DEEP DETERMINISTIC INDEPENDENT COMPONENT ANALYSIS FOR HYPERSPECTRAL UNMIXING

Hongming Li[1], *Shujian Yu*[*2], *José C. Príncipe*[1]

[1]Computational NeuroEngineering Laboratory, University of Florida, Gainesville, FL 32611, USA
[2]Machine Learning Group, UiT - The Arctic University of Norway, 9037 Tromsø, Norway

1. SUPPLEMENTARY MATERIAL

1.1. Structure of Deep ICA Approaches

To the best of our knowledge, there are only two recent proposals on deep ICA that address this issue. Both of them use parametric approaches inspired by adversarial learning.

In [1], the authors use an adversarial autoencoder to minimize JS-divergence (a substitute of MI) and reconstruction error simultaneously. The detailed structure is shown in Fig. 1, where R denotes reconstruction error, JS denotes JS-divergence. There are three deep neural modules, i.e., encoder (unmixing function), decoder and discriminator. The autoencoder is updated by minimizing $JS + \lambda R$, where λ is a hyper-parameter, while the discriminator is updated by maximizing JS.

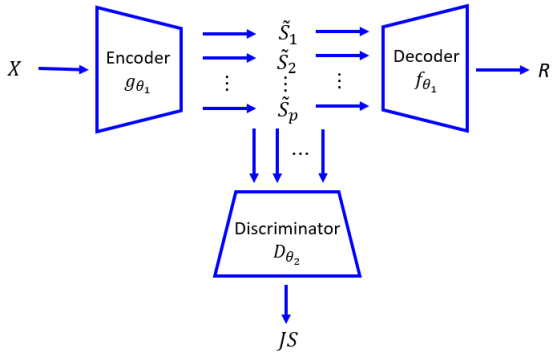


Fig. 1. Architecture of the adversarial autoencoder ICA. The autoencoder minimizes the JS while the discriminator maximizes it.

Mutual information neural estimator (MINE) [2] is a more straightforward approach which estimates MI directly. As demonstrated in Fig. 2, where MI denotes mutual information, the MINE architecture has two neural modules. The objective function of MINE is a lower bound of MI. The encoder learns to minimize this lower bound while estimator maximizes it. Our empirical study use MINE as a baseline representing neural ICA approach. We use the codes of [3] which apply MINE on ICA directly.

*Contact author: yusj9011@gmail.com.

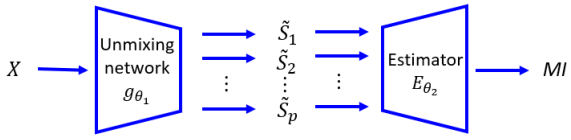


Fig. 2. Architecture of the MINE-based ICA. The encoder (Unmixing function) minimizes the MI while the estimator maximizes it.

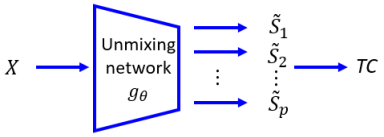


Fig. 3. Architecture of our proposed deep deterministic independent component analysis (DDICA). We directly minimize the total dependence of all predicted sources $\{\tilde{S}_i\}|_{i=1}^{p}$ with the objective $\arg\min_{\theta} \mathrm{TC}\left(\tilde{S}_1, \cdots, \tilde{S}_p\right)$ by the matrix-based Rényi's α-order total correlation (TC) [4].

Unlike parametric independence estimation approaches, we propose to exploit a matrix-based non-parametric Renyi's entropy estimator [5, 6, 4] for MI estimation. We utilize TC, a variant of MI, as our objective function and a single deep neural network as our unmixing function. As shown in Fig. 1 2 and 3, our architecture has only one neural module which promises much less parameters.

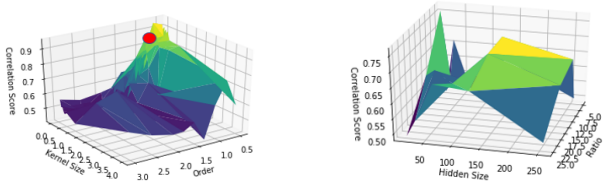
Besides, adversarial learning is hard to train because it introduce more hyper-parameters (learning rate, number of hidden units and layers of discriminator, etc.). In contrast, there are only two hyper-parameters in our mutual information estimation, i.e., order of Rényi's entropy, kernel size. Hence our framework is easier to train compared to variational approaches.

1.2. Less Hyper-parameters, Easier Training

In this section, we conduct experiments to show the simplicity of hyper-parameters selection of our architecture and compares it with the MINE-based ICA.

We validate it using the PNL data shown in the paper. We train our model with different order of Rényi's entropy α, kernel size σ and plot the 3D figure of it (See the left panel of Fig. 4). The z axis is the absolute correlation coefficient $|\rho|$. Normally, we can first set the order to be 1.01 to approximate Shannon entropy and apply Sliverman rule [7] for kernel size, that is $N^{-1/(4+p)}$, where p is the dimension of the latent vector, N can be considered as the batch size. We set the batch size to be 2000, and the σ based on Sliverman rule should be 0.2187. The point $(1.01, 0.2187)$ is demonstrated using a red point in the the left panel of Fig. 4, which is very close to the global optimal of hyper-parameter selection $(0.75, 1.584)$. Therefore, we claim that there is a simple hyper-parameters searching policy for our architecture. We can start from 1.01 and Sliverman rule, and then search a small space near the start point to obtain best performance.

On the contrary, neural estimation has more hyper-parameters mainly including learning rate, hidden layers, hidden units in each layer, corporation of two neural networks. In our experiment, we fix the learning rate to be 0.0001, the number of hidden layers to be 2 but change the hidden units in each layer and updating ratio to plot the 3D figure. The updating ratio indicates how many times we update discriminator after we update the encoder for a single time. As demonstrated in the right panel of Fig. 4, there are multiple peaks in the hyper-parameter space, and we don't have a start point policy for our searching. It indicates that we have to search a large space for best performance.



(a) DDICA (b) MINE

Fig. 4. Hyper parameter space of DDICA and MINE. The red point indicates the 1.01 for order and Silverman rule for kernel size

In summary, the major advantages of DDICA compared to variational approaches are twofold:

- We have much less parameters in our architecture.
- We have much less hyper-parameters in our mutual information estimator, and a simple searching policy make it easy find the optimal of hyper-parameters selection.

2. ACKNOWLEDGMENT

This work was funded in part by the U.S. ONR under grants N00014-18-1-2306, N00014-21-1-2324, N00014-21-1-2295, the DARPA under grant FA9453-18-1-0039, and the Research Council of Norway (RCN) under grant 309439.

3. REFERENCES

[1] Philemon Brakel and Yoshua Bengio, "Learning independent features with adversarial nets for non-linear ica," *arXiv preprint arXiv:1710.05050*, 2017.

[2] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm, "Mutual information neural estimation," in *ICML*. PMLR, 2018, pp. 531–540.

[3] Hlynur Daví Hlynsson and Laurenz Wiskott, "Learning gradient-based ica by neurally estimating mutual information," in *KI*. Springer, 2019, pp. 182–187.

[4] Shujian Yu, Luis Gonzalo Sanchez Giraldo, Robert Jenssen, and Jose C. Principe, "Multivariate extension of matrix-based rényi's α-order entropy functional," *IEEE PAMI*, vol. 42, no. 11, pp. 2960–2966, 2020.

[5] Alfréd Rényi, "On measures of entropy and information," in *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1961, vol. 1, pp. 547–561.

[6] Luis Gonzalo Sanchez Giraldo, Murali Rao, and Jose C. Principe, "Measures of entropy from data using infinitely divisible Kernels," *IEEE Trans. Inf. Theory*, vol. 61, no. 1, pp. 535–548, 2015.

[7] B. W. Silverman, *Density estimation: For statistics and data analysis*, Routledge, 2018.