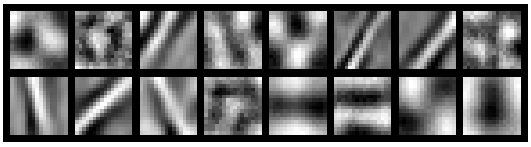

(a)

| Methods | p=2, d=4 | | p=3, d=6 | | p=4, d=8 | | p=5, d=10 | |
|---|---|---|---|---|---|---|---|---|
| | std. | p-value | std. | p-value | std. | p-value | std. | p-value |
| RICA | 3.72e-03 | 2.47e-10 | 3.73e-03 | 1.58e-07 | 2.95e-03 | 4.54e-07 | 2.98e-03 | 6.33e-05 |
| MFICA_Gauss | 3.04e-02 | 1.67e-05 | 1.28e-02 | 3.72e-07 | 1.05e-02 | 1.03e-09 | 1.27e-02 | 2.44e-07 |
| MFICA_MoG | 2.08e-03 | 3.01e-11 | 1.07e-02 | 1.14e-03 | 2.08e-02 | 7.95e-03 | 3.97e-03 | 1.74e-05 |
| NG-EM | 1.84e-02 | 4.71e-05 | 1.27e-02 | 1.21e-05 | 7.08e-03 | 2.18e-07 | 7.26e-03 | 2.11e-08 |
| SMICA | 5.83e-02 | 1.59e-01 | 4.12e-03 | 1.30e-03 | 6.01e-02 | 1.97e-02 | 6.58e-03 | 1.07e-05 |
| LFOICA | 1.88e-03 | - | 1.14e-03 | - | 9.07e-04 | - | 1.95e-03 | - |

(b)

Figure 1: (a) Some of the features extracted by LFOICA. (b) std. of the results of each method and the p-value of Welch's $t$-test between each baseline method and our LFOICA.

We would like to thank all reviewers for the constructive comments. We will fully address all the review concerns in the revision. The citation indices below are consistent with those in the main paper.

**To R#1: MoG and sparsity have been studied.** It is true that MoG and sparsity have been extensively studied in existing ICA methods. Aside from the two references suggested by the reviewer, we also cited several representative works in this direction ([14,15,16]). In fact, one major motivation of the proposed LFOICA is that the classical MoG-based ICA methods use maximum likelihoood to estimate the parameters, which involve computationally intractable posterior inference. In our practical considerations, we disscussed in the paper that LFOICA can be further specified with additional constraints, such as MoG and sparsity, in a likelihood-free way that enjoys both statistical and computational efficiency. We will further discuss the suggested references in the "practical consideration" subsection.

**Identifiability and convergence of our model.** The identifiability of the mixing matrix $\mathbf{A}$ in our model ($\mathbf{x} = G_{\mathbf{A},\boldsymbol{\theta}}(\mathbf{z}) = \mathbf{A}[f_{\theta_1}(z_1), \ldots, f_{\theta_d}(z_d)]^\intercal$) follows the identifiability results for OICA [29], which is summarized in the following Theorem. **Theorem 1** Given two OICA models $\mathbf{x} = G_{\mathbf{A},\boldsymbol{\theta}}(\mathbf{z})$ and $\mathbf{x}' = G_{\mathbf{A}',\boldsymbol{\theta}'}(\mathbf{z}')$ that specify distributions $\mathbb{P}(\mathbf{x})$ and $\mathbb{P}(\mathbf{x}')$, respectively. Under the condition that none of the distributions of $f_i(\mathbf{z}_i)$ is Gaussian, if $MMD(\mathbb{P}(\mathbf{x}), \mathbb{P}(\mathbf{x}')) = 0$, then $\mathbf{A}' = \mathbf{A}\mathbf{P}_p\mathbf{S}_p$, where $\mathbf{P}_p$ is a $p \times p$ column permutation matrix and $\mathbf{S}_p$ is a $p \times p$ scaling matrix. The proof is almost the same as that of Theorem 3 in [29], except that in order to guarantee $\mathbb{P}(\mathbf{x}) = \mathbb{P}(\mathbf{x}')$, we use $MMD = 0$ while [29] uses maximum likelihood (KL divergence). Given the identifiability results, the estimated mixing matrix converges to the scaled and permuted version of the true mixing matrix and so does the source distributions. With additional constraints on $\mathbf{A}$, the permutation and scaling ambiguity can be further reduced. The proofs to this are provided in the original papers of the two causal discovery problems evaluated in our paper. **The parameters in our MLPs (*i.e.*,$\theta$) are not identifiable ($\theta \neq \theta'$), but we do not need the identifiability of $\theta$ to perform certain downstream tasks, such as the two causal discovery tasks mentioned in our paper.**

**GAN-based methods to solve nonlinear ICA [21].** Although both our work and [21] use a GAN style approach to solve ICA, they are largely different to each other. **First**, the main purpose of [21] is to recover the ICs instead of how the ICs are mixed (*i.e.*,the mixing matrix). It models the mixing and unmixing procedure implicitly with an encoder-decoder architecture. As a consequence of non-linearity, there is no guarantee for identifiability. In contrast, we concentrate on the mixing matrix estimation for causal discovery purpose. **Second**, the encoder-decoder architecture in [21] cannot be easily extended for OICA because the posterior of ICs cannot be modeled by a deterministic encoder. **Third**, the adversarial training target of LFOICA and [21] are different. While [21] aims at matching the joint distribution and product of marginal distribution of the recovered ICs (this is also how [21] makes the components independent), LFOICA is trained to match the distributions of the generated mixtures and true mixtures. And the estimated ICs by LFOICA are naturally independent because they are generated from independent latent noises with separate networks.

**To R#2: Describe the details of standard overcomplete ICA.** Thanks for the suggestion. We will add an algorithmic description that summarizes maximum likelihood (the EM-type) algorithms for OICA in the Supplementary Material.

**To R#3: The generative model does not have an inference algorithm.** Since the main focus of our work is OICA and it's applications to causal discovery, we put significant effort on mixing matrix estimation instead of latent variable inference. Nevertheless, this is a valuable suggestion and we'll consider extending our LFOICA.

**Comparison to score-matching ICA**. We conduct additional experiments with the suggested score matching ICA (SMICA) under the same conditions. For each experiment setting in Table 1 in the paper, MSE of the recovered mixing matrix obtained by SMICA are 3.52e-2, 6.70e-2, 8.11e-3 and 1.94e-2 respectively, which is not as good as LFOICA.

**Scalability to higher-dimensional data such as images.** LFOICA is originally designed for causal discovery, which is different from image feature extraction. To demonstrate the generality of the method to higher dimensional problems, we apply it to extract features from image patches. The features form a dictionary which can be used with the recovered ICs for image restoration tasks, such as image denoising. Example features are shown in the Figure 1(a). Patch size is 16x16 and the number of atoms is set to 400. We use DCT dictionary for initialization.

**"empirical estimator".** It refers to Maximum Mean Discrepancy (MMD), which is mentioned in line 83.

**"s.e.m.s? How significant?".** Here we interpret "s.e.m.s" as standard deviation of the MSE values in Table 1 of our paper and report it as well as the p-values by conduct Welch's $t$-test between baseline works and our LFOICA in Figure 1(b) (this figure corresponds to Table 1 in our paper). As one can see, the differences are significant enough.