CAUCHY-SCHWARZ DIVERGENCE BASED INDEPENDENT COMPONENT ANALYSIS WITH AN APPLICATION TO SINGLE CHANNEL SPEECH ENHANCEMENT

Yuanle Li[1†], *Zhenghan Chen*[2†], *Hongqing Liu*[1], *Yi Zhou*[1*], *Xiaoxuan Liang*[3]

[1]Chongqing University of Posts and Telecommunications, Chongqing, China
[2]Peking University, Beijng, China [3]University of Massachusetts Amherst, Amherst MA, USA

ABSTRACT

We develop a new neural independent component analysis (ICA) approach by directly minimizing the dependence amongst all extracted components. To this end, we measure dependence with the Cauchy-Schwarz (CS) divergence between the joint distribution of all components and the product of their marginal distributions, which has a closed-form sample estimator but is computationally cumbersome. Following this, we consider an approximation which significantly reduces the computational burden. The theoretical justification on our approximation is provided. We also empirically demonstrate its superiority over other popular dependence measures that have been used in ICA literature before, such as Shannon's mutual information, Renyi's mutual information, and Hilbert-Schmidt Independence Criterion (HSIC). We finally evaluate the performance of our framework on the problem of single channel speech enhancement.

Index Terms— Independent component analysis, Cauchy-Schwarz divergence, single channel speech enhancement

1. INTRODUCTION

The aim of independent component analysis (ICA) is to extract statistically independent hidden factors (or components) from a mixed signal, by learning a possibly nonlinear unmixing function. It has been widely applied to different types of signals including speech [1], medical images [2], biological assays [3]. Two learning objectives dominate in existing ICA methods [4]: 1) maximizing the non-Gaussianity of the predicted components; 2) minimizing the dependence or mutual information (MI) between different components.

MI is for the ICA for the most part expressed in terms of the Kullback-Leibler (KL) divergence, which corresponds to the Shannon's MI. However, Shannon's MI is notoriously difficult to estimate, especially in high-dimensional space. Over the last decades, substantial efforts have been made to pursue an easy-to-estimate and data-efficient MI or dependence measure that is also differentiable, such that it could be directly optimized in modern machine learning models with stochastic gradient descent (SGD). Notable measures in this line of research include the Rényi's MI [5], the Hilbert-Schmidt Independence Criterion (HSIC) [6], the matrix-based entropy functional [7, 4], and the recently proposed mutual information neural estimator (MINE) [8, 9] that uses a neural network to estimate a lower bound of MI.

In this paper, we aim to draw the attention of community to the Cauchy-Schwarz (CS) divergence [10, 11] based dependence measure [12, 13], which has witnessed great success in traditional signal processing tasks, but its performance in deep learning era is largely underrated and unknown. As a complement to previous literature, we discuss the computational limitation of the sample estimator of CS divergence based dependence measure (when there are more than two variables) and suggest a fast approximation with theoretical guarantee. Using this approximated dependence measure, we further propose a much simpler neural ICA framework without adversarial training and apply it to the problem of single channel speech enhancement (SCSE) which uses noisy speech collected from a single microphone. Specifically,

1. We analyze the computational limitation of CS divergence based dependence measures (in case of more than 2 variables) and provide a sample estimator to its approximation. We also provide a sandwich bound to justify our approximation.
2. We further develop a new neural ICA approach by directly minimizing CS-TC over all predicted components, which can be optimized with SGD, without any variational approximation or adversarial training.
3. Using our developed neural ICA as a main ingredient, we showcase its effectiveness in the problem of SCSE. This is the worst-case scenario of a system which is underdetermined with two unknowns and one equation.

2. BACKGROUND AND PROBLEM FORMULATION

ICA assumes the observed random vector $\boldsymbol{x} = [x_1, x_2, \ldots, x_d]$ is generated by p independent latent variables (which are also called the independent sources or components) $\boldsymbol{s} = [s_1, s_2, \ldots, s_p]$ as:

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

*Contact author: zhouy@cqupt.edu.cn; †Co-first authors.

in which f is a mixing function, which is usually assumed to be linear in classic ICA approaches [14, 15], that is, $\boldsymbol{x} = W^T \boldsymbol{s}$. The goal of ICA is then to recover the inverse function f^{-1} (which is characterized by a possibly nonlinear unmixing function g) as well as the statistically independent components $\tilde{\boldsymbol{s}} = [\tilde{s}_1, \tilde{s}_2, \ldots, \tilde{s}_p]$ solely based on observations $\boldsymbol{x}$:

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

Neural ICA relaxes the linear assumption of f by modeling g with a neural network [16, 4]. To encourage independence, [9] uses MINE [8] to estimate a lower bound of MI, which requires training an additional neural network. [17] maximizes the non-Gaussianity of latent components via an autoencoder in which a decoder is used for reconstruction.

3. METHOD

3.1. CS-ICA framework

In this work, we use a deep neural network g, parameterized by θ, as the nonlinear unmixing function. Our goal is to minimize the total dependence amongst all predicted components $[\tilde{s}_1, \tilde{s}_2, \ldots, \tilde{s}_p]$, without using an additional neural network to estimate mutual information (e.g., [9]) or reconstruct input data $\mathbf{x}$ (e.g., [16, 17]). The overall framework is shown in Fig. 1 (only blue and red modules, without yellow one).

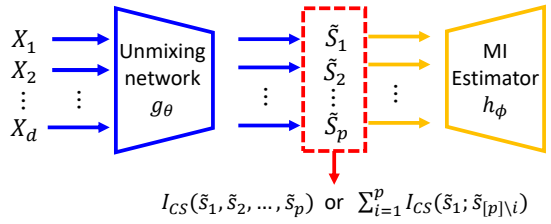

Fig. 1: Our neural ICA (blue module) avoids introduction of an additional neural network h_ϕ to estimate (a lower bound of) mutual information or reconstruct input data (yellow module). We directly optimize $I_{CS}\left(\tilde{s}_1, \tilde{s}_2, \ldots, \tilde{s}_p\right)$ or its approximation $\sum_{k=1}^{p} I_{CS}(s_k; s_{[p]\setminus k})$ (red module).

To encourage the independence of p predicted components $\{\tilde{s}_i\}_{i=1}^{p}$, we directly minimize their total dependence with Cauchy-Schwarz divergence Total correlation (CS-TC) $I_{CS}\left(\tilde{s}_1, \tilde{s}_2, \ldots, \tilde{s}_p\right)$, which is defined in Eq. (6) and can be approximated with Eq. (8).

3.2. CS divergence and it induced dependence measures

Motivated by the famed Cauchy-Schwarz (CS) inequality:

$$\left(\int p(x)q(x)\,dx\right)^2 \leq \int p(x)^2\,dx \int q(x)^2\,dx, \tag{3}$$

with equality if and only if $p(x)$ and $q(x)$ are linearly dependent, the CS divergence [12, 11] is defined as a measure of distance between the probability density functions:

$$D_{CS}(p;q) = -\log\left(\frac{\left(\int p(x)q(x)\,dx\right)^2}{\int p(x)^2\,dx \int q(x)^2\,dx}\right). \tag{4}$$

The CS divergence is symmetric and has closed-form expression for mixture-of-Gaussians (MoG) [18], a desirable property that does not hold for the well-known Kullback-Leibler (KL) divergence. We recommend interested readers to [19] for a recent tutorial in ICASSP-23 regarding this topic.

Two variables s_1 and s_2 are said to be independent if and only if their joint distribution equals to the product of marginal distributions, that is, $p\left(s_1, s_2\right) = p\left(s_1\right) p\left(s_2\right)$. This conclusion can be extended to $p > 2$ variables. Specifically, variables $s_1, s_2, \ldots, s_p$ are mutually independent if and only if $p\left(s_1, s_2, \ldots, s_p\right) = p\left(s_1\right) p\left(s_2\right) \ldots p\left(s_p\right)$. Hence, a reliable way to measure the total dependence amongst p variables is by computing the distance between $p\left(s_1, s_2, \cdots, s_p\right)$ and $p\left(s_1\right) p\left(s_2\right) \cdots p\left(s_p\right)$. If we use the KL divergence, we obtain the total correlation (TC) [20]:

$$\begin{aligned} I(s_1, s_2, \ldots, s_p) &:= D_{KL}\left(p(s_1, s_2, \cdots, s_p); p(s_1)p(s_2)\cdots p(s_p)\right) \\ &= \sum_{k=1}^{p} H(s_k) - H(s_1, s_2, \ldots, s_p), \end{aligned} \tag{5}$$

in which H refers to Shannon entropy or joint entropy.

By replacing KL divergence with CS divergence, one obtains a new measure of dependence in Eq. (6), which is also called the Cauchy-Schwarz Total Correlation (CS-TC) [12] and denoted as I_{CS}. When $p = 2$, CS-TC reduces to the Cauchy-Schwarz Mutual information (CS-MI) [19], which has been used in [13] in a classic 2-source-2-sensor ICA problem with a linear unmixing function. Later, [21] builds the connection between CS-MI and kernel ICA [22]. [23] and [24] extends the general idea of CS-MI and apply, respectively, the Jensen' inequality and the convex CS inequality on $p(s_1, s_2)$ and $p(s_1)p(s_2)$ to measure the dependence.

$$\begin{aligned} &I_{CS}(s_1, s_2, \ldots, s_p) := D_{CS}\left(p(s_1, s_2, \ldots, s_p); p(s_1)p(s_2)\ldots p(s_p)\right) \\ &= -2\log\left(\int p(s_1, s_2, \ldots, s_p)p(s_1)p(s_2)\ldots p(s_p)\,ds_1 ds_2 \ldots ds_p\right) \\ &+ \log\left(\int p(s_1, s_2, \ldots, s_p)^2\,ds_1 ds_2 \ldots ds_p\right) \\ &+ \log\left(\int \left(p(s_1)p(s_2)\ldots p(s_p)\right)^2\,ds_1 ds_2 \ldots ds_p\right). \end{aligned} \tag{6}$$

Distinct to these earlier literature, our paper concerns about the empirical estimator of CS-TC in case of more than 2 variables and its utility in deep learning framework with nonlinear unmixing function. Hence, the scalability and efficiency of the used estimator becomes a priority.

The CS-TC can be estimated from given observations by Proposition 1. Proof in our GitHub repository `https://github.com/Yuanle-Lee/ICASSP_CS_TC`.

Proposition 1 (Empirical Estimator of CS-TC) *Given N observations $\left\{\left(s_1^i, s_2^i, \ldots, s_p^i\right)\right\}_{i=1}^N$, each observation contains p different types of measurements $s_1 \in \mathcal{S}_1, s_2 \in \mathcal{S}_2, \ldots, s_p \in \mathcal{S}_p$. Let $Q_k \in \mathbb{R}^{N\times N}$ denote the Gram matrix for the k-th $(1 \le k \le p)$ measurement, i.e., $Q_k(i,j) = G_\sigma\left(s_k^i - s_k^j\right)$, in which G_σ refers to a Gaussian kernel with width σ, i.e., $G_\sigma\left(s_k^i - s_k^j\right) = \exp\left(-\frac{\left\|s_k^i - s_k^j\right\|^2}{2\sigma^2}\right)$. The empirical estimator of CS-TC is given by:*

$$\begin{aligned}\hat{I}_{CS}(s_1, s_2, \ldots, s_p) &= \log\left(\frac{1}{N^2}\sum_{(i,j)\in \mathbf{i}_2^N}\prod_{k=1}^{p} Q_k(i,j)\right)\\ &+ \log\left(\frac{1}{N^{2p}}\sum_{(i_1,j_1,i_2,j_2,\cdots,i_p,j_p)\in \mathbf{i}_{2p}^N}\prod_{k=1}^{p} Q_k(i_k,j_k)\right)\\ &- 2\log\left(\frac{1}{N^{p+1}}\sum_{(i,j_1,j_2,\cdots,j_p)\in \mathbf{i}_{p+1}^N}\prod_{k=1}^{p} Q_k(i,j_k)\right),\end{aligned} \tag{7}$$

*where the index set $\mathbf{i}_r^N$ denotes the set of all r-tuples drawn **with** replacement from $\{1, 2, \cdots, N\}$.*

An exact and naïve computation of Eq. (7) would require $\mathcal{O}(N^{2p})$ operations. To reduce computational burden, we use Eq. (8) as an alternative, i.e., the sum of Cauchy-Schwarz based mutual information (CS-MI) between the k-th component s_k and the rest of all components $s_{[p]\backslash k} = [s_1, \cdots, s_{k-1}, s_{k+1}, \cdots, s_p]$:

$$\sum_{k=1}^{p} I_{CS}(s_k; s_{[p]\backslash k}). \tag{8}$$

The following Lemmas guarantee the effectiveness of Eq. (8).

Lemma 1 *Both $I_{CS}(s_1, s_2, \cdots, s_p)$ and $\sum_{k=1}^p I_{CS}(s_k; s_{[p]\backslash k})$ reduce to zero if and only if all components $\{s_1, s_2, \cdots, s_p\}$ are independent to each other.*

Lemma 2 *Total correlation is closely related to the sum of mutual information between individual component s_i and all rest components $s_{[p]\backslash i}$, in particular:*

$$\frac{p}{p-1} I(s_1, s_2, \cdots, s_p) \le \sum_{k=1}^{p} I(s_k; s_{[p]\backslash k}) \le p I(s_1, s_2, \cdots, s_p). \tag{9}$$

Lemmas 1 and 2 jointly indicate that Eq. (8) is both lower and upper bounded by Eq. (6). Both measures reach to 0 when all components are independent. Hence, Eq. (8) serves as a reliable alternative or approximation to Eq. (6).

Proposition 2 (Empirical Estimator of $I_{CS}(s_k; s_{[p]\backslash k})$ [19, 12]) *Given N observations $\left\{\left(s_1^i, s_2^i, \ldots, s_p^i\right)\right\}_{i=1}^N$, each observation contains p different types of measurements $s_1 \in \mathcal{S}_1, s_2 \in \mathcal{S}_2, \ldots, s_p \in \mathcal{S}_p$. Let Q and L denote, respectively, the Gram matrices for variable s_k and all rest variables $s_{[p]\backslash k} = [s_1, \cdots, s_{k-1}, s_{k+1}, \cdots, s_p]$, e.g., $L(i,j) = \exp\left(-\frac{\left\|s_{[p]\backslash k}^i - s_{[p]\backslash k}^j\right\|^2}{2\sigma^2}\right)$. The empirical estimator of $I_{CS}(s_k; s_{[p]\backslash k})$ is given by:*

$$\begin{aligned}\widehat{I}_{CS}(s_k; s_{[p]\backslash k}) &= \log\left(\frac{1}{N^2}\sum_{i,j}^{N} Q_{ij}L_{ij}\right) + \log\left(\frac{1}{N^4}\sum_{i,j,q,r}^{N} Q_{ij}L_{qr}\right)\\ &- 2\log\left(\frac{1}{N^3}\sum_{i,j,q}^{N} Q_{ij}L_{iq}\right).\end{aligned} \tag{10}$$

A direct computation of Eq. (10) still requires $\mathcal{O}(n^4)$ operations. However, an equivalent form which needs $\mathcal{O}(n^2)$ operations can be formulated as [25]:

$$\begin{aligned}\widehat{I}_{\text{CS}}(s_k; s_{[p]\backslash k}) &= \log\left(\frac{1}{N^2}\operatorname{tr}(QL)\right) + \log\left(\frac{1}{N^4}\mathbb{1}^T Q \mathbb{1}\mathbb{1}^T L \mathbb{1}\right)\\ &- 2\log\left(\frac{1}{N^3}\mathbb{1}^T Q L \mathbb{1}\right),\end{aligned} \tag{11}$$

where $\mathbb{1}$ is a vector of 1s of relevant dimension. Hence, the overall computational complexity to estimate Eq. (8) is $\mathcal{O}\left(N^2 p\right)$. To avoid confusion, the following experiments are performed with approximated CS-TC, i.e., Eq. (8), in which $I_{\text{CS}}(s_k; s_{[p]\backslash k})$ is estimated by Eq. (11).

3.3. Advantage of CS-TC.

We explain the advantages of approximated CS-TC in terms of computational complexity and statistical power.

Computational Complexity: The computational complexity of approximated CS-TC is $\mathcal{O}\left(N^2 p\right)$. For HSIC, the total dependence for $p > 2$ variables can be obtained by $\sum_{i=1}^p \sum_{j=i+1}^p \text{HSIC}\left(s_i, s_j\right)$, which results in least $\mathcal{O}\left(N^2 p^2\right)$ operations. For Rényi's MI, the computational complexity with KDE is again $\mathcal{O}\left(N^2 p\right)$. However, its statistical power is lower than CS-TC, as will be demonstrated later. For Shannon's MI, if we estimate it with the state-of-the-art (SOTA) MINE [8], it requires training an additional neural network, which usually takes much longer training time.

Statistical power. To quantitatively demonstrate the advantage of approximated CS-TC over other dependence measures including HSIC, Shannon's and Renyi's MI, we run two simulations in scope to this in [6, 26].

Data A: First, we randomly generated $N = 256$ 1-dimensional *i.i.d.* samples from two randomly picked densities in the ICA benchmark densities [22]. Second, we mixed these random variables using a rotation matrix parameterized by an angle θ, varying from 0 to $\pi/4$ (a zero angle means

the data are independent, while dependence becomes easier to detect as the angle increases to $\pi/4$). Hence, a reliable dependence measure is expected to have an acceptance rate of the H_0 hypothesis (there is no dependence between two groups of observations) at $\theta = 0$. But this rate has a rapidly decaying as θ increases. According to Fig. 2(a), the CS-TC outperforms other three measures in discovering dependence.

Data B: There is a functional relationship between s_1 and the remaining dimensions: $s_1 = \left(\frac{1}{p-1}\sum_{k=2}^{p} s_k\right)^2$, where $\{s_2, s_3, \cdots, s_p\}$ are uniformly and independently distributed. In this case, the strength of the total dependence should decrease with the increase of dimension. Fig. 2(b) shows the average value of the analyzed measures induced on $p-1 \in [1, 9]$ and $N = 100$ samples. Both CS-TC and Rényi's MI are close to the expected curve. The estimator of Shannon's MI obtains negative values, which is hard to interpret since dependence should always be non-negative.

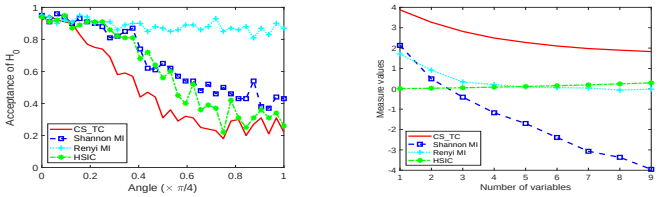

Fig. 2: Statistical power evaluation.

4. EXPERIMENTS

We compare neural ICA optimized by CS-TC with baseline fastICA [14], InformaxICA [15] and the SOTA MINE-based ICA [8, 9]. We then test its practical performance in SCSE.

4.1. ICA on synthetic PNL data

We first test all competing ICA approaches on actual 16kHz audio signals [27] but with synthetic post non-linear (PNL) 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 $\tanh(x)$, $\left(x + x^3\right)/2, e^x$. respectively. We set the batch size to be $2,000$ and results are normalized using Zscore.

The g_θ in our deep ICA is a four-hidden-layer fully connected neural network with ReLU activation in the first three layers and linear activation in the last layer. It is followed by a differentiable whitening layer to avoid trivial solution [28]. The kernel width σ is 0.16 according to the Silverman's rule of thumb [29]. Experiments are repeated for 10 times and the average scores are recorded. According to Table 1, our neural ICA achieves the best performance.

4.2. Application to Single Channel Speech Enhancement

ICA has been used for single channel speech enhancement (SCSE) before [30, 31, 32]. Fig. 3 shows our block diagram of SCSE, which is from [1], but replaces the naïve linear ICA with our neural ICA. In this diagram, the input noisy speech $x(n)$ is passed to a LogMMSE [33] module. The output $\hat{s}_1(n)$ is used as the first input to neural ICA module and the original noisy speech $x(n)$ as the second input. After ICA transformation, the outputs $\hat{s}_2(n)$ and $\hat{d}(n)$ are the estimated speech and noise respectively. To enhance the quality of $\hat{s}_2(n)$, the residual noise in $\hat{s}_2(n)$ is further suppressed with LogMMSE. $\hat{s}_3(n)$ is finally enhanced speech.

Methods	FastICA	Infomax	MINE	Ours
$\lvert\rho\rvert$	0.713	0.665	0.769± 0.037	**0.797**± 0.012

Table 1: Separation performance in terms of absolute correlation coefficient $|\rho|$. The best performance is in bold.

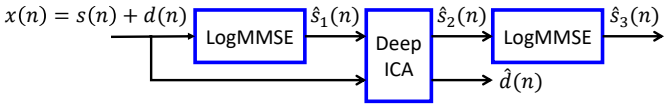

Fig. 3: Block diagram of SCSE with our neural ICA.

We select the same clean speeches as in Section 4.1, and added DrivingCar noise[1] [34] with SNR at 2dB, 5dB, and 10dB. We compare our approach with Reddy *et al.* [1] and Hao *et al.* [32] in 5 independent runs, and use Perceptual Evaluation of Speech Quality (PESQ) as quality measure. Fig. 4 suggests that our neural ICA outperforms others. For our method, its standard deviation (std) is always smaller than 0.015, and is hence ignored herein.

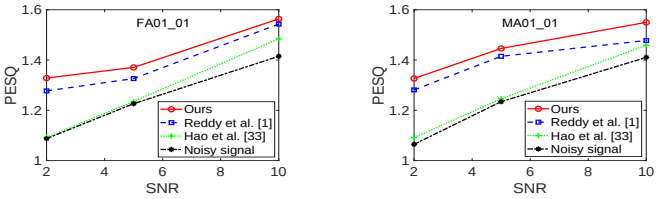

Fig. 4: PESQ under DrivingCar noise of 2dB, 5dB, 10dB. The black dashed line corresponds to signal without enhancement.

5. CONCLUSION.

We consider using CS divergence in neural ICA, rather than MINE. To this end, we analyze the computational issue regarding the sample estimator of Cauchy-Schwarz Total Correlation (CS-TC) and provide an efficient approximation in terms of Cauchy-Schwarz Mutual Information (CS-MI), which reduces the computational complexity from $\mathcal{O}(N^{2p})$ to $\mathcal{O}(N^2 p)$. Second, we numerically demonstrate the advantage of approximated CS-TC over existing independence measures. The use of CS-TC enables us to design a much simpler neural ICA framework without adversarial training or an extra neural network. We finally showcase the practical usage of our neural ICA in single channel speech enhancement.

[1]Results of Machineary noise and descriptions regarding both noises are in `https://github.com/Yuanle-Lee/ICASSP_CS_TC`.

6. REFERENCES

[1] Chandan KA Reddy, Anshuman Ganguly, and Issa Panahi, "Ica based single microphone blind speech separation technique using non-linear estimation of speech," in *IEEE ICASSP*, 2017, pp. 5570–5574.

[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] Hongming Li, Shujian Yu, and José C Príncipe, "Deep deterministic independent component analysis for hyperspectral unmixing," in *IEEE ICASSP*, 2022, pp. 3878–3882.

[5] Kenneth E Hild, Deniz Erdogmus, and José Príncipe, "Blind source separation using renyi's mutual information," *IEEE Signal Processing Letters*, vol. 8, no. 6, pp. 174–176, 2001.

[6] Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola, "A kernel statistical test of independence," *NeurIPS*, vol. 20, 2007.

[7] Shujian Yu, Luis Gonzalo Sanchez Giraldo, Robert Jenssen, and Jose C Principe, "Multivariate extension of matrix-based rényi's α-order entropy functional," *IEEE TPAMI*, vol. 42, no. 11, pp. 2960–2966, 2019.

[8] Mohamed Ishmael Belghazi et al., "Mutual information neural estimation," in *ICML*. PMLR, 2018, pp. 531–540.

[9] Hlynur Davíð Hlynsson and Laurenz Wiskott, "Learning gradient-based ica by neurally estimating mutual information," in *42nd German Conference on AI*, 2019, pp. 182–187.

[10] Robert Jenssen et al., "The cauchy–schwarz divergence and parzen windowing: Connections to graph theory and mercer kernels," *Journal of the Franklin Institute*, vol. 343, no. 6, pp. 614–629, 2006.

[11] Shujian Yu, Hongming Li, Sigurd Løkse, Robert Jenssen, and José C Príncipe, "The conditional cauchy-schwarz divergence with applications to time-series data and sequential decision making," *arXiv preprint arXiv:2301.08970*, 2023.

[12] Jose C Principe, *Information theoretic learning: Renyi's entropy and kernel perspectives*, Springer Science & Business Media, 2010.

[13] Dongxin Xu, Jose C Principe, John Fisher, and Hsiao-Chun Wu, "A novel measure for independent component analysis (ica)," in *IEEE ICASSP*, 1998, vol. 2, pp. 1161–1164.

[14] 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.

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

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

[17] Po-Ting Yeh, Arthur C Tsai, Chia-Ying Hsieh, Chia-Cheng Yang, and Chun-Shu Wei, "Oicnet: A neural network for online eeg source separation using independent component analysis," *bioRxiv*, pp. 2023–05, 2023.

[18] Kittipat Kampa, Erion Hasanbelliu, and Jose C Principe, "Closed-form cauchy-schwarz pdf divergence for mixture of gaussians," in *IEEE IJCNN*, 2011, pp. 2578–2585.

[19] "Information theory meets deep learning," `https://rb.gy/44t0p`, 2023, IEEE ICASSP Tutorial.

[20] Satosi Watanabe, "Information theoretical analysis of multivariate correlation," *IBM Journal of research and development*, vol. 4, no. 1, pp. 66–82, 1960.

[21] Jian-Wu Xu, Deniz Erdogmus, Robert Jenssen, and Jose C Principe, "An information-theoretic perspective to kernel independent components analysis," in *IEEE ICASSP*, 2005, vol. 5, pp. 249–252.

[22] Francis R Bach and Michael I Jordan, "Kernel independent component analysis," *JMLR*, vol. 3, no. Jul, pp. 1–48, 2002.

[23] Jen-Tzung Chien, Hsin-Lung Hsieh, and Sadaoki Furui, "A new mutual information measure for independent component alalysis," in *IEEE ICASSP*, 2008, pp. 1817–1820.

[24] Zaid Albataineh and Fathi M Salem, "Convex cauchy schwarz independent component analysis for blind source separation," *arXiv preprint arXiv:1408.0192*, 2014.

[25] Le Song, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt, "Feature selection via dependence maximization.," *JMLR*, vol. 13, no. 5, 2012.

[26] Simone Romano et al., "Measuring dependency via intrinsic dimensionality," in *IEEE ICPR*, 2016, pp. 1207–1212.

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

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

[29] Bernard W Silverman, *Density estimation for statistics and data analysis*, Routledge, 2018.

[30] Liang Hong, Justinian Rosca, and Radu Balan, "Independent component analysis based single channel speech enhancement," in *IEEE ISSPIT*, 2003, pp. 522–525.

[31] Hong-yan Li, Qing-hua Zhao, Guang-long Ren, and Bao-jin Xiao, "Speech enhancement algorithm based on independent component analysis," in *ICNC*. IEEE, 2009, vol. 2, pp. 598–602.

[32] Xueyuan Hao, Yu Shi, and Xiaohong Yan, "Speech enhancement algorithm based on independent component analysis in noisy environment," in *ICAIIS*. IEEE, 2020, pp. 456–461.

[33] Yariv Ephraim and David Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE TASLP*, vol. 33, no. 2, pp. 443–445, 1985.

[34] "Smartphone-based open research platform for hearing improvement studies," `https://labs.utdallas.edu/ssprl/hearing-aid-project/database/`.