019), which consists of two steps: First, a representation function $\boldsymbol{h}_x$ is
learned with unlabelled data. Second, the learned representation function is fixed (i.e., not learned
anymore), and the feature computed through the representation function is tested on a downstream
linear classification task using labelled data. The classification accuracy is the measure for goodness
of the representation function.
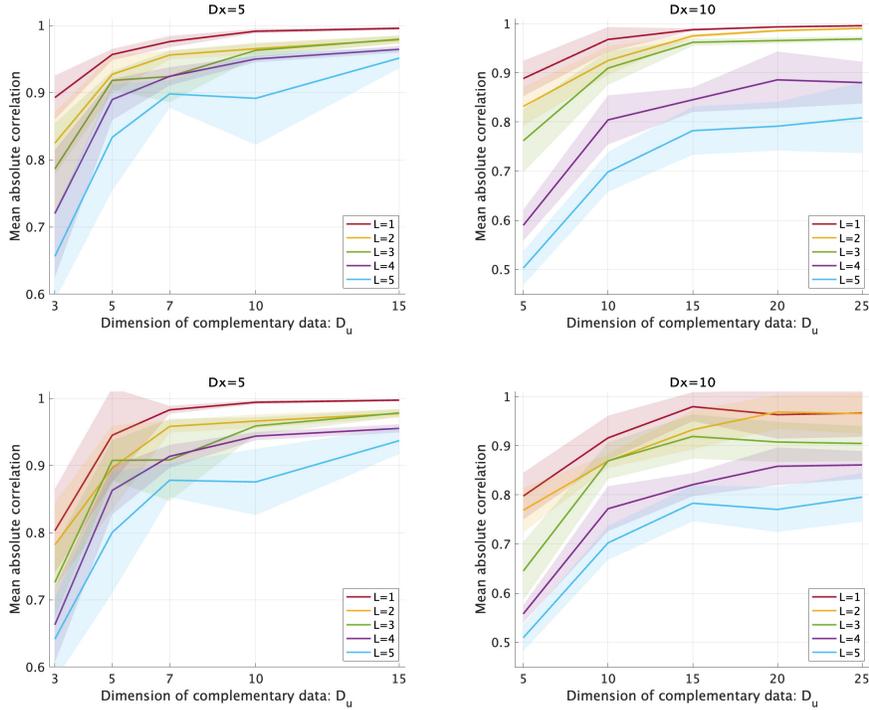
Figure 3: The averages of the mean absolute correlation against the dimensionality of complementary data over 10 runs. The left and right plots in the first (second) row are $D_x = 5$ and $D_x = 10$ for Donsker-Varadhan variational estimation (γ-cross entropy), respectively. L denotes the number of layers both in the nonlinear mixing function $\boldsymbol{f}$ in (8) and representation function $\boldsymbol{h}_x$.

More specifically, we followed the experimental protocol in Tschannen et al. (2019) [5], which has been used in the context of deep canonical correlation analysis as well (Andrew et al., 2013). We first divided a single image in half, and then used the upper and lower half images as input $\boldsymbol{x}(t)$ and complementary data samples $\boldsymbol{u}(t)$, respectively (Left figures in Fig. 4). Based on these data samples $\{\boldsymbol{x}(t), \boldsymbol{u}(t)\}_{t=1}^T$, we applied five methods based on the following objective functions for representation learning:

- *γ-cross entropy (γ-CE) $\widetilde{J}_\gamma$*: In order for initialization, we first updated the parameters in $r(\boldsymbol{x}, \boldsymbol{u})$ with $\gamma = 0.1$ for ten epochs, and then used the updated parameters as the initial parameters for the γ-cross entropy with a larger γ value.

- *Logistic regression (LR) in (5)*: LR can be seen as the limit of $\gamma \to 0$ in the γ-cross entropy.

- *Variational estimation of f-divergence (f-div) in (6) (Nguyen et al., 2008; Sugiyama et al., 2008)*.

---

5. We slightly modified the python codes available at `https://github.com/google-research/google-research/tree/master/mutual_information_representation_learning`.
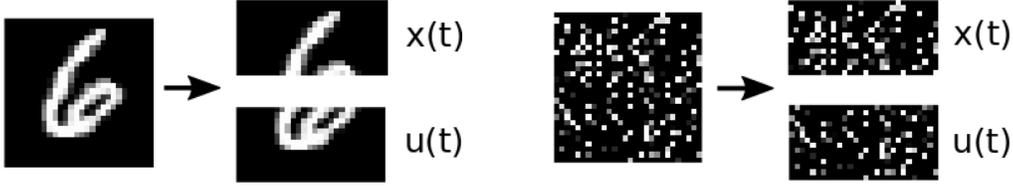
Figure 4: An example of input and complementary data. The half upper images are used for input data samples, while the half below images are complementary ones. The right images are randomly shuffled versions of left images used as outliers.
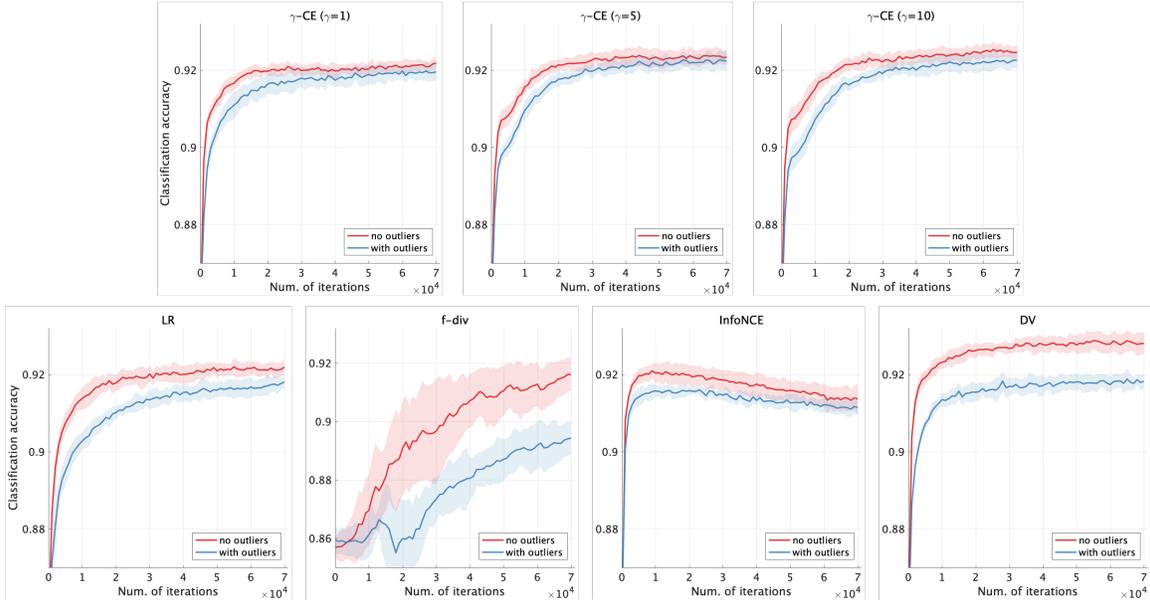
Figure 5: Averaged classification accuracy for MNIST over 10 runs. For the red lines, training is performed without outliers, while the blue lines are training with outliers. Note that the scale of the vertical axis only for f-div is different.

- *InfoNCE* (van den Oord et al., 2018):

$$\widetilde{J}_{\mathrm{NCE}}(r) := E\left[\frac{1}{K}\sum_{i=1}^{K}\log\frac{e^{r(\boldsymbol{x}_i,\boldsymbol{u}_i)}}{\frac{1}{K}\sum_{j=1}^{K}e^{r(\boldsymbol{x}_i,\boldsymbol{u}_j)}}\right], \tag{44}$$

where the expectation is taken over $\prod_{i=1}^{K}p(\boldsymbol{x}_i,\boldsymbol{u}_i)$. In practice, we estimate (44) by averaging over multiple (mini-)batches of samples (Tschannen et al., 2019).

- *Donsker-Varadhan variational estimation of mutual information (DV)* $\widetilde{J}_{\mathrm{DV}}$ (Ruderman et al., 2012; Belghazi et al., 2018): Unlike nonlinear ICA, this experiment excludes DV from the proposed methods because DV has been already used for maximization of MI (Hjelm et al., 2019).

32
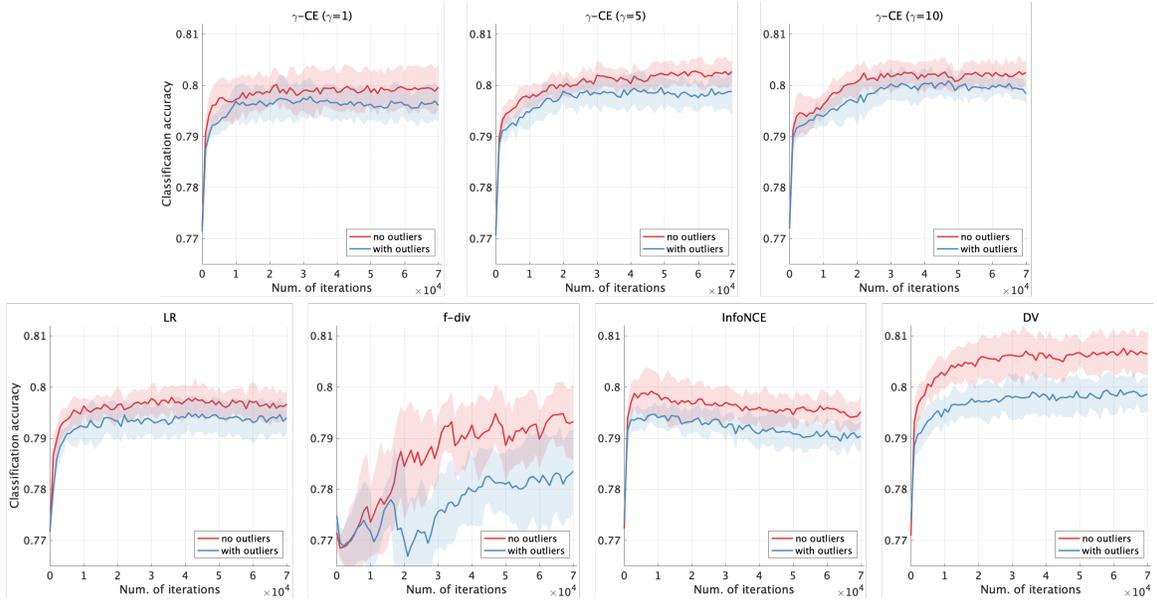
Figure 6: Averaged classification accuracy for Fashion-MNIST over 10 runs.
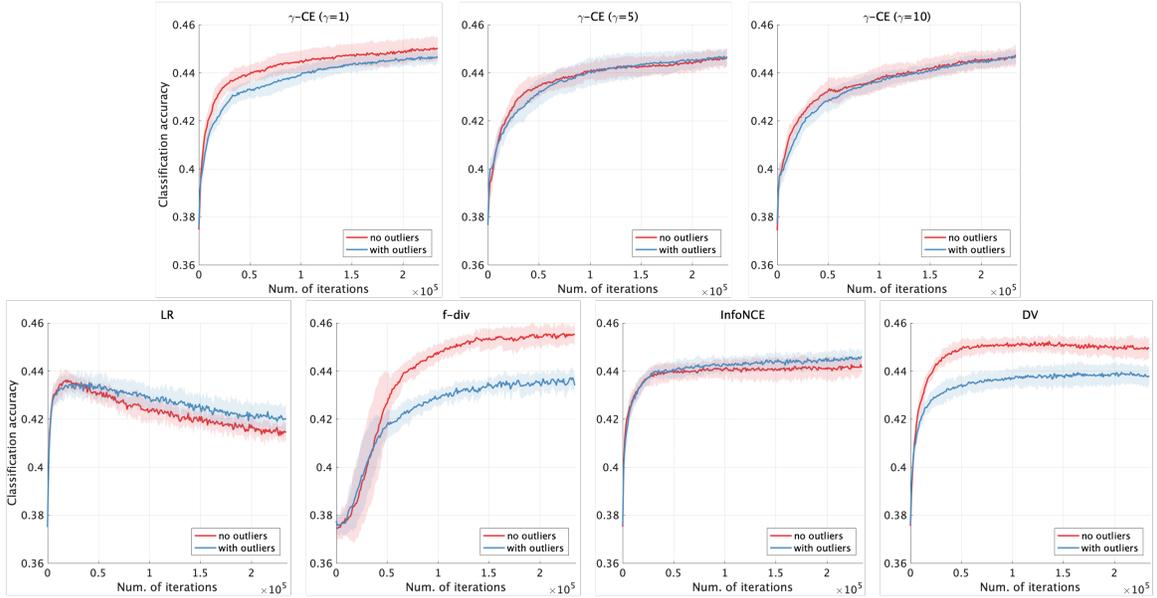
Figure 7: Averaged classification accuracy for CIFAR-10 over 10 runs.

For all methods, we used the same model $r(\boldsymbol{x}, \boldsymbol{u})$ as

$$r(\boldsymbol{x}, \boldsymbol{u}) = \psi(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}), \boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u})) + a(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x})) + b(\boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u})).$$

The representation functions of $\boldsymbol{h}_{\mathrm{x}}$ and $\boldsymbol{h}_{\mathrm{u}}$ were modelled by the same neural architecture without parameter sharing: The first two hidden layers were convolution layers with the ReLU activation function and the third layer, which is the output layer, was a fully connected layer without the activa-

33

Table 6: Averages of classification accuracy for the last $5,000$ iterations over 10 runs. Since we recorded classification accuracy at every $1,000$ iterations, each method includes the 50 points of the accuracy over 10 runs. The numbers in parentheses indicate standard deviations. The best and comparable methods judged by the t-test at the significance level 1% are described in boldface.

γ-CE ($\gamma = 5$)	γ-CE ($\gamma = 10$)	LR	f-div.	InfoNCE	DV
MNIST					
0.924(0.002)	0.925(0.002)	0.921(0.002)	0.915(0.006)	0.914(0.003)	**0.928(0.003)**
MNIST *with outliers*					
0.922(0.002)	**0.922(0.002)**	0.917(0.002)	0.893(0.006)	0.911(0.002)	0.918(0.002)
Fashion-MNIST					
0.802(0.003)	0.802(0.003)	0.796(0.003)	0.794(0.006)	0.795(0.003)	**0.807(0.004)**
Fashion-MNIST *with outliers*					
0.799(0.003)	**0.799(0.002)**	0.794(0.003)	0.783(0.008)	0.791(0.003)	**0.799(0.003)**
CIFAR-10					
0.446(0.003)	0.447(0.004)	0.414(0.004)	**0.455(0.004)**	0.442(0.004)	0.449(0.004)
CIFAR-10 *with outliers*					
0.446(0.003)	**0.446(0.003)**	0.420(0.005)	0.436(0.004)	**0.445(0.003)**	0.438(0.004)

tion function. Before the output layer, we sequentially applied layer normalization (Ba et al., 2016) and average pooling. The output dimensions were fixed at $d_{\mathrm{x}} = d_{\mathrm{u}} = 200$. $a(\boldsymbol{h}_{\mathrm{x}})$ and $b(\boldsymbol{h}_{\mathrm{u}})$ were modelled by a one-layer feedforward network without activation functions. By following van den Oord et al. (2018) and Tian et al. (2019), we set $\psi(\boldsymbol{h}_{\mathrm{x}}, \boldsymbol{h}_{\mathrm{u}}) = \boldsymbol{h}_{\mathrm{x}}^{\top} \boldsymbol{W} \boldsymbol{h}_{\mathrm{u}}$ where $\boldsymbol{W}$ is a d_{x} by d_{u} matrix and learned from data. We optimized all parameters by the Adam optimizer (Kingma and Ba, 2015). After estimating the representation function, $\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x})$ was fixed and not learned anymore on the evaluation phase. As evaluation, we only trained a linear classifier to the learned features $\{\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}(t))\}_{t=1}^{T}$ based on multinomial logistic regression. Classification accuracy on test data was used as the evaluation metric.

As datasets, we used the following three classification datasets with ten classes[6]:

- *MNIST* ($D_{\mathrm{x}} = 392, D_{\mathrm{u}} = 392$ and $T = 60,000$)

- *Fashion-MNIST* ($D_{\mathrm{x}} = 392, D_{\mathrm{u}} = 392$ and $T = 60,000$)

- *CIFAR10* ($D_{\mathrm{x}} = 1,536, D_{\mathrm{u}} = 1,536$ and $T = 50,000$)

For MNIST and Fashion-MNIST, we used the learning rate 10^{-4} and updated the parameters for 300 epochs, while the learning rate for CIFAR10 is 10^{-5} and the number of epochs is 1200. In order to demonstrate the robustness to outliers, the training data was contaminated by randomly shuffled images with respect to pixels (Right figures in Fig. 4). The contamination ratio of outliers was fixed at 0.3.

---

6. All datasets were downloaded through the tensorflow library.

Classification accuracy of MNIST on test data samples over iterations is plotted in Fig. 5. In the case of no outliers (i.e., red lines), DV performs the best, while the classification accuracy of the γ-CE and LR is fairly good. For InfoNCE, the classification accuracy around the $70,000$-th iteration is less than DV, LR and γ-CE, while f-div shows high variance of the classification accuracy over iterations. When data is contaminated by outliers (i.e., blue lines), f-div, LR and DV show clear performance degeneration. On the other hand, the performance of γ-CE is not so influenced by the contamination of outliers. Interestingly, estimation based on InfoNCE is also not strongly hampered by outliers, but the classification accuracy is worse than γ-CE around the 70000-th iteration. The same tendency of the robustness of the γ-cross entropy was observed both for Fashion-MNIST and CIFAR-10 (Figs. 6 and 7). Table 6 quantitatively indicates that the γ-cross entropy yields the best classification accuracy when data is contaminated by outliers, and shows a fairly good permanence even without outliers. Thus, the proposed method based on the γ-cross entropy can be robust against outliers as implied in our theoretical analysis and fairly works well even when data is not contaminated by outliers.

Finally, let us note that these results are not trivial because the connection between mutual information and quality of data representation has been experimentally demonstrated to be rather loose (Tian et al., 2019). Thus, it was unclear that robust estimation always leads to robust representations of data. Nonetheless, the proposed method based on the γ-cross entropy performed the best in the presence of outliers. This means that the proposed robust method for unsupervised representation learning is promising.

6. Conclusion

This paper theoretically showed that density ratio estimation plays a key role in three frameworks for unsupervised representation learning: Maximization of mutual information and nonlinear independent component analysis as well as nonlinear subspace estimation, which is a novel framework proposed in this paper. Furthermore, we made theoretical contributions in each of the three frameworks: We showed that density ratio estimation is necessary and sufficient in maximization of mutual information, while our analysis revealed a novel insight for source recover in nonlinear ICA that the dimensionality of complementary data is an important factor for source recovery, which was clearly supported by numerical experiments. In addition, the proposed generative model in nonlinear subspace estimation is more general than nonlinear ICA in the sense that the latent source components are no longer assumed to be conditionally independent, and theoretical conditions to estimate a nonlinear subspace of the latent source components were given. Motivated by the theoretical results, we developed a nonlinear ICA method by applying an variation lower-bound of mutual information, and proposed an outlier-robust method for unsupervised representation learning through density ratio estimation. We also theoretically investigated their outlier-robustness. The usefulness of the proposed methods were demonstrated through numerical experiments of nonlinear ICA and linear classification in a downstream task.

Acknowledgments

Appendix A. Proof of Theorem 1

Proof We first recall that $\boldsymbol{y}_{\mathrm{x}} := \boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x})$ and $\boldsymbol{y}_{\mathrm{u}} := \boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u})$, and denote $\boldsymbol{h}_{\mathrm{x}}^{\perp}(\boldsymbol{x})$ and $\boldsymbol{h}_{\mathrm{u}}^{\perp}(\boldsymbol{u})$ by $\boldsymbol{y}_{\mathrm{x}}^{\perp}$ and $\boldsymbol{y}_{\mathrm{u}}^{\perp}$, respectively. Based on Assumption (A2), $(\boldsymbol{y}_{\mathrm{x}}, \boldsymbol{y}_{\mathrm{x}}^{\perp}) := (\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}), \boldsymbol{h}_{\mathrm{x}}^{\perp}(\boldsymbol{x}))$ and $(\boldsymbol{y}_{\mathrm{u}}, \boldsymbol{y}_{\mathrm{u}}^{\perp}) := (\boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u}), \boldsymbol{h}_{\mathrm{u}}^{\perp}(\boldsymbol{u}))$ are both invertible in the sense that there exist functions $\boldsymbol{g}_{\mathrm{x}}$ and $\boldsymbol{g}_{\mathrm{u}}$ such that $\boldsymbol{x} = \boldsymbol{g}_{\mathrm{x}}(\boldsymbol{y}_{\mathrm{x}}, \boldsymbol{y}_{\mathrm{x}}^{\perp})$ and $\boldsymbol{u} = \boldsymbol{g}_{\mathrm{u}}(\boldsymbol{y}_{\mathrm{u}}, \boldsymbol{y}_{\mathrm{u}}^{\perp})$. Based on the invertible functions, we first establish the following lemma, which decomposes $I(\mathcal{X}, \mathcal{U})$ into three terms:

Lemma 9. *Assumptions (A1-2) hold. Then, $I(\mathcal{X}, \mathcal{U})$ can be decomposed as follows:*

$$I(\mathcal{X}, \mathcal{U}) = I(\mathcal{Y}_{\mathrm{x}}, \mathcal{Y}_{\mathrm{u}}) + E_{Y_{\mathrm{u}}}[I(\mathcal{Y}_{\mathrm{x}}|\mathcal{Y}_{\mathrm{u}}, \mathcal{Y}_{\mathrm{u}}^{\perp}|\mathcal{Y}_{\mathrm{u}})] + E_{Y_{\mathrm{x}}}[I(\mathcal{Y}_{\mathrm{x}}^{\perp}|\mathcal{Y}_{\mathrm{x}}, \mathcal{U}|\mathcal{Y}_{\mathrm{x}})], \tag{45}$$

where $E_{Y_{\mathrm{u}}}$ and $E_{Y_{\mathrm{x}}}$ denote the expectations over $p(\boldsymbol{y}_{\mathrm{u}})$ and $p(\boldsymbol{y}_{\mathrm{x}})$ respectively,

$$I(\mathcal{Y}_{\mathrm{x}}|\mathcal{Y}_{\mathrm{u}}, \mathcal{Y}_{\mathrm{u}}^{\perp}|\mathcal{Y}_{\mathrm{u}}) = \iint p(\boldsymbol{y}_{\mathrm{x}}, \boldsymbol{y}_{\mathrm{u}}^{\perp}|\boldsymbol{y}_{\mathrm{u}}) \log \frac{p(\boldsymbol{y}_{\mathrm{x}}, \boldsymbol{y}_{\mathrm{u}}^{\perp}|\boldsymbol{y}_{\mathrm{u}})}{p(\boldsymbol{y}_{\mathrm{x}}|\boldsymbol{y}_{\mathrm{u}})p(\boldsymbol{y}_{\mathrm{u}}^{\perp}|\boldsymbol{y}_{\mathrm{u}})} \mathrm{d}\boldsymbol{y}_{\mathrm{x}} \mathrm{d}\boldsymbol{y}_{\mathrm{u}}^{\perp}$$

$$I(\mathcal{Y}_{\mathrm{x}}^{\perp}|\mathcal{Y}_{\mathrm{x}}, \mathcal{U}|\mathcal{Y}_{\mathrm{x}}) := \iint p(\boldsymbol{y}_{\mathrm{x}}^{\perp}, \boldsymbol{u}|\boldsymbol{y}_{\mathrm{x}}) \log \frac{p(\boldsymbol{y}_{\mathrm{x}}^{\perp}, \boldsymbol{u}|\boldsymbol{y}_{\mathrm{x}})}{p(\boldsymbol{y}_{\mathrm{x}}^{\perp}|\boldsymbol{y}_{\mathrm{x}})p(\boldsymbol{u}|\boldsymbol{y}_{\mathrm{x}})} \mathrm{d}\boldsymbol{y}_{\mathrm{x}}^{\perp} \mathrm{d}\boldsymbol{u}.$$

The proof is given in Appendix A.1. The first term on the right-hand side of (45) is mutual information between the representation functions of $\boldsymbol{x}$ and $\boldsymbol{u}$, while $I(\mathcal{Y}_{\mathrm{x}}^{\perp}|\mathcal{Y}_{\mathrm{x}}, \mathcal{U}|\mathcal{Y}_{\mathrm{x}})$ and $I(\mathcal{Y}_{\mathrm{x}}|\mathcal{Y}_{\mathrm{u}}, \mathcal{Y}_{\mathrm{u}}^{\perp}|\mathcal{Y}_{\mathrm{u}})$ measure the conditional independence and are equal to 0 when $\boldsymbol{u} \perp \boldsymbol{y}_{\mathrm{x}}^{\perp}|\boldsymbol{y}_{\mathrm{x}}$ and $\boldsymbol{y}_{\mathrm{x}} \perp \boldsymbol{y}_{\mathrm{u}}^{\perp}|\boldsymbol{y}_{\mathrm{u}}$. It follows from (45) that $I(\mathcal{X}, \mathcal{U}) = I(\mathcal{Y}_{\mathrm{x}}, \mathcal{Y}_{\mathrm{u}})$ if and only if $I(\mathcal{Y}_{\mathrm{x}}|\mathcal{Y}_{\mathrm{u}}, \mathcal{Y}_{\mathrm{u}}^{\perp}|\mathcal{Y}_{\mathrm{u}}) = 0$ and $I(\mathcal{Y}_{\mathrm{x}}^{\perp}|\mathcal{Y}_{\mathrm{x}}, \mathcal{U}|\mathcal{Y}_{\mathrm{x}}) = 0$.

Next, we fix $\boldsymbol{h}_{\mathrm{x}}$ and $\boldsymbol{h}_{\mathrm{u}}$ at $\boldsymbol{h}_{\mathrm{x}}^{\star}$ and $\boldsymbol{h}_{\mathrm{u}}^{\star}$ respectively, and substitute them as $\boldsymbol{y}_{\mathrm{x}} = \boldsymbol{h}_{\mathrm{x}}^{\star}(\boldsymbol{x})$ and $\boldsymbol{y}_{\mathrm{u}} = \boldsymbol{h}_{\mathrm{u}}^{\star}(\boldsymbol{u})$. Then, it is shown that (10) implies both $I(\mathcal{Y}_{\mathrm{x}}|\mathcal{Y}_{\mathrm{u}}, \mathcal{Y}_{\mathrm{u}}^{\perp}|\mathcal{Y}_{\mathrm{u}}) = 0$ and $I(\mathcal{Y}_{\mathrm{x}}^{\perp}|\mathcal{Y}_{\mathrm{x}}, \mathcal{U}|\mathcal{Y}_{\mathrm{x}}) = 0$. We first multiply $p(\boldsymbol{u})$ to both sides and rewrite (10) as

$$p(\boldsymbol{u}|\boldsymbol{x}) = p(\boldsymbol{u}|\boldsymbol{y}_{\mathrm{x}}), \tag{46}$$

where with the partition function $Z(\boldsymbol{y}_{\mathrm{x}}) := \int \exp\left(\psi^{\star}(\boldsymbol{y}_{\mathrm{x}}, \boldsymbol{h}_{\mathrm{u}}^{\star}(\boldsymbol{u})) + a^{\star}(\boldsymbol{y}_{\mathrm{x}}) + b^{\star}(\boldsymbol{h}_{\mathrm{u}}^{\star}(\boldsymbol{u}))\right) p(\boldsymbol{u}) \mathrm{d}\boldsymbol{u}$,

$$p(\boldsymbol{u}|\boldsymbol{y}_{\mathrm{x}}) := \frac{\exp\left(\psi^{\star}(\boldsymbol{y}_{\mathrm{x}}, \boldsymbol{h}_{\mathrm{u}}^{\star}(\boldsymbol{u}))\right) + a^{\star}(\boldsymbol{y}_{\mathrm{x}}) + b^{\star}(\boldsymbol{h}_{\mathrm{u}}^{\star}(\boldsymbol{u}))) p(\boldsymbol{u})}{Z(\boldsymbol{y}_{\mathrm{x}})}.$$

By the invertible assumption for $(\boldsymbol{y}_{\mathrm{x}}, \boldsymbol{y}_{\mathrm{x}}^{\perp})$, (46) can be equivalently expressed as the following conditional independence:

$$\boldsymbol{u} \perp \boldsymbol{y}_{\mathrm{x}}^{\perp}|\boldsymbol{y}_{\mathrm{x}}, \tag{47}$$

implying that

$$I(\mathcal{Y}_{\mathrm{x}}^{\perp}|\mathcal{Y}_{\mathrm{x}}, \mathcal{U}|\mathcal{Y}_{\mathrm{x}}) = 0.$$

Similarly, we can derive the following equation from (10):

$$p(\boldsymbol{x}|\boldsymbol{u}) = p(\boldsymbol{x}|\boldsymbol{y}_{\mathrm{u}}), \tag{48}$$

where with the partition function $Z(\boldsymbol{y}_{\mathrm{u}}) := \int \exp\left(\psi^{\star}(\boldsymbol{h}_{\mathrm{x}}^{\star}(\boldsymbol{x}), \boldsymbol{y}_{\mathrm{u}}) + a^{\star}(\boldsymbol{h}_{\mathrm{x}}^{\star}(\boldsymbol{x})) + b^{\star}(\boldsymbol{y}_{\mathrm{u}})\right) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$,

$$p(\boldsymbol{x}|\boldsymbol{y}_{\mathrm{u}}) := \frac{\exp\left(\psi^{\star}(\boldsymbol{h}_{\mathrm{x}}^{\star}(\boldsymbol{x}), \boldsymbol{y}_{\mathrm{u}}) + a^{\star}(\boldsymbol{h}_{\mathrm{x}}^{\star}(\boldsymbol{x})) + b^{\star}(\boldsymbol{y}_{\mathrm{u}})\right) p(\boldsymbol{x})}{Z(\boldsymbol{y}_{\mathrm{u}})}.$$

Again, (48) implies the conditional independence as

$$\boldsymbol{x} \perp \boldsymbol{y}_{\mathrm{u}}^{\perp}|\boldsymbol{y}_{\mathrm{u}}. \tag{49}$$

Data processing inequality and conditional independence (49) lead to

$$I(\mathcal{Y}_{\mathrm{x}}|\mathcal{Y}_{\mathrm{u}}, \mathcal{Y}_{\mathrm{u}}^{\perp}|\mathcal{Y}_{\mathrm{u}}) \le I(\mathcal{X}|\mathcal{Y}_{\mathrm{u}}, \mathcal{Y}_{\mathrm{u}}^{\perp}|\mathcal{Y}_{\mathrm{u}}) \underset{(49)}{=} 0.$$

Both $I(\mathcal{Y}_{\mathrm{x}}|\mathcal{Y}_{\mathrm{u}}, \mathcal{Y}_{\mathrm{u}}^{\perp}|\mathcal{Y}_{\mathrm{u}}) = 0$ and $I(\mathcal{Y}_{\mathrm{x}}^{\perp}|\mathcal{Y}_{\mathrm{x}}, \mathcal{U}|\mathcal{Y}_{\mathrm{x}}) = 0$ enable us to ensure $I(\mathcal{X}, \mathcal{U}) = I(\mathcal{Y}_{\mathrm{x}}, \mathcal{Y}_{\mathrm{u}})$.

Conversely, we suppose that $I(\mathcal{X}, \mathcal{U}) = I(\mathcal{Y}_{\mathrm{x}}, \mathcal{Y}_{\mathrm{u}})$ at $\boldsymbol{h}_{\mathrm{x}} = \boldsymbol{h}_{\mathrm{x}}^{\star}$ and $\boldsymbol{h}_{\mathrm{u}} = \boldsymbol{h}_{\mathrm{u}}^{\star}$. Then, it follows from (45) that $I(\mathcal{Y}_{\mathrm{x}}|\mathcal{Y}_{\mathrm{u}}, \mathcal{Y}_{\mathrm{u}}^{\perp}|\mathcal{Y}_{\mathrm{u}}) = 0$ and $I(\mathcal{Y}_{\mathrm{x}}^{\perp}|\mathcal{Y}_{\mathrm{x}}, \mathcal{U}|\mathcal{Y}_{\mathrm{x}}) = 0$. $I(\mathcal{Y}_{\mathrm{x}}^{\perp}|\mathcal{Y}_{\mathrm{x}}, \mathcal{U}|\mathcal{Y}_{\mathrm{x}}) = 0$ means the conditional independence (47), which can be equivalently expressed as (46). On the other hand, under the invertibility of $(\boldsymbol{y}_{\mathrm{u}}, \boldsymbol{y}_{\mathrm{u}}^{\perp})$, $I(\mathcal{Y}_{\mathrm{x}}|\mathcal{Y}_{\mathrm{u}}, \mathcal{Y}_{\mathrm{u}}^{\perp}|\mathcal{Y}_{\mathrm{u}}) = 0$ implies

$$\boldsymbol{y}_{\mathrm{x}} \perp \boldsymbol{y}_{\mathrm{u}}^{\perp}|\boldsymbol{y}_{\mathrm{u}} \quad \text{or equivalently} \quad p(\boldsymbol{y}_{\mathrm{x}}|\boldsymbol{u}) = p(\boldsymbol{y}_{\mathrm{x}}|\boldsymbol{y}_{\mathrm{u}}). \tag{50}$$

A simple calculation yields

$$p(\boldsymbol{x}, \boldsymbol{u}) = p(\boldsymbol{u}|\boldsymbol{x})p(\boldsymbol{x}) \underset{(46)}{=} p(\boldsymbol{u}|\boldsymbol{y}_{\mathrm{x}})p(\boldsymbol{x}) \underset{(\star)}{=} \frac{p(\boldsymbol{y}_{\mathrm{x}}|\boldsymbol{u})p(\boldsymbol{u})}{p(\boldsymbol{y}_{\mathrm{x}})}p(\boldsymbol{x}) \underset{(50)}{=} \frac{p(\boldsymbol{y}_{\mathrm{x}}|\boldsymbol{y}_{\mathrm{u}})}{p(\boldsymbol{y}_{\mathrm{x}})}p(\boldsymbol{x})p(\boldsymbol{u}), \tag{51}$$

where we applied Bayes' theorem on $(\star)$. By dividing both sides of (51) by $p(\boldsymbol{x})p(\boldsymbol{u})$, we have

$$\log \frac{p(\boldsymbol{x}, \boldsymbol{u})}{p(\boldsymbol{x})p(\boldsymbol{u})} = \log p(\boldsymbol{y}_{\mathrm{x}}, \boldsymbol{y}_{\mathrm{u}}) - \log p(\boldsymbol{y}_{\mathrm{x}}) - \log p(\boldsymbol{y}_{\mathrm{u}}).$$

The existence of the functions is obvious by denoting $\log p(\boldsymbol{y}_{\mathrm{x}}, \boldsymbol{y}_{\mathrm{u}})$, $-\log p(\boldsymbol{y}_{\mathrm{x}})$ and $-\log p(\boldsymbol{y}_{\mathrm{u}})$ by $\psi^{\star}(\boldsymbol{y}_{\mathrm{x}}, \boldsymbol{y}_{\mathrm{u}})$, $a^{\star}(\boldsymbol{y}_{\mathrm{x}})$ and $b^{\star}(\boldsymbol{y}_{\mathrm{u}})$ respectively. The proof is completed. ∎

A.1 Proof of Lemma 9

Proof We first decompose $I(\mathcal{X}, \mathcal{U})$ based on the invertibility assumption of $(\boldsymbol{y}_{\mathrm{x}}, \boldsymbol{y}_{\mathrm{x}}^{\perp})$ as follows:

$$
\begin{aligned}
I(\mathcal{X}, \mathcal{U}) &= \iiint p(\boldsymbol{y}_{\mathrm{x}}, \boldsymbol{y}_{\mathrm{x}}^{\perp}, \boldsymbol{u}) \log \frac{p(\boldsymbol{y}_{\mathrm{x}}, \boldsymbol{y}_{\mathrm{x}}^{\perp}, \boldsymbol{u})}{p(\boldsymbol{y}_{\mathrm{x}}, \boldsymbol{y}_{\mathrm{x}}^{\perp})p(\boldsymbol{u})} \mathrm{d}\boldsymbol{y}_{\mathrm{x}}\mathrm{d}\boldsymbol{y}_{\mathrm{x}}^{\perp}\mathrm{d}\boldsymbol{u} \\
&= \iint p(\boldsymbol{y}_{\mathrm{x}}, \boldsymbol{u}) \log \frac{p(\boldsymbol{y}_{\mathrm{x}}, \boldsymbol{u})}{p(\boldsymbol{y}_{\mathrm{x}})p(\boldsymbol{u})} \mathrm{d}\boldsymbol{y}_{\mathrm{x}}\mathrm{d}\boldsymbol{u} \\
&\quad + \int p(\boldsymbol{y}_{\mathrm{x}}) \left[\iint p(\boldsymbol{y}_{\mathrm{x}}^{\perp}, \boldsymbol{u}|\boldsymbol{y}_{\mathrm{x}}) \log \frac{p(\boldsymbol{y}_{\mathrm{x}}^{\perp}, \boldsymbol{u}|\boldsymbol{y}_{\mathrm{x}})}{p(\boldsymbol{y}_{\mathrm{x}}^{\perp}|\boldsymbol{y}_{\mathrm{x}})p(\boldsymbol{u}|\boldsymbol{y}_{\mathrm{x}})} \mathrm{d}\boldsymbol{u}\mathrm{d}\boldsymbol{y}_{\mathrm{x}}^{\perp} \right] \mathrm{d}\boldsymbol{y}_{\mathrm{x}} \\
&= I(\mathcal{Y}_{\mathrm{x}}, \mathcal{U}) + E_{Y_{\mathrm{x}}}[I(\mathcal{Y}_{\mathrm{x}}^{\perp}|\mathcal{Y}_{\mathrm{x}}, \mathcal{U}|\mathcal{Y}_{\mathrm{x}})], \tag{52}
\end{aligned}
$$

where we used the following relation:

$$\frac{p(\boldsymbol{y}_{\mathrm{x}}, \boldsymbol{y}_{\mathrm{x}}^{\perp}, \boldsymbol{u})}{p(\boldsymbol{y}_{\mathrm{x}}, \boldsymbol{y}_{\mathrm{x}}^{\perp})p(\boldsymbol{u})} = \frac{p(\boldsymbol{y}_{\mathrm{x}}^{\perp}, \boldsymbol{u}|\boldsymbol{y}_{\mathrm{x}})p(\boldsymbol{y}_{\mathrm{x}})}{p(\boldsymbol{y}_{\mathrm{x}}^{\perp}|\boldsymbol{y}_{\mathrm{x}})p(\boldsymbol{y}_{\mathrm{x}})p(\boldsymbol{u})} \cdot \frac{p(\boldsymbol{y}_{\mathrm{x}}, \boldsymbol{u})}{p(\boldsymbol{y}_{\mathrm{x}}, \boldsymbol{u})} = \frac{p(\boldsymbol{y}_{\mathrm{x}}, \boldsymbol{u})}{p(\boldsymbol{y}_{\mathrm{x}})p(\boldsymbol{u})} \cdot \frac{p(\boldsymbol{y}_{\mathrm{x}}^{\perp}, \boldsymbol{u}|\boldsymbol{y}_{\mathrm{x}})}{p(\boldsymbol{y}_{\mathrm{x}}^{\perp}|\boldsymbol{y}_{\mathrm{x}})p(\boldsymbol{u}|\boldsymbol{y}_{\mathrm{x}})}.$$

Applying the same decomposition as (52) to $I(\mathcal{Y}_{\mathrm{x}}, \mathcal{U})$ under the change of variables by the invertible $(\boldsymbol{y}_{\mathrm{u}}, \boldsymbol{y}_{\mathrm{u}}^{\perp})$ yields

$$\begin{aligned}
I(\mathcal{Y}_{\mathrm{x}}, \mathcal{U}) &= \iiint p(\boldsymbol{y}_{\mathrm{x}}, \boldsymbol{y}_{\mathrm{u}}, \boldsymbol{y}_{\mathrm{u}}^{\perp}) \log \frac{p(\boldsymbol{y}_{\mathrm{x}}, \boldsymbol{y}_{\mathrm{u}}, \boldsymbol{y}_{\mathrm{u}}^{\perp})}{p(\boldsymbol{y}_{\mathrm{x}})p(\boldsymbol{y}_{\mathrm{u}}, \boldsymbol{y}_{\mathrm{u}}^{\perp})} \mathrm{d}\boldsymbol{y}_{\mathrm{x}}\mathrm{d}\boldsymbol{y}_{\mathrm{u}}\mathrm{d}\boldsymbol{y}_{\mathrm{u}}^{\perp} \\
&= \iint p(\boldsymbol{y}_{\mathrm{x}}, \boldsymbol{y}_{\mathrm{u}}) \log \frac{p(\boldsymbol{y}_{\mathrm{x}}, \boldsymbol{y}_{\mathrm{u}})}{p(\boldsymbol{y}_{\mathrm{x}})p(\boldsymbol{y}_{\mathrm{u}})} \mathrm{d}\boldsymbol{y}_{\mathrm{x}}\mathrm{d}\boldsymbol{y}_{\mathrm{u}} \\
&\quad + \int p(\boldsymbol{y}_{\mathrm{u}}) \left[\iint p(\boldsymbol{y}_{\mathrm{x}}, \boldsymbol{y}_{\mathrm{u}}^{\perp}|\boldsymbol{y}_{\mathrm{u}}) \log \frac{p(\boldsymbol{y}_{\mathrm{x}}, \boldsymbol{y}_{\mathrm{u}}^{\perp}|\boldsymbol{y}_{\mathrm{u}})}{p(\boldsymbol{y}_{\mathrm{u}}^{\perp}|\boldsymbol{y}_{\mathrm{u}})p(\boldsymbol{y}_{\mathrm{x}}|\boldsymbol{y}_{\mathrm{u}})} \mathrm{d}\boldsymbol{y}_{\mathrm{x}}\mathrm{d}\boldsymbol{y}_{\mathrm{u}}^{\perp} \right] \mathrm{d}\boldsymbol{y}_{\mathrm{u}} \\
&= I(\mathcal{Y}_{\mathrm{x}}, \mathcal{Y}_{\mathrm{u}}) + E_{\mathrm{Y}_{\mathrm{u}}}[I(\mathcal{Y}_{\mathrm{x}}|\mathcal{Y}_{\mathrm{u}}, \mathcal{Y}_{\mathrm{u}}^{\perp}|\mathcal{Y}_{\mathrm{u}})],
\end{aligned} \tag{53}$$

where we used

$$\frac{p(\boldsymbol{y}_{\mathrm{x}}, \boldsymbol{y}_{\mathrm{u}}, \boldsymbol{y}_{\mathrm{u}}^{\perp})}{p(\boldsymbol{y}_{\mathrm{x}})p(\boldsymbol{y}_{\mathrm{u}}, \boldsymbol{y}_{\mathrm{u}}^{\perp})} = \frac{p(\boldsymbol{y}_{\mathrm{x}}, \boldsymbol{y}_{\mathrm{u}}^{\perp}|\boldsymbol{y}_{\mathrm{u}})p(\boldsymbol{y}_{\mathrm{u}})}{p(\boldsymbol{y}_{\mathrm{x}})p(\boldsymbol{y}_{\mathrm{u}}^{\perp}|\boldsymbol{y}_{\mathrm{u}})p(\boldsymbol{y}_{\mathrm{u}})} \cdot \frac{p(\boldsymbol{y}_{\mathrm{x}}, \boldsymbol{y}_{\mathrm{u}})}{p(\boldsymbol{y}_{\mathrm{x}}, \boldsymbol{y}_{\mathrm{u}})} = \frac{p(\boldsymbol{y}_{\mathrm{x}}, \boldsymbol{y}_{\mathrm{u}})}{p(\boldsymbol{y}_{\mathrm{x}})p(\boldsymbol{y}_{\mathrm{u}})} \cdot \frac{p(\boldsymbol{y}_{\mathrm{x}}, \boldsymbol{y}_{\mathrm{u}}^{\perp}|\boldsymbol{y}_{\mathrm{u}})}{p(\boldsymbol{y}_{\mathrm{u}}^{\perp}|\boldsymbol{y}_{\mathrm{u}})p(\boldsymbol{y}_{\mathrm{x}}|\boldsymbol{y}_{\mathrm{u}})}$$

Substituting (53) into (52) completes the proof. ∎

Appendix B. Proof of Proposition 2

Proof Let us denote the inverse of the mixing function $\boldsymbol{f}$ in the generative model (8) as $\boldsymbol{g} := \boldsymbol{f}^{-1}$. From the conditional independence of $\boldsymbol{s}$ given $\boldsymbol{u}$ in Assumption (B1), we obtain the conditional density of $\boldsymbol{x}$ given $\boldsymbol{u}$ under the change of variables by $\boldsymbol{s} = \boldsymbol{g}(\boldsymbol{x})$ as follows:

$$\log p(\boldsymbol{x}|\boldsymbol{u}) = \sum_{i=1}^{D_{\mathrm{x}}} q_i(g_i(\boldsymbol{x}), \boldsymbol{u}) + \log|\det \boldsymbol{J}_{\boldsymbol{g}}(\boldsymbol{x})| - \log Z(\boldsymbol{u}), \tag{54}$$

where $\boldsymbol{g}(\boldsymbol{x}) = (g_1(\boldsymbol{x}), g_2(\boldsymbol{x}), \dots, g_{D_{\mathrm{x}}}(\boldsymbol{x}))^{\top}$ and $\boldsymbol{J}_{\boldsymbol{g}}(\boldsymbol{x})$ is the Jacobian of $\boldsymbol{g}$ at $\boldsymbol{x}$. Similarly, by applying the same change of variables for the marginal density of $\boldsymbol{s}$, the marginal density of $\boldsymbol{x}$ is given by

$$\log p(\boldsymbol{x}) = \log p(\boldsymbol{g}(\boldsymbol{x})) + \log|\det \boldsymbol{J}_{\boldsymbol{g}}(\boldsymbol{x})|. \tag{55}$$

On the other hand, the log-density ratio equation (11) yields

$$\log p(\boldsymbol{x}|\boldsymbol{u}) - \log p(\boldsymbol{x}) = \sum_{i=1}^{D_{\mathrm{x}}} \psi_i^{\star}(h_{\mathrm{x},i}^{\star}(\boldsymbol{x}), h_{\mathrm{u}}^{\star}(\boldsymbol{u})) + a^{\star}(h_{\mathrm{x}}^{\star}(\boldsymbol{x})) + b^{\star}(h_{\mathrm{u}}^{\star}(\boldsymbol{u})). \tag{56}$$

Substitution of (54) and (55) into the left-hand side of (56) cancels out the Jacobian term and gives the following equation:

$$\sum_{i=1}^{D_{\mathrm{x}}} q_i(g_i(\boldsymbol{x}), \boldsymbol{u}) - \log Z(\boldsymbol{u}) - \log p(\boldsymbol{g}(\boldsymbol{x})) = \sum_{i=1}^{D_{\mathrm{x}}} \psi_i^{\star}(h_{\mathrm{x},i}^{\star}(\boldsymbol{x}), h_{\mathrm{u}}^{\star}(\boldsymbol{u})) + a^{\star}(h_{\mathrm{x}}^{\star}(\boldsymbol{x})) + b^{\star}(h_{\mathrm{u}}^{\star}(\boldsymbol{u})).$$

$$\tag{57}$$

Then, we compute the gradient of the both sides on (57) with respect to $\boldsymbol{u}$ as

$$\sum_{i=1}^{D_{\mathrm{x}}} \nabla_{\boldsymbol{u}} q_i(g_i(\boldsymbol{x}), \boldsymbol{u}) - \nabla_{\boldsymbol{u}} \log Z(\boldsymbol{u}) = \sum_{i=1}^{D_{\mathrm{x}}} \nabla_{\boldsymbol{u}} \psi_i^\star(h_{\mathrm{x},i}^\star(\boldsymbol{x}), \boldsymbol{h}_{\mathrm{u}}^\star(\boldsymbol{u})) + \nabla_{\boldsymbol{u}} b^\star(\boldsymbol{h}_{\mathrm{u}}^\star(\boldsymbol{u})).$$

By using the notations of $\boldsymbol{z} := \boldsymbol{h}_{\mathrm{x}}^\star(\boldsymbol{x})$ and $\boldsymbol{v}(\boldsymbol{z}) := \boldsymbol{g}((\boldsymbol{h}_{\mathrm{x}}^\star)^{-1}(\boldsymbol{z}))$ where $(\boldsymbol{h}_{\mathrm{x}}^\star)^{-1}$ denotes the inverse of $\boldsymbol{h}_{\mathrm{x}}^\star$,

$$\sum_{i=1}^{D_{\mathrm{x}}} \nabla_{\boldsymbol{u}} q_i(v_i(\boldsymbol{z}), \boldsymbol{u}) - \nabla_{\boldsymbol{u}} \log Z(\boldsymbol{u}) = \sum_{i=1}^{D_{\mathrm{x}}} \nabla_{\boldsymbol{u}} \psi_i^\star(z_i, \boldsymbol{h}_{\mathrm{u}}^\star(\boldsymbol{u})) + \nabla_{\boldsymbol{u}} b^\star(\boldsymbol{h}_{\mathrm{u}}^\star(\boldsymbol{u})), \tag{58}$$

where z_i and v_i are the i-th elements in $\boldsymbol{z}$ and $\boldsymbol{v}$, respectively.

Next, we show from (58) that each element in $\boldsymbol{v}(\boldsymbol{z})(=\boldsymbol{s})$ is a function of a distinct and single element in $\boldsymbol{z}(=\boldsymbol{h}_{\mathrm{x}}^\star(\boldsymbol{x}))$. To this end, we first take the partial derivative to both sides on (58) with respect to z_l as

$$\sum_{i=1}^{D_{\mathrm{x}}} v_i^{(l)}(\boldsymbol{z}) \nabla_{\boldsymbol{u}} q_i'(v_i(\boldsymbol{z}), \boldsymbol{u}) = \frac{\partial}{\partial z_l} \left[\nabla_{\boldsymbol{u}} \psi_l^\star(z_l, \boldsymbol{h}_{\mathrm{u}}^\star(\boldsymbol{u})) \right], \tag{59}$$

where $q_i'(t, \boldsymbol{u}) := \frac{\partial q_i(t, \boldsymbol{u})}{\partial t}$ and $v_i^{(l)}(\boldsymbol{z}) := \frac{\partial v_i(\boldsymbol{z})}{\partial z_l}$. We further take the partial derivative of both sides on (59) with respect to z_m for $m \neq l$ as

$$\sum_{i=1}^{D_{\mathrm{x}}} \left[v_i^{(l,m)}(\boldsymbol{z}) \nabla_{\boldsymbol{u}} q_i'(v_i(\boldsymbol{z}), \boldsymbol{u}) + v_i^{(l)}(\boldsymbol{z}) v_i^{(m)}(\boldsymbol{z}) \nabla_{\boldsymbol{u}} q_i''(v_i(\boldsymbol{z}), \boldsymbol{u}) \right] = \boldsymbol{0}, \tag{60}$$

where $q_i''(t, \boldsymbol{u}) := \frac{\partial^2}{\partial t^2} q_i(t, \boldsymbol{u})$ and $v_i^{(l,m)}(\boldsymbol{z}) := \frac{\partial^2}{\partial z_l \partial z_m} v_i(\boldsymbol{z})$. In order to express (60) as a matrix form, we define the $D_{\mathrm{x}}(D_{\mathrm{x}} - 1)/2$ by $2D_{\mathrm{x}}$ matrix consisting of a collection of $v_i^{(l)}(\boldsymbol{z}) v_i^{(m)}(\boldsymbol{z})$ and $v_i^{(l,m)}(\boldsymbol{z})$ with respect to $i, l, m \ (l \neq m)$ as

$$\boldsymbol{M}(\boldsymbol{v}) := \begin{pmatrix} v_1^{(1,2)} & \cdots & v_{D_{\mathrm{x}}}^{(1,2)} & v_1^{(1)} v_1^{(2)} & \cdots & v_{D_{\mathrm{x}}}^{(1)} v_{D_{\mathrm{x}}}^{(2)} \\ v_1^{(1,3)} & \cdots & v_{D_{\mathrm{x}}}^{(1,3)} & v_1^{(1)} v_1^{(3)} & \cdots & v_{D_{\mathrm{x}}}^{(1)} v_{D_{\mathrm{x}}}^{(3)} \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ v_1^{(D_{\mathrm{x}}-1,D_{\mathrm{x}})} & \cdots & v_{D_{\mathrm{x}}}^{(D_{\mathrm{x}}-1,D_{\mathrm{x}})} & v_1^{(D_{\mathrm{x}}-1)} v_1^{(D_{\mathrm{x}})} & \cdots & v_{D_{\mathrm{x}}}^{(D_{\mathrm{x}}-1)} v_{D_{\mathrm{x}}}^{(D_{\mathrm{x}})} \end{pmatrix} \in \mathbb{R}^{\frac{D_{\mathrm{x}}(D_{\mathrm{x}}-1)}{2} \times 2D_{\mathrm{x}}}.$$

We also define a collection of $q_i'(v_i(\boldsymbol{z}), \boldsymbol{u})$ and $q_i''(v_i(\boldsymbol{z}), \boldsymbol{u})$ by the following $2D_{\mathrm{x}}$-dimensional vector:

$$\boldsymbol{w}(\boldsymbol{v}, \boldsymbol{u}) := \left(q_1'(v_1, \boldsymbol{u}), \ldots, q_{D_{\mathrm{x}}}'(v_{D_{\mathrm{x}}}, \boldsymbol{u}), q_1''(v_1, \boldsymbol{u}), \ldots, q_{D_{\mathrm{x}}}''(v_{D_{\mathrm{x}}}, \boldsymbol{u}) \right)^\top \in \mathbb{R}^{2D_{\mathrm{x}}}.$$

By fixing $\boldsymbol{u}$ at $\boldsymbol{u} = \boldsymbol{u}_1$, (60) can be expressed by using $\boldsymbol{M}(\boldsymbol{v})$ and $\boldsymbol{w}(\boldsymbol{v}, \boldsymbol{u})$ as

$$\boldsymbol{M}(\boldsymbol{v}) \boldsymbol{W}(\boldsymbol{v}) = \boldsymbol{O}, \tag{61}$$

where $\boldsymbol{W}(\boldsymbol{v}) := \nabla_{\boldsymbol{u}} \boldsymbol{w}(\boldsymbol{v}, \boldsymbol{u})\big|_{\boldsymbol{u}=\boldsymbol{u}_1} \in \mathbb{R}^{2D_{\mathrm{x}} \times D_{\mathrm{u}}}$ and $\boldsymbol{O}$ denotes the null matrix. Since Assumptions (B3-4) ensure that the rank of $\boldsymbol{W}(\boldsymbol{v})$ is $2D_{\mathrm{x}}$, $\boldsymbol{W}(\boldsymbol{v})\boldsymbol{W}(\boldsymbol{v})^\top$ is a $2D_{\mathrm{x}}$ by $2D_{\mathrm{x}}$ invertible matrix. Thus, multiplying $\boldsymbol{W}(\boldsymbol{v})^\top$ to (61) and taking the inverse of $\boldsymbol{W}(\boldsymbol{v})\boldsymbol{W}(\boldsymbol{v})^\top$ on the right yields

$$\boldsymbol{M}(\boldsymbol{v}) = \boldsymbol{O}. \tag{62}$$

Eq.(62) indicates that $v_i^{(l,m)}(\boldsymbol{z}) = 0$ and $v_i^{(l)}(\boldsymbol{z})v_i^{(m)}(\boldsymbol{z}) = 0$ for all $i, l, m = 1, \ldots, D_{\mathrm{x}}$ with $l \neq m$. This means that each element in $\boldsymbol{v}(\boldsymbol{z})$ is a function of a distinct and single element in $\boldsymbol{z}$ because $\boldsymbol{v}(\boldsymbol{z})$ is an invertible function. Thus, the proof is completed. ∎

Appendix C. Rank condition in the exponential family

We first express the exponential family (12) as

$$\log p(\boldsymbol{s}|\boldsymbol{u}) = \sum_{k=1}^{K} \boldsymbol{\lambda}_k(\boldsymbol{u})^\top \boldsymbol{q}_k(\boldsymbol{s}) - \log Z(\boldsymbol{u}),$$

where $\boldsymbol{\lambda}_k(\boldsymbol{u}) := (\lambda_{1k}(\boldsymbol{u}), \ldots, \lambda_{D_{\mathrm{x}}k}(\boldsymbol{u}))^\top$ and $\boldsymbol{q}_k(\boldsymbol{s}) := (q_{1k}(s_1), \ldots, q_{D_{\mathrm{x}}k}(s_{D_{\mathrm{x}}}))$. Then, a simple computation yields

$$\nabla_{\boldsymbol{u}} \boldsymbol{w}(\boldsymbol{v}, \boldsymbol{u}) = \sum_{k=1}^{K} \boldsymbol{\Lambda}_k \odot \boldsymbol{Q}_k, \tag{63}$$

where $\odot$ denotes the Hadamard product (i.e., elementwise multiplication of two matrices)

$$\boldsymbol{\Lambda}_k := \begin{pmatrix} \nabla_{\boldsymbol{u}} \boldsymbol{\lambda}_k(\boldsymbol{u}) \\ \nabla_{\boldsymbol{u}} \boldsymbol{\lambda}_k(\boldsymbol{u}) \end{pmatrix},$$

and with $q'_{ik}(t) := \frac{\mathrm{d}}{\mathrm{d}t} q_{ik}(t)$ and $q''_{ik}(t) := \frac{\mathrm{d}^2}{\mathrm{d}t^2} q_{ik}(t)$ for $i = 1, \ldots, D_{\mathrm{x}}$,

$$\boldsymbol{Q}_k = \begin{pmatrix} q'_{1k}(v_1) & q'_{1k}(v_1) & \cdots & q'_{1k}(v_1) \\ \vdots & \vdots & \cdots & \vdots \\ q'_{D_{\mathrm{x}}k}(v_{D_{\mathrm{x}}}) & q'_{D_{\mathrm{x}}k}(v_{D_{\mathrm{x}}}) & \cdots & q'_{D_{\mathrm{x}}k}(v_{D_{\mathrm{x}}}) \\ q''_{1k}(v_1) & q''_{1k}(v_1) & \cdots & q''_{1k}(v_1) \\ \vdots & \vdots & \cdots & \vdots \\ q''_{D_{\mathrm{x}}k}(v_{D_{\mathrm{x}}}) & q''_{D_{\mathrm{x}}k}(v_{D_{\mathrm{x}}}) & \cdots & q''_{D_{\mathrm{x}}k}(v_{D_{\mathrm{x}}}) \end{pmatrix}.$$

The rank of $\boldsymbol{\Lambda}_k$ is at most D_{x} under Assumption (B3) because $\boldsymbol{\Lambda}_k$ is a vertical concatenation of the two same matrices of $\nabla_{\boldsymbol{u}} \boldsymbol{\lambda}_k(\boldsymbol{u}) \in \mathbb{R}^{D_{\mathrm{x}} \times D_{\mathrm{u}}}$. On the other hand, $\boldsymbol{Q}_k$ is a rank-one matrix because the column vectors in $\boldsymbol{Q}_k$ are all same. Thus, the rank factorization ensures that the following decompositions of $\boldsymbol{\Lambda}_k$ and $\boldsymbol{Q}_k$ exist:

$$\boldsymbol{\Lambda}_k = \sum_{i=1}^{D_{\mathrm{x}}} \boldsymbol{a}_{ik} \boldsymbol{b}_{ik}^\top \quad \text{and} \quad \boldsymbol{Q}_k = \boldsymbol{c}_k \boldsymbol{d}_k^\top, \tag{64}$$

where $\boldsymbol{a}_{ik}, \boldsymbol{c}_k \in \mathbb{R}^{2D_\mathrm{x}}$ and $\boldsymbol{b}_{ik}, \boldsymbol{d}_k \in \mathbb{R}^{D_\mathrm{u}}$. Substituting (64) into (63) yields

$$\nabla_{\boldsymbol{u}} w(\boldsymbol{v}, \boldsymbol{u}) = \sum_{k=1}^{K} \sum_{i=1}^{D_\mathrm{x}} (\boldsymbol{a}_{ik} \odot \boldsymbol{c}_k)(\boldsymbol{b}_{ik} \odot \boldsymbol{d}_k)^\top. \tag{65}$$

indicating that the rank of $\nabla_{\boldsymbol{u}} w(\boldsymbol{v}, \boldsymbol{u})$ is at most $2D_\mathrm{x}$ when $K > 1$. On the other hand, when $K = 1$, (65) is given by

$$\nabla_{\boldsymbol{u}} w(\boldsymbol{v}, \boldsymbol{u}) = \sum_{i=1}^{D_\mathrm{x}} (\boldsymbol{a}_{i1} \odot \boldsymbol{c}_1)(\boldsymbol{b}_{i1} \odot \boldsymbol{d}_1)^\top,$$

and thus the rank is at most D_x.

Appendix D. Proof of Theorem 3

Our proof is based on the following lemmas:

Lemma 10. *Suppose that $\boldsymbol{D}(\boldsymbol{v})$ is an n by n diagonal matrix whose diagonals $d_i(\boldsymbol{v})$ is a function of $\boldsymbol{v} \in \mathbb{R}^n$, and $\boldsymbol{A}$ and $\boldsymbol{B}$ are n by n constant matrices. Furthermore, the following assumptions are made:*

(1) There exist n points $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$ such that $\boldsymbol{d}(\boldsymbol{v}_1), \boldsymbol{d}(\boldsymbol{v}_2), \ldots, \boldsymbol{d}(\boldsymbol{v}_n)$ are linearly independent where $\boldsymbol{d}(\boldsymbol{v}) := (d_1(\boldsymbol{v}), d_2(\boldsymbol{v}), \ldots, d_n(\boldsymbol{v}))^\top$ is the vector of the diagonal elements in $\boldsymbol{D}(\boldsymbol{v})$.

(2) $\boldsymbol{A}$ and $\boldsymbol{B}$ are of full-rank.

Then, when $\boldsymbol{A} \boldsymbol{D}(\boldsymbol{v}) \boldsymbol{B}$ is a diagonal matrix at least at n points $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$, then both $\boldsymbol{A}$ and $\boldsymbol{B}$ are diagonal matrices multiplied by a permutation matrix.

Lemma 11 (Theorem 4.4.8 in Harville (2006))**.** *Suppose that $\boldsymbol{A}$ is an m by n matrix of rank r. For any m by r matrix $\boldsymbol{B}$ and r by n matrix $\boldsymbol{T}$ such that $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{T}$, both $\boldsymbol{B}$ and $\boldsymbol{T}$ have rank r.*

Lemma 10 is proved in Appendix D.2, while the reader may refer to the proof of Theorem 4.4.8 in Harville (2006) for Lemma 11.

D.1 Main proof

Proof We start by using the same notations and following the same line of the proof until (57) in Section B. Then, we obtain the following equation under Assumptions (B′1-2,6):

$$\sum_{i=1}^{D_\mathrm{x}} q_i(g_i(\boldsymbol{x}), \lambda_i(\boldsymbol{u})) - \log Z(\boldsymbol{u}) - \log p(\boldsymbol{g}(\boldsymbol{x}))$$

$$= \sum_{i=1}^{D_\mathrm{x}} \psi_i^\star(h_{\mathrm{x},i}^\star(\boldsymbol{x}), h_{\mathrm{u},i}^\star(\boldsymbol{u})) + a^\star(\boldsymbol{h}_\mathrm{x}^\star(\boldsymbol{x})) + b^\star(\boldsymbol{h}_\mathrm{u}^\star(\boldsymbol{u})), \tag{66}$$

where $\boldsymbol{g}$ denotes the inverse of the mixing function $\boldsymbol{f}$ and $\boldsymbol{g}(\boldsymbol{x}) = (g_1(\boldsymbol{x}), g_2(\boldsymbol{x}), \ldots, g_{D_{\mathrm{x}}}(\boldsymbol{x}))^{\top}$.
Let us denote $\boldsymbol{z} := \boldsymbol{h}_{\mathrm{x}}^{\star}(\boldsymbol{x})$, $\boldsymbol{v}(\boldsymbol{z}) := \boldsymbol{g}((\boldsymbol{h}_{\mathrm{x}}^{\star})^{-1}(\boldsymbol{z}))$, $r_{\mathrm{u},i} := h_{\mathrm{u},i}^{\star}(\boldsymbol{u})$ and $r_{\mathrm{q},i} := \lambda_i(\boldsymbol{u})$. Then, taking
a cross-derivative with respect to v_l and $r_{\mathrm{q},m}$ to both side on (66) yields

$$\sum_{i=1}^{D_{\mathrm{x}}} \frac{\partial^2 q_i(v_i, r_{\mathrm{q},i})}{\partial v_l \partial r_{\mathrm{q},m}} = \sum_{i=1}^{D_{\mathrm{x}}} \frac{\partial z_i}{\partial v_l} \frac{\partial^2 \psi_i^{\star}(z_i, r_{\mathrm{u},i})}{\partial z_i \partial r_{\mathrm{u},i}} \sum_{j=1}^{D_{\mathrm{x}}} \frac{\partial r_{\mathrm{u},i}}{\partial u_j} \frac{\partial u_j}{\partial r_{\mathrm{q},m}}. \tag{67}$$

To express (67) as a matrix form, we define the following D_{x} by D_{x} diagonal matrices:

$$\boldsymbol{D}_q(\boldsymbol{v}, \boldsymbol{r}_{\mathrm{q}}) := \begin{pmatrix} \frac{\partial^2 q_1(v_1, r_{\mathrm{q},1})}{\partial v_1 \partial r_{\mathrm{q},1}} & 0 & \cdots & 0 \\ 0 & \frac{\partial^2 q_2(v_2, r_{\mathrm{q},2})}{\partial v_2 \partial r_{\mathrm{q},2}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{\partial^2 q_{D_{\mathrm{x}}}(v_{D_{\mathrm{x}}}, r_{\mathrm{q},D_{\mathrm{x}}})}{\partial v_{D_{\mathrm{x}}} \partial r_{\mathrm{q},D_{\mathrm{x}}}} \end{pmatrix} \in \mathbb{R}^{D_{\mathrm{x}} \times D_{\mathrm{x}}}$$

$$\boldsymbol{D}_\psi(\boldsymbol{v}, \boldsymbol{r}_{\mathrm{u}}) := \begin{pmatrix} \frac{\partial^2 \psi_1^{\star}(v_1, r_{\mathrm{u},1})}{\partial v_1 \partial r_{\mathrm{u},1}} & 0 & \cdots & 0 \\ 0 & \frac{\partial^2 \psi_2^{\star}(v_2, r_{\mathrm{u},2})}{\partial v_2 \partial r_{\mathrm{u},2}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{\partial^2 \psi_{D_{\mathrm{x}}}^{\star}(v_{D_{\mathrm{x}}}, r_{\mathrm{u},D_{\mathrm{x}}})}{\partial v_{D_{\mathrm{x}}} \partial r_{\mathrm{u},D_{\mathrm{x}}}} \end{pmatrix} \in \mathbb{R}^{D_{\mathrm{x}} \times D_{\mathrm{x}}},$$

where $\boldsymbol{r}_{\mathrm{q}} = (r_{\mathrm{q},1}, r_{\mathrm{q},2}, \ldots, r_{\mathrm{q},D_{\mathrm{x}}})^{\top}$ and $\boldsymbol{r}_{\mathrm{u}} = (r_{\mathrm{u},1}, r_{\mathrm{u},2}, \ldots, r_{\mathrm{u},D_{\mathrm{x}}})^{\top}$. Then, (67) can be expressed
as

$$\boldsymbol{D}_{\mathrm{q}}(\boldsymbol{v}, \boldsymbol{r}_{\mathrm{q}}) = \boldsymbol{J}_{\boldsymbol{z}}(\boldsymbol{v}) \boldsymbol{D}_\psi(\boldsymbol{v}, \boldsymbol{r}_{\mathrm{u}}) \boldsymbol{J}_{\boldsymbol{r}_{\mathrm{u}}}(\boldsymbol{u}) \boldsymbol{J}_{\boldsymbol{u}}(\boldsymbol{r}_{\mathrm{q}}), \tag{68}$$

where $\boldsymbol{J}_{\boldsymbol{z}}(\boldsymbol{v}) := \nabla_{\boldsymbol{v}} \boldsymbol{z} \in \mathbb{R}^{D_{\mathrm{x}} \times D_{\mathrm{x}}}$, $\boldsymbol{J}_{\boldsymbol{r}_{\mathrm{u}}}(\boldsymbol{u}) := \nabla_{\boldsymbol{u}} \boldsymbol{r}_{\mathrm{u}} \in \mathbb{R}^{D_{\mathrm{x}} \times D_{\mathrm{u}}}$ and $\boldsymbol{J}_{\boldsymbol{u}}(\boldsymbol{r}_{\mathrm{q}}) := \nabla_{\boldsymbol{r}_{\mathrm{q}}} \boldsymbol{u} \in \mathbb{R}^{D_{\mathrm{u}} \times D_{\mathrm{x}}}$
are the Jacobian matrices of $\boldsymbol{z}$, $\boldsymbol{r}_{\mathrm{u}}$ and $\boldsymbol{u}$, respectively. We fix $\boldsymbol{u}$ at $\boldsymbol{u} = \boldsymbol{u}_1$ and express (68) as

$$\boldsymbol{D}_{\mathrm{q}}^1(\boldsymbol{v}) = \boldsymbol{J}_{\boldsymbol{z}}(\boldsymbol{v}) \boldsymbol{D}_\psi^1(\boldsymbol{v}) \boldsymbol{J}^1, \tag{69}$$

where $\boldsymbol{D}_{\mathrm{q}}^1(\boldsymbol{v}) := \boldsymbol{D}_{\mathrm{q}}(\boldsymbol{v}, \boldsymbol{r}_{\mathrm{q}})$ at $\boldsymbol{r}_{\mathrm{q}} = \boldsymbol{\lambda}(\boldsymbol{u}_1)$, $\boldsymbol{D}_\psi^1(\boldsymbol{v}) := \boldsymbol{D}_\psi(\boldsymbol{v}, \boldsymbol{r}_{\mathrm{u}})$ at $\boldsymbol{r}_{\mathrm{u}} = \boldsymbol{h}_{\mathrm{u}}^{\star}(\boldsymbol{u}_1)$, and $\boldsymbol{J}^1 :=$
$\boldsymbol{J}_{\boldsymbol{r}_{\mathrm{u}}}(\boldsymbol{u}_1) \boldsymbol{J}_{\boldsymbol{u}}(\boldsymbol{r}_{\mathrm{q}})$ at $\boldsymbol{r}_{\mathrm{q}} = \boldsymbol{\lambda}(\boldsymbol{u}_1)$. We can have a similar matrix expression as (69) at $\boldsymbol{u} = \boldsymbol{u}_2$ as

$$\boldsymbol{D}_{\mathrm{q}}^2(\boldsymbol{v}) = \boldsymbol{J}_{\boldsymbol{z}}(\boldsymbol{v}) \boldsymbol{D}_\psi^2(\boldsymbol{v}) \boldsymbol{J}^2. \tag{70}$$

Here, we note that both $\boldsymbol{J}^1 \in \mathbb{R}^{D_{\mathrm{x}} \times D_{\mathrm{x}}}$ and $\boldsymbol{J}^2 \in \mathbb{R}^{D_{\mathrm{x}} \times D_{\mathrm{x}}}$ have at most rank D_{x} by Assumption (B'3).

Next, we confirm that $\boldsymbol{J}^1$, $\boldsymbol{J}^2$, $\boldsymbol{D}_\psi^1(\boldsymbol{v})$ and $\boldsymbol{D}_\psi^2(\boldsymbol{v})$ for all $\boldsymbol{v}$ have rank D_{x}, and thus are invertible.
To this end, we denote the i-th diagonals in $\boldsymbol{D}_{\mathrm{q}}^1(\boldsymbol{v})$ and $\boldsymbol{D}_{\mathrm{q}}^2(\boldsymbol{v})$ by

$$\alpha_i^1(v_i) := \frac{\partial^2 q_i(v_i, r)}{\partial v_i \partial r}\bigg|_{r=\lambda_i(\boldsymbol{u}_1)} \quad \text{and} \quad \alpha_i^2(v_i) := \frac{\partial^2 q_i(v_i, r)}{\partial v_i \partial r}\bigg|_{r=\lambda_i(\boldsymbol{u}_2)},$$

respectively. From Assumptions (B'4), $\boldsymbol{D}_{\mathrm{q}}^1(\boldsymbol{v})$ has nonzero diagonals $\alpha_i^1(v_i)$ and rank D_{x} for all $\boldsymbol{v}$.
Then, applying Lemma 11 to (69) ensures that $\boldsymbol{D}_\psi^1(\boldsymbol{v}) \boldsymbol{J}^1$ have rank D_{x}. By applying Lemma 11

to $\boldsymbol{D}_\psi^1(\boldsymbol{v})\boldsymbol{J}^1$ again, it can be shown that both $\boldsymbol{D}_\psi^1(\boldsymbol{v})$ and $\boldsymbol{J}^1$ also have rank D_{x}. Similarly, we can prove that both $\boldsymbol{D}_\psi^2(\boldsymbol{v})$ and $\boldsymbol{J}^2$ have rank D_{x} under Assumption (B$'$4).

Finally, we show that $\boldsymbol{J}_{\boldsymbol{z}}(\boldsymbol{v})$ is the product of a permutation and diagonal matrices. Since $\boldsymbol{D}_\psi^1(\boldsymbol{v})$ and $\boldsymbol{J}^1$ are invertible, from (69), we have

$$\boldsymbol{J}_{\boldsymbol{z}}(\boldsymbol{v}) = \boldsymbol{D}_{\mathrm{q}}^1(\boldsymbol{v})[\boldsymbol{J}^1]^{-1}[\boldsymbol{D}_\psi^1(\boldsymbol{v})]^{-1}. \tag{71}$$

Substituting (71) into (70), and multiplying $\boldsymbol{J}^1$ and the inverses of $\boldsymbol{D}_{\mathrm{q}}^1(\boldsymbol{v})$ and $\boldsymbol{J}^2$ to both sides yields

$$\boldsymbol{J}^1\left[[\boldsymbol{D}_{\mathrm{q}}^1(\boldsymbol{v})]^{-1}\boldsymbol{D}_{\mathrm{q}}^2(\boldsymbol{v})\right][\boldsymbol{J}^2]^{-1} = [\boldsymbol{D}_\psi^1(\boldsymbol{v})]^{-1}\boldsymbol{D}_\psi^2(\boldsymbol{v}). \tag{72}$$

Then, we compactly express all of diagonals in $[\boldsymbol{D}_{\mathrm{q}}^1(\boldsymbol{v})]^{-1}\boldsymbol{D}_{\mathrm{q}}^2(\boldsymbol{v})$ as the following vector:

$$\boldsymbol{\alpha}(\boldsymbol{v}) := \left(\frac{\alpha_1^2(v_1)}{\alpha_1^1(v_1)}, \frac{\alpha_2^2(v_2)}{\alpha_2^1(v_2)}, \ldots, \frac{\alpha_{D_{\mathrm{x}}}^2(v_{D_{\mathrm{x}}})}{\alpha_{D_{\mathrm{x}}}^1(v_{D_{\mathrm{x}}})}\right)^\top.$$

By Assumption (B$'$5), $\boldsymbol{\alpha}(\boldsymbol{v}_1), \boldsymbol{\alpha}(\boldsymbol{v}_2), \ldots, \boldsymbol{\alpha}(\boldsymbol{v}_{D_{\mathrm{x}}})$ are linearly independent. Since $\boldsymbol{J}^1$ and $\boldsymbol{J}^2$ are proved to have rank D_{x} and thus constant full-rank matrices, applying Lemma 10 ensures that $\boldsymbol{J}^1$ and $\boldsymbol{J}^2$ are diagonal matrices multiplied by a permutation matrix. Thus, it follows from (71) that $\boldsymbol{J}_{\boldsymbol{z}}(\boldsymbol{v})$ is also a diagonal matrix multiplied by a permutation matrix. Thus, each $v_i(\boldsymbol{z})(= s_i)$ corresponds to a single and distinct element in $\boldsymbol{z}(= \boldsymbol{h}_{\mathrm{x}}^\star(\boldsymbol{x}))$. The proof is completed. ∎

D.2 Proof of Lemma 10

Proof Let us first denote the (l, m)-th elements of $\boldsymbol{A}$ and $\boldsymbol{B}$ by a_{lm} and b_{lm}. The (l, m)-th element in $\boldsymbol{A}\boldsymbol{D}(\boldsymbol{v})\boldsymbol{B}$ is given by

$$\sum_{i=1}^n d_i(\boldsymbol{v})a_{li}b_{im}.$$

Since off-diagonal elements in $\boldsymbol{A}\boldsymbol{D}(\boldsymbol{v})\boldsymbol{B}$ are all zeros at $\boldsymbol{v} = \boldsymbol{v}_j$, $j = 1, \ldots, n$,

$$\sum_{i=1}^n d_i(\boldsymbol{v}_j)a_{li}b_{im} = \boldsymbol{d}(\boldsymbol{v}_j)^\top\boldsymbol{c}^{lm} = 0 \quad (l \neq m), \tag{73}$$

where

$$\boldsymbol{c}^{lm} := (a_{l1}b_{1m}, a_{l2}b_{2m}, \ldots, a_{ln}b_{nm})^\top.$$

Collecting $\boldsymbol{d}(\boldsymbol{v})$ over the n points $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$ based on (73) yields

$$\widetilde{\boldsymbol{D}}^\top\boldsymbol{c}^{lm} = \boldsymbol{0}, \tag{74}$$

where

$$\widetilde{\boldsymbol{D}} := [\boldsymbol{d}(\boldsymbol{v}_1), \boldsymbol{d}(\boldsymbol{v}_2), \ldots, \boldsymbol{d}(\boldsymbol{v}_n)] \in \mathbb{R}^{n \times n}.$$

By Assumption (1), $\widetilde{\boldsymbol{D}}$ is of full-rank and invertible. Thus, from (74), we obtain $\boldsymbol{c}^{lm} = \boldsymbol{0}$ for $l, m = 1, \ldots, n$ with $l \neq m$, indicating that

$$a_{li}b_{im} = 0 \quad (i, l, m = 1, \ldots, n \text{ and } l \neq m). \tag{75}$$

Next, we show that both the i-th column vector in $\boldsymbol{A}$ and the i-th row vector in $\boldsymbol{B}$ have a single nonzero element, while the other elements are zeros. We first suppose that the i-th column vector in $\boldsymbol{A}$ has at least two nonzeros elements such that $a_{li} \neq 0$ and $a_{l'i} \neq 0$ for $l \neq l'$. $a_{li} \neq 0$ implies that $b_{im} = 0$ except for $m = l$, but $a_{l'i} \neq 0$ ensures $b_{im} = 0$ at $m = l$. Thus, the i-th row vector in $\boldsymbol{B}$ must be the zero vector. However, this contradicts to the assumption that $\boldsymbol{B}$ is of full-rank. Therefore, the i-th column vector of $\boldsymbol{A}$ must have the single nonzero element. Similarly, we can prove that the i-th row vector in $\boldsymbol{B}$ also has a single nonzero element. Thus, by the full-rank assumption of $\boldsymbol{A}$ and $\boldsymbol{B}$, all column and row vectors in $\boldsymbol{A}$ and $\boldsymbol{B}$ must have the single nonzero element at distinct positions. Thus, $\boldsymbol{A}$ and $\boldsymbol{B}$ are equal to diagonal matrices multiplied by a permutation matrix. ∎

Appendix E. Proof of Theorem 4

Proof Taking the gradient of both sides on (18) with respect to $\boldsymbol{u}$ yields

$$\nabla_{\boldsymbol{u}} \log p(\boldsymbol{x}|\boldsymbol{u}) = \nabla_{\boldsymbol{u}} \psi^\star(\boldsymbol{h}_{\mathrm{x}}^\star(\boldsymbol{x}), \boldsymbol{h}_{\mathrm{u}}^\star(\boldsymbol{u})) + \nabla_{\boldsymbol{u}} b^\star(\boldsymbol{h}_{\mathrm{u}}^\star(\boldsymbol{u})). \tag{76}$$

By the change of variables by $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{s}, \boldsymbol{n})$ in the generative model (17), we re-express (76) as

$$\nabla_{\boldsymbol{u}} \log p(\boldsymbol{s}, \boldsymbol{n}|\boldsymbol{u}) = \nabla_{\boldsymbol{u}} \psi^\star(\boldsymbol{h}_{\mathrm{x}}^\star(\boldsymbol{f}(\boldsymbol{s}, \boldsymbol{n})), \boldsymbol{h}_{\mathrm{u}}^\star(\boldsymbol{u})) + \nabla_{\boldsymbol{u}} b^\star(\boldsymbol{h}_{\mathrm{u}}^\star(\boldsymbol{u})), \tag{77}$$

where we note that the Jacobian term due to the change of variables is deleted by the differential operator $\nabla_{\boldsymbol{u}}$.

Next, with $\boldsymbol{v} := \boldsymbol{h}_{\mathrm{x}}^\star(\boldsymbol{f}(\boldsymbol{s}, \boldsymbol{n}))$, we compute the gradient of the left-hand side on (77) with respect to $\boldsymbol{n}$ as

$$\begin{aligned}
\nabla_{\boldsymbol{n}} \nabla_{\boldsymbol{u}} \log p(\boldsymbol{s}, \boldsymbol{n}|\boldsymbol{u}) &= \nabla_{\boldsymbol{n}}[\nabla_{\boldsymbol{u}}\{\log p(\boldsymbol{s}|\boldsymbol{u}) + \log p(\boldsymbol{n})\}] \\
&= \nabla_{\boldsymbol{n}}[\nabla_{\boldsymbol{u}} \log p(\boldsymbol{s}|\boldsymbol{u})] \\
&= \boldsymbol{O},
\end{aligned} \tag{78}$$

where we used Assumption (C1) that $\boldsymbol{s} \perp \boldsymbol{n}$, $\boldsymbol{u} \perp \boldsymbol{n}$ and $\boldsymbol{s} \not\perp \boldsymbol{u}$. On the other hand, the gradient of the right-hand side on (77) is given by

$$\nabla_{\boldsymbol{n}}[\nabla_{\boldsymbol{u}} \psi^\star(\boldsymbol{v}, \boldsymbol{h}_{\mathrm{u}}^\star(\boldsymbol{u}))] = [\nabla_{\boldsymbol{v}} \nabla_{\boldsymbol{u}} \psi^\star(\boldsymbol{v}, \boldsymbol{h}_{\mathrm{u}}^\star(\boldsymbol{u}))] \boldsymbol{J}_{\boldsymbol{v}}(\boldsymbol{n}), \tag{79}$$

where $\nabla_{\boldsymbol{v}} \nabla_{\boldsymbol{u}} \psi^\star(\boldsymbol{v}, \boldsymbol{h}_{\mathrm{u}}^\star(\boldsymbol{u})) \in \mathbb{R}^{D_{\mathrm{u}} \times d_{\mathrm{x}}}$ and $\boldsymbol{J}_{\boldsymbol{v}}(\boldsymbol{n}) := \nabla_{\boldsymbol{n}} \boldsymbol{v} \in \mathbb{R}^{d_{\mathrm{x}} \times (D_{\mathrm{x}} - d_{\mathrm{x}})}$ is the Jacobian matrix of $\boldsymbol{v}$ at $\boldsymbol{n}$. We fix $\boldsymbol{u}$ at $\boldsymbol{u} = \boldsymbol{u}_1$ and equate (79) with (78) based on (77) as

$$\boldsymbol{M}(\boldsymbol{v}) \boldsymbol{J}_{\boldsymbol{v}}(\boldsymbol{n}) = \boldsymbol{O}, \tag{80}$$

where $\boldsymbol{M}(\boldsymbol{v}) := \nabla_{\boldsymbol{v}} \nabla_{\boldsymbol{u}} \psi^\star(\boldsymbol{v}, \boldsymbol{h}_{\mathrm{u}}^\star(\boldsymbol{u}))\big|_{\boldsymbol{u}=\boldsymbol{u}_1} \in \mathbb{R}^{D_{\mathrm{u}} \times d_{\mathrm{x}}}$. By Assumption (C4), the rank of $\boldsymbol{M}(\boldsymbol{v})$ is d_{x}, and thus the inverse of $\boldsymbol{M}(\boldsymbol{v})^\top \boldsymbol{M}(\boldsymbol{v})$ exists. Multiplying $\boldsymbol{M}(\boldsymbol{v})^\top$ on the left and taking the inverse of $\boldsymbol{M}(\boldsymbol{v})^\top \boldsymbol{M}(\boldsymbol{v})$ to (80) indicates $\boldsymbol{J}_{\boldsymbol{v}}(\boldsymbol{n}) = \boldsymbol{O}$. Recall that $\boldsymbol{v} = \boldsymbol{h}_{\mathrm{x}}^\star \circ \boldsymbol{f}(\boldsymbol{s}, \boldsymbol{n}) = \boldsymbol{h}_{\mathrm{x}}^\star(\boldsymbol{x})$. Since $\boldsymbol{f}(\boldsymbol{s}, \boldsymbol{n})$ and $\boldsymbol{h}_{\mathrm{x}}^\star(\boldsymbol{x})$ are invertible and surjective respectively, $\boldsymbol{J}_{\boldsymbol{v}}(\boldsymbol{n}) = \boldsymbol{O}$ ensures that $\boldsymbol{v}$ is a (vector-valued) function of only $\boldsymbol{s}$. Thus, the proof is completed. ∎

Appendix F. Derivation of (19)

By Assumption (C1) that $s \perp n$, $u \perp n$ and $s \not\perp u$, (77) can be written as

$$\nabla_{\boldsymbol{u}} \log p(\boldsymbol{s}|\boldsymbol{u}) = \nabla_{\boldsymbol{u}}\psi^{\star}(\boldsymbol{v}, \boldsymbol{h}_{\mathrm{u}}^{\star}(\boldsymbol{u})) + \nabla_{\boldsymbol{u}}b^{\star}(\boldsymbol{h}_{\mathrm{u}}^{\star}(\boldsymbol{u})),$$

where $\boldsymbol{v} := \boldsymbol{h}_{\mathrm{x}}^{\star}(\boldsymbol{f}(\boldsymbol{s}, \boldsymbol{n}))$. Taking the gradient with respect to $\boldsymbol{v}$ yields

$$\boldsymbol{J}_{\boldsymbol{v}}(\boldsymbol{s})[\nabla_{\boldsymbol{s}}\nabla_{\boldsymbol{u}} \log p(\boldsymbol{s}|\boldsymbol{u})] = \nabla_{\boldsymbol{v}}\nabla_{\boldsymbol{u}}\psi^{\star}(\boldsymbol{v}, \boldsymbol{h}_{\mathrm{u}}^{\star}(\boldsymbol{u})),$$

where $\boldsymbol{J}_{\boldsymbol{v}}(\boldsymbol{s}) := \nabla_{\boldsymbol{s}}\boldsymbol{v}$ denotes the Jacobian of $\boldsymbol{v}$ at $\boldsymbol{s}$. Thus, we obtain (19) by fixing $\boldsymbol{u}$ at $\boldsymbol{u} = \boldsymbol{u}_1$.

Appendix G. Proof of Proposition 5

Proof Firstly, we consider the contamination model 2. By definition, the estimator $\boldsymbol{\theta}^{\star}$ satisfies the following equation:

$$\nabla_{\boldsymbol{\theta}} \left\{ E_{\mathrm{xu}}[r_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{U})] - \log E_{\mathrm{x}\times\mathrm{u}}[e^{r_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{U})}] \right\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{\star}} = \boldsymbol{0}, \tag{81}$$

while $\boldsymbol{\theta}_{\epsilon}$ associated with the contaminated densities fulfills

$$\nabla_{\boldsymbol{\theta}} \left\{ \bar{E}_{\mathrm{xu}}[r_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{U})] - \log \bar{E}_{\mathrm{x}\times\mathrm{u}}[e^{r_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{U})}] \right\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\epsilon}} = \boldsymbol{0}, \tag{82}$$

where $\bar{E}_{\mathrm{xu}}$ and $\bar{E}_{\mathrm{x}\times\mathrm{u}}$ denote the expectations over $\bar{p}(\boldsymbol{x}, \boldsymbol{u})$ and $\bar{p}(\boldsymbol{x})\bar{p}(\boldsymbol{u})$, respectively. Applying the Taylor series expansion of (82) around $\boldsymbol{\theta}^{\star}$ yields

$$\boldsymbol{0} = \nabla_{\boldsymbol{\theta}} \left\{ \bar{E}_{\mathrm{xu}}[r_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{U})] - \log \bar{E}_{\mathrm{x}\times\mathrm{u}}[e^{r_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{U})}] \right\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{\star}} + \bar{\boldsymbol{V}}_{\mathrm{DV}}(\boldsymbol{\theta}_{\epsilon} - \boldsymbol{\theta}^{\star}) + O(\|\boldsymbol{\theta}_{\epsilon} - \boldsymbol{\theta}^{\star}\|^2). \tag{83}$$

where $\bar{\boldsymbol{V}}_{\mathrm{DV}}$ is the Hessian matrix at $\boldsymbol{\theta} = \boldsymbol{\theta}^{\star}$ and defined by

$$\bar{\boldsymbol{V}}_{\mathrm{DV}} := \nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}} \left\{ \bar{E}_{\mathrm{xu}}[r_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{U})] - \log \bar{E}_{\mathrm{x}\times\mathrm{u}}[e^{r_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{U})}] \right\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{\star}}.$$

Let us denote $\nabla_{\boldsymbol{\theta}}r_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{u})$ by $\boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{u})$. We recall that the contaminated densities in the contamination model 2 are defined as

$$\bar{p}(\boldsymbol{x}, \boldsymbol{u}) = (1 - \epsilon)p(\boldsymbol{x}, \boldsymbol{u}) + \epsilon\delta_{\bar{\boldsymbol{x}}}(\boldsymbol{x})\delta_{\bar{\boldsymbol{u}}}(\boldsymbol{u})$$
$$\bar{p}(\boldsymbol{x}) = (1 - \epsilon)p(\boldsymbol{x}) + \epsilon\delta_{\bar{\boldsymbol{x}}}(\boldsymbol{x})$$
$$\bar{p}(\boldsymbol{u}) = (1 - \epsilon)p(\boldsymbol{u}) + \epsilon\delta_{\bar{\boldsymbol{u}}}(\boldsymbol{u}).$$

Then, with sufficiently small ϵ, the first term on the right-hand side of (83) is given by

$$\nabla_{\boldsymbol{\theta}} \left\{ \bar{E}_{\mathrm{xu}}[r_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{U})] - \log \bar{E}_{\mathrm{x}\times\mathrm{u}}[e^{r_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{U})}] \right\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{\star}} = \bar{E}_{\mathrm{xu}}[\boldsymbol{g}_{\boldsymbol{\theta}^{\star}}(\boldsymbol{X}, \boldsymbol{U})] - \frac{\bar{E}_{\mathrm{x}\times\mathrm{u}}[e^{r_{\boldsymbol{\theta}^{\star}}(\boldsymbol{X}, \boldsymbol{U})}\boldsymbol{g}_{\boldsymbol{\theta}^{\star}}(\boldsymbol{X}, \boldsymbol{U})]}{\bar{E}_{\mathrm{x}\times\mathrm{u}}[e^{r_{\boldsymbol{\theta}^{\star}}(\boldsymbol{X}, \boldsymbol{U})}]}$$
$$\simeq \bar{E}_{\mathrm{xu}}[\boldsymbol{g}_{\boldsymbol{\theta}^{\star}}(\boldsymbol{X}, \boldsymbol{U})] - \bar{E}_{\mathrm{x}\times\mathrm{u}}[e^{r_{\boldsymbol{\theta}^{\star}}(\boldsymbol{X}, \boldsymbol{U})}\boldsymbol{g}_{\boldsymbol{\theta}^{\star}}(\boldsymbol{X}, \boldsymbol{U})], \tag{84}$$

45

where $\simeq$ denotes the equality up to terms of ϵ^2, and we applied the following relation:

$$\bar{E}_{\mathrm{x}\times\mathrm{u}}[e^{r_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})}] \simeq E_{\mathrm{x}\times\mathrm{u}}[e^{r_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})}] + \epsilon\left(E_{\mathrm{x}}[e^{r_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\bar{\boldsymbol{u}})}] + E_{\mathrm{u}}[e^{r_{\boldsymbol{\theta}^\star}(\bar{\boldsymbol{x}},\boldsymbol{U})}] - 2E_{\mathrm{x}\times\mathrm{u}}[e^{r_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})}]\right) = 1,$$

where we used

$$E_{\mathrm{x}\times\mathrm{u}}[e^{r_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})}] = E_{\mathrm{x}\times\mathrm{u}}\left[\frac{p(\boldsymbol{X},\boldsymbol{U})}{p(\boldsymbol{X})p(\boldsymbol{U})}\right] = 1, \tag{85}$$

$$E_{\mathrm{x}}[e^{r_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\bar{\boldsymbol{u}})}] = \int p(\boldsymbol{x}|\bar{\boldsymbol{u}})d\boldsymbol{x} = 1 \quad \text{and} \quad E_{\mathrm{u}}[e^{r_{\boldsymbol{\theta}^\star}(\bar{\boldsymbol{x}},\boldsymbol{U})}] = \int p(\boldsymbol{u}|\bar{\boldsymbol{x}})d\boldsymbol{u} = 1, \tag{86}$$

based on the assumption that $r_{\boldsymbol{\theta}^\star}(\boldsymbol{x},\boldsymbol{u}) = \log\frac{p(\boldsymbol{x},\boldsymbol{u})}{p(\boldsymbol{x})p(\boldsymbol{u})}$. Then, we substitute

$$\bar{E}_{\mathrm{xu}}[\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})] = (1-\epsilon)E_{\mathrm{xu}}[\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})] + \epsilon\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\bar{\boldsymbol{x}},\bar{\boldsymbol{u}}),$$

and

$$\bar{E}_{\mathrm{x}\times\mathrm{u}}[e^{r_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})}\boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{X},\boldsymbol{U})] \simeq (1-\epsilon)E_{\mathrm{x}\times\mathrm{u}}[e^{r_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})}\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})] + \epsilon\left(E_{\mathrm{x}}[e^{r_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\bar{\boldsymbol{u}})}\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\bar{\boldsymbol{u}})]\right.$$
$$\left. + E_{\mathrm{u}}[e^{r_{\boldsymbol{\theta}^\star}(\bar{\boldsymbol{x}},\boldsymbol{U})}\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\bar{\boldsymbol{x}},\boldsymbol{U})] - E_{\mathrm{x}\times\mathrm{u}}[e^{r_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})}\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})]\right),$$

into (84), and have

$$\nabla_{\boldsymbol{\theta}}\left\{\bar{E}_{\mathrm{xu}}[r_{\boldsymbol{\theta}}(\boldsymbol{X},\boldsymbol{U})] - \log\bar{E}_{\mathrm{x}\times\mathrm{u}}[e^{r_{\boldsymbol{\theta}}(\boldsymbol{X},\boldsymbol{U})}]\right\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\star}$$
$$\simeq (1-\epsilon)\left(E_{\mathrm{xu}}[\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})] - E_{\mathrm{x}\times\mathrm{u}}[e^{r_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})}\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})]\right)$$
$$+ \epsilon\left(\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\bar{\boldsymbol{x}},\bar{\boldsymbol{u}}) + E_{\mathrm{x}\times\mathrm{u}}[e^{r_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})}\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})] - E_{\mathrm{x}}[e^{r_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\bar{\boldsymbol{u}})}\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\bar{\boldsymbol{u}})] - E_{\mathrm{u}}[e^{r_{\boldsymbol{\theta}^\star}(\bar{\boldsymbol{x}},\boldsymbol{U})}\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\bar{\boldsymbol{x}},\boldsymbol{U})]\right)$$
$$= \epsilon\left(\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\bar{\boldsymbol{x}},\bar{\boldsymbol{u}}) + E_{\mathrm{xu}}[\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})] - E_{\mathrm{x}}[e^{r_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\bar{\boldsymbol{u}})}\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\bar{\boldsymbol{u}}) - E_{\mathrm{u}}[e^{r_{\boldsymbol{\theta}^\star}(\bar{\boldsymbol{x}},\boldsymbol{U})}\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\bar{\boldsymbol{x}},\boldsymbol{U})]\right), \tag{87}$$

where we applied $E_{\mathrm{xu}}[\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})] = E_{\mathrm{x}\times\mathrm{u}}[e^{r_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})}\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})]$, which is proved from (81) and (85) as follows:

$$E_{\mathrm{xu}}[\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})] - E_{\mathrm{x}\times\mathrm{u}}[e^{r_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})}\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})] = E_{\mathrm{xu}}[\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})] - \frac{E_{\mathrm{x}\times\mathrm{u}}[e^{r_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})}\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})]}{E_{\mathrm{x}\times\mathrm{u}}[e^{r_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})}]}$$
$$= \nabla_{\boldsymbol{\theta}}\left\{E_{\mathrm{xu}}[r_{\boldsymbol{\theta}}(\boldsymbol{X},\boldsymbol{U})] - \log E_{\mathrm{x}\times\mathrm{u}}[e^{r_{\boldsymbol{\theta}}(\boldsymbol{X},\boldsymbol{U})}]\right\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\star}$$
$$= \boldsymbol{0}.$$

Substituting (87) into (83) yields

$$\boldsymbol{0} \simeq \epsilon\left\{\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\bar{\boldsymbol{x}},\bar{\boldsymbol{u}}) + E_{\mathrm{xu}}[\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})] - E_{\mathrm{x}}[e^{r_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\bar{\boldsymbol{u}})}\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\bar{\boldsymbol{u}})] - E_{\mathrm{u}}[e^{r_{\boldsymbol{\theta}^\star}(\bar{\boldsymbol{x}},\boldsymbol{U})}\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\bar{\boldsymbol{x}},\boldsymbol{U})]\right\}$$
$$+ \bar{\boldsymbol{V}}_{\mathrm{DV}}(\boldsymbol{\theta}_\epsilon - \boldsymbol{\theta}^\star) + O(\|\boldsymbol{\theta}_\epsilon - \boldsymbol{\theta}^\star\|^2) \tag{88}$$

In the limit of $\epsilon \to 0$, $O(\|\boldsymbol{\theta}_\epsilon - \boldsymbol{\theta}^\star\|^2)$ in (88) quickly converges to zero and can be negligible, and $\bar{\boldsymbol{V}}_{\mathrm{DV}}$ approaches

$$-E_{\mathrm{xu}}[\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})^\top] + E_{\mathrm{xu}}[\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})]E_{\mathrm{xu}}[\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X},\boldsymbol{U})]^\top =: \boldsymbol{V}_{\mathrm{DV}},$$

where we applied (85). By dividing both sides on (88) by ϵ and taking the limit of $\epsilon \to 0$, we obtain

$$
\begin{aligned}
\boldsymbol{V}_{\mathrm{DV}}\mathrm{IF}_{\mathrm{DV}}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}) \\
= \boldsymbol{g}_{\boldsymbol{\theta}^\star}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}) + E_{\mathrm{xu}}[\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X}, \boldsymbol{U})] - E_{\mathrm{x}}[e^{r_{\boldsymbol{\theta}^\star}(\boldsymbol{X}, \bar{\boldsymbol{u}})}\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\boldsymbol{X}, \bar{\boldsymbol{u}})] - E_{\mathrm{u}}[e^{r_{\boldsymbol{\theta}^\star}(\bar{\boldsymbol{x}}, \boldsymbol{U})}\boldsymbol{g}_{\boldsymbol{\theta}^\star}(\bar{\boldsymbol{x}}, \boldsymbol{U})].
\end{aligned}
$$

Thus, applying the inverse of $\boldsymbol{V}_{\mathrm{DV}}$ yields (31).

For the contamination model 1, we recall that the contaminated densities are given as follows:

$$
\begin{aligned}
\bar{p}(\boldsymbol{x}, \boldsymbol{u}) &= (1 - \epsilon)p(\boldsymbol{x}, \boldsymbol{u}) + \epsilon\delta_{\bar{\boldsymbol{u}}}(\boldsymbol{u})p(\boldsymbol{x}) \\
\bar{p}(\boldsymbol{x}) &= p(\boldsymbol{x}) \\
\bar{p}(\boldsymbol{u}) &= (1 - \epsilon)p(\boldsymbol{u}) + \epsilon\delta_{\bar{\boldsymbol{u}}}(\boldsymbol{u}).
\end{aligned}
$$

By following the same line of the proof, we reach the influence function (29) for the contamination model 1. $\blacksquare$

Appendix H. Proof of Proposition 8

H.1 Main proof

We employ the following lemma, which is proved in Appendix H.2:

Lemma 12. *Constant ν_1 is sufficiently small under Assumptions (D1-2).*

Thus, we next focus on proving that (38) holds when ν_1 is sufficiently small. Denoting $e^{r(\boldsymbol{x}, \boldsymbol{u})}$ by $\varphi(\boldsymbol{x}, \boldsymbol{u})$, for the contamination model 1, the expectations over $\bar{p}(\boldsymbol{x}, \boldsymbol{u})$ and $\bar{p}(\boldsymbol{x})\bar{p}(\boldsymbol{u})$ can be expressed as

$$
\begin{aligned}
\bar{E}_{\mathrm{xu}}\left[\left(\frac{\varphi(\boldsymbol{X}, \boldsymbol{U})^{\gamma+1}}{1 + \varphi(\boldsymbol{X}, \boldsymbol{U})^{\gamma+1}}\right)^{\frac{\gamma}{\gamma+1}}\right] &= (1 - \epsilon)E_{\mathrm{xu}}\left[\left(\frac{\varphi(\boldsymbol{X}, \boldsymbol{U})^{\gamma+1}}{1 + \varphi(\boldsymbol{X}, \boldsymbol{U})^{\gamma+1}}\right)^{\frac{\gamma}{\gamma+1}}\right] \\
&\quad + \epsilon\iint p(\boldsymbol{x})\delta(\boldsymbol{u}|\boldsymbol{x})\left(\frac{\varphi(\boldsymbol{x}, \boldsymbol{u})^{\gamma+1}}{1 + \varphi(\boldsymbol{x}, \boldsymbol{u})^{\gamma+1}}\right)^{\frac{\gamma}{\gamma+1}}\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u} \\
\bar{E}_{\mathrm{x}\times\mathrm{u}}\left[\left(\frac{1}{1 + \varphi(\boldsymbol{X}, \boldsymbol{U})^{\gamma+1}}\right)^{\frac{\gamma}{\gamma+1}}\right] &= (1 - \epsilon)E_{\mathrm{x}\times\mathrm{u}}\left[\left(\frac{1}{1 + \varphi(\boldsymbol{X}, \boldsymbol{U})^{\gamma+1}}\right)^{\frac{\gamma}{\gamma+1}}\right] \\
&\quad + \epsilon\iint p(\boldsymbol{x})\delta(\boldsymbol{u})\left(\frac{1}{1 + \varphi(\boldsymbol{x}, \boldsymbol{u})^{\gamma+1}}\right)^{\frac{\gamma}{\gamma+1}}\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u}.
\end{aligned}
$$

Substituting these expectations into $\bar{J}_\gamma(\varphi)$ yields

$$
\begin{aligned}
\bar{J}_\gamma(\varphi) = &-\frac{1}{\gamma}\log\left[E_{\mathrm{xu}}\left[\left(\frac{\varphi(\boldsymbol{X}, \boldsymbol{U})^{\gamma+1}}{1 + \varphi(\boldsymbol{X}, \boldsymbol{U})^{\gamma+1}}\right)^{\frac{\gamma}{\gamma+1}}\right] + E_{\mathrm{x}\times\mathrm{u}}\left[\left(\frac{1}{1 + \varphi(\boldsymbol{X}, \boldsymbol{U})^{\gamma+1}}\right)^{\frac{\gamma}{\gamma+1}}\right] + \frac{\epsilon}{1-\epsilon}\nu_1\right] \\
&- \frac{1}{\gamma}\log(1 - \epsilon).
\end{aligned}
$$

47

By applying $\log(z + \nu_1) = \log(z) + O(\nu_1)$ with a sufficiently small ν_1,

$$\bar{J}_\gamma(\varphi) = -\frac{1}{\gamma} \log \left[E_{\mathrm{xu}} \left[\left(\frac{\varphi(\boldsymbol{X}, \boldsymbol{U})^{\gamma+1}}{1 + \varphi(\boldsymbol{X}, \boldsymbol{U})^{\gamma+1}} \right)^{\frac{\gamma}{\gamma+1}} \right] + E_{\mathrm{x \times u}} \left[\left(\frac{1}{1 + \varphi(\boldsymbol{X}, \boldsymbol{U})^{\gamma+1}} \right)^{\frac{\gamma}{\gamma+1}} \right] \right]$$

$$- \frac{1}{\gamma} \log(1 - \epsilon) + O\left(\frac{\epsilon}{1 - \epsilon} \nu_1 \right)$$

$$= J_\gamma(\varphi) - \frac{1}{\gamma} \log(1 - \epsilon) + O\left(\frac{\epsilon}{1 - \epsilon} \nu_1 \right).$$

This completes the proof.

H.2 Proof of Lemma 12

We recall that ν_1 is defined by

$$\nu_1 = \underbrace{\iint p(\boldsymbol{x})\delta(\boldsymbol{u}|\boldsymbol{x}) \left(\frac{e^{(\gamma+1)r(\boldsymbol{x}, \boldsymbol{u})}}{1 + e^{(\gamma+1)r(\boldsymbol{x}, \boldsymbol{u})}} \right)^{\frac{\gamma}{\gamma+1}} \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u}}_{(A)} + \underbrace{\iint p(\boldsymbol{x})\delta(\boldsymbol{u}) \left(\frac{1}{1 + e^{(\gamma+1)r(\boldsymbol{x}, \boldsymbol{u})}} \right)^{\frac{\gamma}{\gamma+1}} \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u}}_{(B)}.$$

Here, we complete the proof by showing below that two terms (A) and (B) on the right-hand side are sufficiently small under Assumptions (D1-2).

Term (A): We apply the Cauchy–Schwartz inequality to Term (A) as

$$\iint \delta(\boldsymbol{u}|\boldsymbol{x})p(\boldsymbol{x}) \left(\frac{e^{(\gamma+1)r(\boldsymbol{x}, \boldsymbol{u})}}{1 + e^{(\gamma+1)r(\boldsymbol{x}, \boldsymbol{u})}} \right)^{\frac{\gamma}{\gamma+1}} \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u}$$

$$\leq C_A \left(\iint \delta(\boldsymbol{u}|\boldsymbol{x})p(\boldsymbol{x})e^{\gamma\{\psi(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}), \boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u})) + a(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}))\}} \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u} \right)^{\frac{1}{2}}, \tag{89}$$

where

$$C_A := \left(\iint \delta(\boldsymbol{u}|\boldsymbol{x})p(\boldsymbol{x}) \left(\frac{e^{(\gamma+1)\{\psi(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}), \boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u})) + a(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}))\}/2}}{e^{-(\gamma+1)b(\boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u}))} + e^{(\gamma+1)\{\psi(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}), \boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u})) + a(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}))\}}} \right)^{\frac{2\gamma}{\gamma+1}} \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u} \right)^{\frac{1}{2}}.$$

By Assumption (D1), the constant C_A is finite. Thus, Assumption (D2) ensures that Term (A) is sufficiently small.

Term (B): Applying the Cauchy–Schwartz inequality yields the upper bound of Term (B) as

$$\iint \delta(\boldsymbol{u})p(\boldsymbol{x}) \left(\frac{1}{1 + e^{(\gamma+1)r(\boldsymbol{x}, \boldsymbol{u})}} \right)^{\frac{\gamma}{\gamma+1}} \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u} \leq C_B \left(\int \delta(\boldsymbol{u})e^{-\gamma b(\boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u}))} \mathrm{d}\boldsymbol{u} \right)^{\frac{1}{2}}, \tag{90}$$

where

$$C_B := \left(\iint \delta(\boldsymbol{u})p(\boldsymbol{x}) \left(\frac{e^{-(\gamma+1)b(\boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u}))/2}}{e^{-(\gamma+1)b(\boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u}))} + e^{(\gamma+1)\{\psi(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}), \boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u})) + a(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}))\}}} \right)^{\frac{2\gamma}{\gamma+1}} \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u} \right)^{\frac{1}{2}}.$$

By (90) and Assumption (D2), Term (B) is also sufficiently small.

Appendix I. Semistrong robustness for the contamination model 2

For the contamination model 2, the following proposition is established:

Proposition 13. *Let us define a constant ν_2 as*

$$\nu_2 := \iint \delta(\boldsymbol{x}, \boldsymbol{u}) \left(\frac{e^{(\gamma+1)r(\boldsymbol{x},\boldsymbol{u})}}{1 + e^{(\gamma+1)r(\boldsymbol{x},\boldsymbol{u})}} \right)^{\frac{\gamma}{\gamma+1}} d\boldsymbol{x} d\boldsymbol{u} + (1 - \epsilon) \iint p(\boldsymbol{x})\delta(\boldsymbol{u}) \left(\frac{1}{1 + e^{(\gamma+1)r(\boldsymbol{x},\boldsymbol{u})}} \right)^{\frac{\gamma}{\gamma+1}} d\boldsymbol{x} d\boldsymbol{u}$$

$$+ (1 - \epsilon) \iint \delta(\boldsymbol{x})p(\boldsymbol{u}) \left(\frac{1}{1 + e^{(\gamma+1)r(\boldsymbol{x},\boldsymbol{u})}} \right)^{\frac{\gamma}{\gamma+1}} d\boldsymbol{x} d\boldsymbol{u} + \epsilon \iint \delta(\boldsymbol{x})\delta(\boldsymbol{u}) \left(\frac{1}{1 + e^{(\gamma+1)r(\boldsymbol{x},\boldsymbol{u})}} \right)^{\frac{\gamma}{\gamma+1}} d\boldsymbol{x} d\boldsymbol{u}.$$

We make the following assumptions:

(D′1) Assume that

$$\iint \delta(\boldsymbol{x}, \boldsymbol{u}) \left(\frac{e^{(\gamma+1)\psi(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}),\boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u}))/2}}{e^{-(\gamma+1)\{a(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}))+b(\boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u}))\}} + e^{(\gamma+1)\psi(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}),\boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u}))}} \right)^{\frac{2\gamma}{\gamma+1}} d\boldsymbol{x} d\boldsymbol{u} < \infty,$$

$$\iint \{p(\boldsymbol{x})\delta(\boldsymbol{u}) + \delta(\boldsymbol{x})p(\boldsymbol{u}) + \delta(\boldsymbol{x})\delta(\boldsymbol{u})\}$$

$$\times \left(\frac{e^{-(\gamma+1)\{a(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}))+b(\boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u}))\}/2}}{e^{-(\gamma+1)\{a(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}))+b(\boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u}))\}} + e^{(\gamma+1)\psi(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}),\boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u}))}} \right)^{\frac{2\gamma}{\gamma+1}} d\boldsymbol{x} d\boldsymbol{u} < \infty$$

(D′2) The following integrals are sufficiently small.

$$\iint \delta(\boldsymbol{x}, \boldsymbol{u})e^{\gamma\psi(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}),\boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u}))} d\boldsymbol{u} d\boldsymbol{x}, \quad \int \delta(\boldsymbol{x})e^{-\gamma a(\boldsymbol{h}_{\mathrm{x}}(\boldsymbol{x}))} d\boldsymbol{u} d\boldsymbol{x} \quad and \quad \int \delta(\boldsymbol{u})e^{-\gamma b(\boldsymbol{h}_{\mathrm{u}}(\boldsymbol{u}))} d\boldsymbol{u}.$$

Then, it holds that ν_2 is sufficiently small and

$$\bar{J}_\gamma(r) = -\frac{1}{\gamma} \log \left[\exp\left(-\gamma J_\gamma(r)\right) - \epsilon I(r) \right] - \frac{1}{\gamma} \log(1 - \epsilon) + O\left(\frac{\epsilon}{1 - \epsilon}\nu_2 \right), \qquad (91)$$

where

$$I(r) := \iint p(\boldsymbol{x})p(\boldsymbol{u}) \left(\frac{1}{1 + e^{(\gamma+1)r(\boldsymbol{x},\boldsymbol{u})}} \right)^{\frac{\gamma}{\gamma+1}} d\boldsymbol{x} d\boldsymbol{u}.$$

The proof is essentially the same as Proposition 8, and thus is omitted. Proposition 13 indicates when ν_2 is sufficiently small, minimization of $\bar{J}_\gamma(r)$ is almost equal to minimization of

$$-\frac{1}{\gamma} \log \left[e^{-\gamma J_\gamma(r)} - \epsilon I(r) \right]. \qquad (92)$$

Unlike the contamination model 1 (Proposition 8), there exists an extra term $\epsilon I(r)$ inside the logarithm in (92). However, this extra term may not cause any bias in terms of representation learning because the minimizer of (92) is given by

$$\log \frac{p(\boldsymbol{x}, \boldsymbol{u})}{p(\boldsymbol{x})p(\boldsymbol{u})} - \log(1 - \epsilon). \qquad (93)$$

Since our goal of representation learning is to estimate the log-density ratio up to a constant, the minimizer (93) means that robust representation learning would be possible even for the contamination model 2 under the condition that ν_2 is sufficiently small. In addition, again, the contamination ratio ϵ is not necessarily assumed to be small.

Assumptions (D$'$1-2) have the same implications as Assumptions (D1-2) in Proposition 8 (See Section 4.3.2), and thus would reflect the typical contamination of outliers. Thus, the condition that ν_2 is sufficiently small would be fairly reasonable. Indeed, Section 5 experimentally demonstrates that our method for representation learning is very robust against outliers.

References

L. B. Almeida. MISEP—linear and nonlinear ICA based on mutual information. *Journal of Machine Learning Research*, 4:1297–1318, 2003.

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

G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 1247–1255. PMLR, 2013.

R. Arandjelovic and A. Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, pages 609–617, 2017.

J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

P. Bachman, R. D. Hjelm, and W. Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 15535–15545, 2019.

A. Banerjee. On Bayesian bounds. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 81–88, 2006.

H. B. Barlow. Possible principles underlying the transformations of sensory messages. *Sensory Communicatio*, 1961.

M. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, D. Hjelm, and A. Courville. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80, pages 530–539, 2018.

A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

T. Blaschke, T. Zito, and L. Wiskott. Independent slow feature analysis and nonlinear blind source separation. *Neural Computation*, 19(4):994–1021, 2007.

D. A. Blythe, P. Von Bunau, F. C. Meinecke, and K.-R. Müller. Feature extraction for change-point detection using stationary subspace analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 23(4):631–643, 2012.

M. Caron, P. Bojanowski, A. Joulin, and M. Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018.

P. Chen, H. Hung, O. Komori, S.-Y. Huang, and S. Eguchi. Robust independent component analysis via minimum γ-divergence estimation. *IEEE Journal of Selected Topics in Signal Processing*, 7 (4):614–624, 2013.

X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2172–2180, 2016.

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.

R. D. Cook. *Regression Graphics: Ideas for Studying Regressions Through Graphics*. John Wiley & Sons, 1998.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional trans-formers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.

C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1422–1430, 2015.

G. Dornhege, J. d. R. Millán, T. Hinterberger, D. McFarland, and K.-R. Müller. *Toward brain-computer interfacing*. MIT press Cambridge, 2007.

A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, and T. Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 766–774, 2014.

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

K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

I. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio. Maxout networks. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, Proceedings of Machine Learning Research, pages 1319–1327. PMLR, 2013.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2672–2680, 2014.

L. Gresele, P. Rubenstein, A. Mehrjou, F. Locatello, and B. Schölkopf. The incomplete Rosetta Stone problem: Identifiability results for multi-view nonlinear ICA. In *Proceedings of the 35th International Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 296–313, 2019.

M. Gutmann and A. Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13:307–361, 2012.

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.

D. A. Harville. *Matrix Algebra From a Statistician's Perspective*. Springer Science & Business Media, 2006.

I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR)*, 2017.

R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P.Bachman, A. Trischler, and Y. Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations (ICLR)*, 2019.

K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2): 251–257, 1991.

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. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.

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 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.

L. Jing and Y. Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4037–4058, 2020.

T. Kanamori, T. Suzuki, and M. Sugiyama. Theoretical analysis of density ratio estimation. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 93(4): 787–798, 2010.

I. Khemakhem, D. P. Kingma, R. P. Monti, and A. Hyvärinen. Variational autoencoders and non-linear ICA: A unifying framework. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108, pages 2207–2217. PMLR, 2020.

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.

D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

H.-Y. Lee, J.-B. Huang, M. Singh, and M.-H. Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 667–676, Oct 2017.

K. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

Y. Li, M. Yang, and Z. Zhang. A survey of multi-view representation learning. *IEEE Transactions on Knowledge and Data Engineering*, 31(10):1863–1883, 2018.

R. Linsker. An application of the principle of maximum information preservation to linear systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 186–194, 1989.

X. Liu, F. Zhang, Z. Hou, Z. Wang, L. Mian, J. Zhang, and J. Tang. Self-supervised learning: Generative or contrastive. *arXiv preprint arXiv:2006.08218*, 2020.

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.

D. McAllester and K. Stratos. Formal limitations on the measurement of mutual information. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pages 875–884, 2020.

I. Misra, C. L. Zitnick, and M. Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 527–544. Springer, 2016.

R. P. Monti, K. Zhang, and A. Hyvärinen. Causal discovery with general non-linear relationships using non-linear ICA. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 186–195, 2020.

X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1089–1096, 2008.

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.

M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2227–2237, 2018.

R. A. Poldrack. The future of fMRI in cognitive neuroscience. *Neuroimage*, 62(2):1216–1220, 2012.

B. Poole, S. Ozair, A. van den Oord, A. Alemi, and G. Tucker. On variational bounds of mutual information. In *Proceedings of the 36th International conference on Machine learning (ICML)*, pages 5171–5180, 2019.

R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng. Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, pages 759–766, 2007.

A. Ruderman, M. D. Reid, D. García-García, and J. Petterson. Tighter variational representations of f-divergences via restriction to probability measures. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 1155–1162, 2012.

H. Sasaki, T. Takenouchi, R. Monti, and A. Hyvärinen. Robust contrastive learning and nonlinear ICA in the presence of outliers. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124, pages 659–668. PMLR, 2020.

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.

M. Sugiyama, S. Nakajima, H. Kashima, P. V. Buenau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1433–1440, 2008.

M. Sugiyama, T. Suzuki, and T. Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012.

C. Sun, F. Baradel, K. Murphy, and C. Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019.

Y. Tan, J. Wang, and J. M. Zurada. Nonlinear blind source separation using a radial basis function network. *IEEE Transactions on Neural Networks*, 12(1):124–134, 2001.

T. Teshima, I. Sato, and M. Sugiyama. Few-shot domain adaptation by causal mechanism transfer. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119, pages 9458–9469, 2020.

Y. Tian, D. Krishnan, and P. Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, and M. Lucic. On mutual information maximization for representation learning. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, 2019.

Y. Tsuboi, H. Kashima, S. Hido, S. Bickel, and M. Sugiyama. Direct density ratio estimation for large-scale covariate shift adaptation. *Journal of Information Processing*, 17:138–155, 2009.

H.-Y. Tung, H.-W. Tung, E. Yumer, and K. Fragkiadaki. Self-supervised learning of motion capture. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 5236–5246, 2017.

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

P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11:3371–3408, 2010.

X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2794–2802, 2015.

Z. Wang, E. Simoncelli, and A. Bovik. Multiscale structural similarity for image quality assessment. In *Proceedings of the 37th Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402, 2003.

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

P. Wu and K. Fukumizu. Causal mosaic: Cause-effect inference via nonlinear ICA and ensemble method. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108, pages 1157–1167, 2020.