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

ABSTRACT

We develop a new neural network based independent component analysis (ICA) method by directly minimizing the dependence amongst all extracted components. Using the matrix-based Rényi's α-order entropy functional, our network can be directly optimized by stochastic gradient descent (SGD), without any variational approximation or adversarial training. As a solid application, we evaluate our ICA in the problem of hyperspectral unmixing (HU) and refute a statement that "*ICA does not play a role in unmixing hyperspectral data*", which was initially suggested by [1]. Code and additional remarks of our DDICA is available at `https://github.com/hongmingli1995/DDICA`.

Index Terms— Independent component analysis, deep neural networks, matrix-based Rényi's α-order entropy functional, hyperspectral unmixing

1. INTRODUCTION

Independent component analysis (ICA) is a statistical approach for extracting hidden factors (or sources) from a mixed signal. It has been widely applied in areas including medical image processing [2], biological assays [3], etc. Existing ICA methods can be roughly divided into two categories by their objective functions. The first type of approaches attempt to maximize the non-Gaussianity of predicted components, since signal mixtures tend to be Gaussian distributed due to the central limit theorem [4]. By contrast, the second type of approaches directly minimize the dependence or mutual information between different components, following the motivation that mutual information value reduces to zero if and only if all components are independent.

Hyperspectual unmixing (HU) is a common problem in remote sensing and medical hyperspectral image analysis [5]. It can be thought of as an undercomplete ICA problem where the number of mixtures is larger than sources. However, ICA-based approaches are not popular for HU mainly for two criticisms: (**C1**) the sources are sometimes mixed in a nonlinear manner, whereas most existing ICA methods assume either a
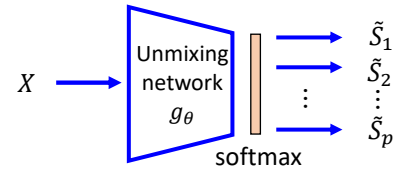
---
*Contact author: yusj9011@gmail.com.

Fig. 1. 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 $\arg\min_\theta \text{TC}\left(\tilde{S}_1, \cdots, \tilde{S}_p\right)$ by the matrix-based Rényi's α-order total correlation (TC) [9].

linear mixing matrix (e.g., [6]) or the equivalence between the number of sources and the number of observations [7, 8]; and (**C2**) the HU itself is not a strict ICA problem, because the sum of abundances are required to be 1. Due to these limitations, [1] claims that "*ICA does not play a competitive role in unmixing hyperspectral data*".

In this paper, we challenge this point of view by developing a new deep neural network (DNN) based ICA with a softmax activation before the output and trained it with advanced information theoretic criteria. We argue that, for **C1**, the nonlinear activation functions, such as ReLU, in DNNs naturally introduce nonlinearity in a hierarchical manner. On the other hand, for **C2**, it can be simply addressed by introducing a softmax layer before the network output.

We term our method the deep deterministic ICA (DDICA), as shown in Fig. 1. DDICA uses a single feed-forward neural network as the unmixing function and estimate the total mutual information non-parametrically with the recently proposed matrix-based Rényi's α-order total correlation [10, 9]. DDICA can be trained directly with stochastic gradient descident (SGD) and its variants, without the necessity of variational approximation or auxiliary discriminator module.

Our main contributions are twofold:

- Mutual information is hard to estimate. Unlike recent proposals that rely on an auxiliary discriminator network which is hard to control in practice, our DDICA can be optimized directly with SGD or its variants.

- Practically, we show that DDICA with a softmax layer is suitable for unmixing hyperspectral data. We show that our DDICA outperforms other traditional ICA and deep ICA methods. Our performance on HU also matches well with that achieved by state-of-the-art non-negative matrix factorization (NMF)-based approaches. This result gives new insight to the future direction of HU.

2. BACKGROUND

2.1. Traditional ICA

ICA assumes the observed random vector $\mathbf{x} = [x_1, x_2, \cdots, x_d] \in \mathbb{R}^d$ is generated by p independent latent variables (also called the independent components) $\mathbf{s} = [s_1, s_2, \cdots, s_p]$ as:

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

in which f is a linear or nonlinear mixing function. The goal of ICA is then to recover the inverse function f^{-1} (characterized by an unmixing function g) as well as the independent components $[s_1, s_2, \cdots, s_p]$ solely based on observations $\mathbf{x}$:

$$\tilde{\mathbf{s}} = g(\mathbf{x}). \tag{2}$$

The identifiability cannot be guaranteed in general nonlinear ICA problem, as there is an infinite number of solutions if the space of mixing functions is unconstrained [11]. However, some specific types of non-linear mixtures like post nonlinear (PNL) ICA [12] is solvable. In this work, we consider the case of $d > p$, which is also called the undercomplete ICA. Hyperspectral unmixing can be thought of as an undercomplete ICA problem [13], as the number of mixtures d is significantly larger than that of endmembers (i.e., sources) p.

Let the predicted sources be $[\tilde{s}_1, \tilde{s}_2, \cdots, \tilde{s}_p]$, the underlying idea of ICA can be divided into two categories. The first type is maximizing the non-Gaussianity of predicted sources, as signal mixtures tend to be Gaussian [14]. FastICA [6] is one of the most well-known approaches in this category. The second type is minimizing mutual information value because it reduces to zero if two random variables are independent. Notable examples in this category include Infomax [15]. Although the original Infomax assumes linearity, it can be extended to the non-linear scenario using either neural networks [7] or kernel tricks [8]. However, the number of sources is required to be the same to that of mixtures in [7, 8]. Therefore they both cannot be applied to HU directly.

2.2. Neural ICA using parametric estimator

DNN is a competitive functional space to represent the unmixing function g. To perform deep ICA, a differentiable mutual information (or dependence) estimator is desirable. Unfortunately, many classic mutual information estimators, e.g., kNN [16], do not satisfy this requirement. 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 [17], the authors use an adversarial autoencoder to minimize JS-divergence (a substitute of MI) and reconstruction error simultaneously. Later, [18] applies the mutual information neural estimator (MINE) [19] on ICA directly. MINE is a straightforward approach that estimates a lower bound of MI by an auxiliary neural network. However, these parametric approaches introduce more hyperparameters (e.g., learning rate, number of hidden units, etc.), and their adversarial architectures, similar to GAN, are well-known for the difficulty of training and replication. These drawbacks motivate us to develop a new neural network based ICA with non-parametric mutual information estimators that is easy to optimize. Additional remarks about architecture differences between our ICA and [17, 18] are further illustrated in supplementary material at https://github.com/hongmingli1995/DDICA.

3. METHODOLOGY

3.1. Deep Deterministic ICA

In this work, we use a DNN g_θ, parameterized by θ, as the nonlinear unmixing function. Our goal is to minimize the total dependence amongst all predicted sources $[\tilde{s}_1, \tilde{s}_2, \cdots, \tilde{s}_p]$.

One popular way to define the total dependence for $[\tilde{s}_1, \tilde{s}_2, \cdots, \tilde{s}_p]$ can be expressed as the Kullback–Leibler (KL) divergence from the joint distribution $P(\tilde{s}_1, \tilde{s}_2, \cdots, \tilde{s}_p)$ to the product of marginal distributions $P(\tilde{s}_1)P(\tilde{s}_2)\cdots P(\tilde{s}_p)$:

$$\begin{aligned} \texttt{TC} &= D_{KL}\left[P(\tilde{s}_1, \tilde{s}_2, \cdots, \tilde{s}_p); P(\tilde{s}_1)P(\tilde{s}_2)\cdots P(\tilde{s}_p)\right], \\ &= \left[\sum_{i=1}^{p} H(\tilde{s}_i)\right] - H(\tilde{s}_1, \tilde{s}_2, \cdots, \tilde{s}_p), \end{aligned} \tag{3}$$

in which $H(\tilde{s}_i)$ is the entropy of the i-th source, $H(\tilde{s}_1, \tilde{s}_2, \cdots, \tilde{s}_p)$ is the joint entropy for $[\tilde{s}_1, \tilde{s}_2, \cdots, \tilde{s}_p]$. Eq. (3) is also called the total correlation (TC) [20], and has been recently applied for disentangled representation learning [21] and the understanding to the dynamics of learning of DNNs [22].

An alternative way to define the total dependence is by the following formula:

$$\texttt{D} = \sum_{i=1}^{p} I(\tilde{s}_i; \tilde{\mathbf{s}}_{-i}), \tag{4}$$

in which I denotes mutual information and $\tilde{\mathbf{s}}_{-i}$ refers to the set of all sources except the i-th source.

Eq. (4) has been used in [18] to train a ICA network, and the (lower bound) of $I(\tilde{s}_i; \tilde{\mathbf{s}}_{-i})$ can be evaluated by MINE. However, the poor assumptions of MINE (requires absolute continuity) and its parametric nature leads to training difficulty. Similar to GAN, numerous "hyper-parameters", e.g.,

structure of estimator neural networks, learning rates, etc, need to be selected much carefully in the implementation.

Unlike [18] that requires the estimation of p mutual information values in high-dimensional space, we directly optimize Eq. (3) for simplicity. Moreover, instead of using MINE or adversarial training strategy which is hard to control in practice, we compute entropy and joint entropy terms in Eq. (3) directly from data by the matrix-based Rényi's α-order entropy functional, without any variational approximation or distributional assumption.

Suppose there are N samples[1] for the i-th predicted source, i.e., $\tilde{s}_i = [\tilde{s}_i^1, \tilde{s}_i^2, \cdots, \tilde{s}_i^N]$ where the subscript denotes view index and superscript denotes sample index. A Gram matrix $K_i \in \mathbb{R}^{N \times N}$ can be obtained by computing $K_i(n, m) = \kappa(\tilde{s}_i^n, \tilde{s}_i^m)$, where κ is a infinite divisible kernel [23] which is usually assumed to be Gaussian. The entropy of $\tilde{s}_i$ can be expressed by [10]:

$$H_\alpha(\tilde{s}_i) = H_\alpha(A_i) = \frac{1}{1-\alpha} \log_2 [\sum_{n=1}^{N} \lambda_n(A_i)^\alpha], \quad (5)$$

where $A_i = K_i / \operatorname{trace}(K_i)$ is the normalized Gram matrix and $\lambda_n(A_i)$ denotes n-th eigenvalue of A_i.

Further, the joint entropy for $\{\tilde{s}_i\}_{i=1}^p$ is defined as [9]:

$$H(\tilde{s}_1, \tilde{s}_2, \cdots, \tilde{s}_p) = H_\alpha \left(\frac{A_1 \circ A_2 \cdots \circ A_p}{\operatorname{trace}(A_1 \circ A_2 \cdots \circ A_p)} \right). \quad (6)$$

The `TC` in Eq. (3) can be directly obtained with Eqs. (5) and (6). The differentiability of this matrix-based Renyi's entropy estimator has been proved and empirically validated in [22]. In practice, the automatic SVD is embedded in mainstream deep learning APIs including TensorFlow and PyTorch. In this sense, our DDICA enjoys a simple and tractable objective `min TC`, which can be directly trained with SGD or its variants.

3.2. Application in Hyperspectral Unmixing (HU)

HU can be considered as an undercomplete ICA problem where $d \gg p$. Mixtures $\mathbf{x}$ are recorded by a camera with d different spectral channels, and our aim is to separate them into p classes. The DDICA achieves this goal by simply changing the number of output units. However, deep learning methods are usually sensitive to weight initialization [24]. Unfortunately, such dependence to weight initialization is also observed in our experiments on HU with both MINE and our DDICA, i.e., the predicted sources of hyperspectral images are sometimes partially correct. To alleviate this problem, we train g_θ with different weight initialization independently by T times. These neural networks generate $p \times T$ predicted sources. We finally cluster the predicted sources using k means ($k = p$) and view the clustering centers as predicted sources. In practice, one can also select k slightly larger than p.

---

[1]One can understand N as the mini-batch size in network training.

4. EXPERIMENT

In this section, we conduct experiments on both synthetic and real-world hyperspectral data to demonstrate the advantages of our DDICA. We compare DDICA with a set of popular references. **FastICA** [6] is the most well-known ICA method based on maximizing non-Gaussianity by kurtosis. **InfomaxICA** [15] follows the idea of minimizing mutual information. It maximizes joint entropy of predicted components as an alternative. ICA based on mutual information neural estimator (**MINE**) [19, 18] estimates a lower bound of mutual information values by an auxiliary network. **MISEP** [7] is a post non-linear (PNL) extension of InfomaxICA but cannot be directly applied to undercomplete ICA, i.e., HU. Additionally, we also compare the performance of DDICA with that of **min-vol NMF** [25]. This is a recently proposed NMF-based approach and outperforms several state-of-the-art (SOTA) NMF-based algorithms in HU [26].

4.1. Validation on Synthetic PNL Dataset

We first test all competing ICA methods on actual 16kHz audio signals [27] but use synthetic PNL mixing function. We select two speech signals (FA01_01 and MA01_01) plus a uniform noise as our sources. Three 1D noisy sources are first mixed linearly by a 3×3 mixing matrix and then processed by $tanh(x)$, $(x + x^3)/2$, e^x, respectively. We set the batch size to be $2,000$, and results are normalized using Z-score.

The g_θ in our DDICA is a four-hidden-layer DNN with ReLU activation in the first three layers and linear activation in the last layer. It is followed by a differentiable whitening layer [28] to avoid trivial solutions. Notice that, we do not add a softmax layer here, because there is no constrains on the sum of magnitude of predicted components. The order α is 0.75 and the kernel width σ is 0.1584 according to Silverman's rule [29]. Experiments are repeated for 10 times and the average scores are recorded. As demonstrated in Fig. 2 and Table 1, our DDICA outperforms other linear (FastICA, Infomax) and non-linear ICAs (MISEP, MINE). Besides, compared with MINE, our network is much easier to optimize (see supplementary material for details).

Methods	FastICA	Infomax	MISEP	MINE	Ours
$|\rho|$	0.7132	0.6646	_0.9133_	0.7690	**0.9527**

Table 1. Separation performance in terms of absolute correlation coefficient $|\rho|$. The best performance is in bold and the second-best performance is underlined.

4.2. Results of Hyperspectral Unmixing

We select two real-world hyperspectral datasets: Samson and Urban. Each pixel of Samson is recorded by 156 spectral channels, i.e., $d = 156$, whereas Urban contains 162 channels. The number of sources p of Samson and Urban is 3
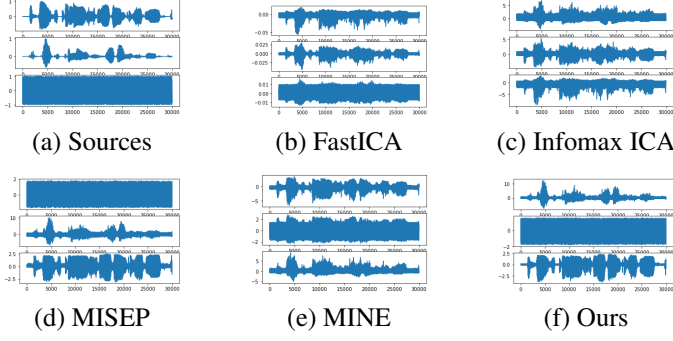
3880

(a) Sources (b) FastICA (c) Infomax ICA

(d) MISEP (e) MINE (f) Ours

Fig. 2. Predicted sources of different ICA methods.

(a) Ground Truth (b) FastICA

(c) Infomax ICA (d) MINE

(e) min-vol NMF (f) Ours (10 units)

Fig. 3. Abundance maps extracted by ICA methods on Urban.

Method	FastICA	Infomax	min-vol NMF	MINE	Ours
Soil	0.2942	0.4198	0.1945	0.3319	**0.1524**
Tree	0.4632	0.4426	0.2248	0.3828	**0.1948**
Water	0.4704	0.4913	**0.1528**	0.3911	0.2401
Average	0.4093	0.4512	**0.1907**	0.3686	0.1958

Table 2. Unmixing performance in terms of RMSE on Sansom dataset. The best performance is in bold and the second-best performance is underlined
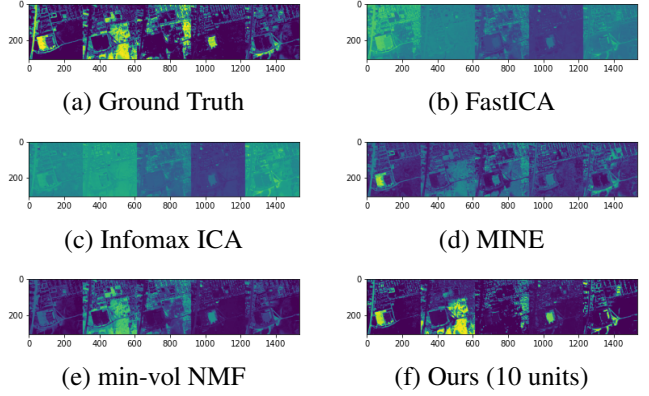
and 5, respectively. In hyperspectral unmixing, the first step is usually the determination of the number of sources using techniques like virtual dimensionality algorithm [30]. However, the aim of this work is to show the advantages of our ICA over other existing ICA methods. Therefore, we assume the true number of sources is known.

The network g_θ is a six-hidden-layer DNN with ReLU activation for the first five layers and softmax activation in the last layer. The α of our estimator is 1.01 to approximate Shannon entropy. The kernel width σ is fixed to be 0.1. We trained $1,000$ sets of parameters with different initialization to obtain $1,000 \times p$ unmixed abundances and then apply k-means. Clustering centers are treated as unmixing results.

Tables 2 and 3 compare the root mean square error (RMSE) between true and predicted sources (see 3). We

repeat the experiments 10 times and average the scores. From the table we observe that our method outperforms others ICA-based methods and achieves comparable results to the NMF-based one. Compared to Samson dataset, there are more classes and details in Urban datasets. Hence our method cannot achieve perfect results if number of output channels p equals to the true number of classes. To address this problem, we apply a strategy inspired by unsupervised neural classification algorithms [31]. We increase the number of outputs units of the neural networks to approximate an over-classification scenario. Then k means is applied to over-dimensioned unmixed maps. Fig. 3 (f) illustrates results with 10 units. We conduct the experiments on Urban dataset with five different number of output units: $[5, 6, 7, 8, 9, 10]$. The corresponding RMSE are $[0.2453, 0.2410, 0.2274, 0.2184, 0.2201, 0.2163]$, respectively. There is a negative correlation between RMSE and the number of output channels, and it is also reported in [1].

5. CONCLUSION

In this paper, we propose a novel deep neural network-based ICA without parametric estimation of mutual information. 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. We validate our model on PNL and hyper-spectral unmixing datasets. Our method outperforms other ICA-based approaches and achieves comparable results to the SOTA NMF-based ones.

Methods	FastICA	Infomax	min-vol NMF	MINE	Ours
Asphalt	0.2502	0.3203	0.1699	0.2995	**0.1548**
Grass	0.3133	0.3333	**0.1690**	0.3281	0.1860
Tree	0.3712	0.4177	**0.1264**	0.3404	0.1927
Roof	0.4299	0.4458	0.2573	0.3978	**0.2359**
Dirt	0.5163	0.4883	0.4075	0.3961	**0.3225**
Average	0.3762	0.4011	0.2228	0.3544	**0.2163**

Table 3. Unmixing performance in terms of RMSE on Urban dataset. The best performance is in bold and the second-best performance is underlined.

6. ACKNOWLEDGMENT

7. REFERENCES

[1] José MP Nascimento and Jose MB Dias, "Does independent component analysis play a role in unmixing hyperspectral data?," *IEEE TGRS*, vol. 43, no. 1, pp. 175–187, 2005.

[2] Vince D Calhoun, Jingyu Liu, and Tülay Adalı, "A review of group ica for fmri data and ica for joint inference of imaging, genetic, and erp data," *Neuroimage*, vol. 45, no. 1, pp. S163–S172, 2009.

[3] Rabia Aziz, CKa Verma, and Namita Srivastava, "A fuzzy based feature selection from independent component subspace for machine learning classification of microarray data," *Genomics data*, vol. 8, pp. 4–15, 2016.

[4] Seungjin Choi, Andrzej Cichocki, Hyung-Min Park, and Soo-Young Lee, "Blind source separation and independent component analysis: A review," *Neural Process. Lett*, vol. 6, no. 1, pp. 1–57, 2005.

[5] José M Bioucas-Dias, Antonio Plaza, Nicolas Dobigeon, Mario Parente, Qian Du, Paul Gader, and Jocelyn Chanussot, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE J Sel Top Appl Earth Obs Remote Sens*, vol. 5, no. 2, pp. 354–379, 2012.

[6] Aapo Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[7] Luís B Almeida, "Misep—linear and nonlinear ica based on mutual information," *J. Mach. Learn. Res.*, vol. 4, pp. 1297–1318, 2003.

[8] Francis R Bach and Michael I Jordan, "Kernel independent component analysis," *J. Mach. Learn. Res.*, vol. 3, no. Jul, pp. 1–48, 2002.

[9] 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.

[10] 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.

[11] Aapo Hyvärinen and Petteri Pajunen, "Nonlinear independent component analysis: Existence and uniqueness results," *Neural networks*, vol. 12, no. 3, pp. 429–439, 1999.

[12] Anisse Taleb and Christian Jutten, "Source separation in post-nonlinear mixtures," *IEEE TSP*, vol. 47, no. 10, pp. 2807–2820, 1999.

[13] Feiyun Zhu, "Hyperspectral unmixing: ground truth labeling, datasets, benchmark performances and survey," *arXiv preprint arXiv:1708.05125*, 2017.

[14] Rasmikanta Pati, Arun K Pujari, Padmavati Gahan, and Vikas Kumar, "Independent component analysis: A review, with emphasis on commonly used algorithms and contrast function," *Computación y Sistemas*, vol. 25, no. 1, 2021.

[15] Anthony J Bell and Terrence J Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, no. 6, pp. 1129–1159, 1995.

[16] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger, "Estimating mutual information," *Physical review E*, vol. 69, no. 6, pp. 066138, 2004.

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

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

[19] 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.

[20] Satosi Watanabe, "Information theoretical analysis of multivariate correlation," *IBM J. Res. Dev.*, vol. 4, no. 1, pp. 66–82, 1960.

[21] Ricky TQ Chen, Xuechen Li, Roger Grosse, and David Duvenaud, "Isolating sources of disentanglement in variational autoencoders," in *NeurIPS*, 2018.

[22] Shujian Yu, Francesco Alesiani, Xi Yu, Robert Jenssen, and Jose Principe, "Measuring dependence with matrix-based entropy functional," in *AAAI*, 2021, vol. 35, pp. 10781–10789.

[23] Rajendra Bhatia, "Infinitely divisible matrices," *Am. Math. Mon*, vol. 113, no. 3, pp. 221–235, 2006.

[24] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton, "On the importance of initialization and momentum in deep learning," in *ICML*. PMLR, 2013, pp. 1139–1147.

[25] Valentin Leplat, Andersen MS Ang, and Nicolas Gillis, "Minimum-volume rank-deficient nonnegative matrix factorizations," in *ICASSP*. IEEE, 2019, pp. 3402–3406.

[26] EMMB Ekanayake, Bhathiya Rathnayake, DYL Ranasinghe, S Herath, GMRI Godaliyadda, HMVR Herath, and MPB Ekanayake, "Constrained nonnegative matrix factorization for blind hyperspectral unmixing incorporating endmember independence," *arXiv preprint arXiv:2003.01041*, 2020.

[27] Peter Kabal, "Tsp speech database," *McGill University, Database Version*, vol. 1, no. 0, pp. 09–02, 2002.

[28] Merlin Schüler, Hlynur Daví Hlynsson, and Laurenz Wiskott, "Gradient-based training of slow feature analysis by differentiable approximate whitening," in *ACML*. PMLR, 2019, pp. 316–331.

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

[30] Chein-I Chang and Qian Du, "Estimation of number of spectrally distinct signal sources in hyperspectral imagery," *IEEE TGRS*, vol. 42, no. 3, pp. 608–619, 2004.

[31] Xu Ji, Joao F Henriques, and Andrea Vedaldi, "Invariant information clustering for unsupervised image classification and segmentation," in *ICCV*, 2019, pp. 9865–9874.