UNIVERSIDADE FEDERAL DO CEARÁ
CENTRO DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA DE TELEINFOMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE TELEINFORMÁTICA
MESTRADO ACADÊMICO EM ENGENHARIA DE TELEINFORMÁTICA

LUCAS DE PAULA DAMASCENO

INDEPENDENT VECTOR ANALYSIS USING SEMI-PARAMETRIC DENSITY ESTIMATION VIA MULTIVARIATE ENTROPY MAXIMIZATION

FORTALEZA
2021

LUCAS DE PAULA DAMASCENO

INDEPENDENT VECTOR ANALYSIS USING SEMI-PARAMETRIC DENSITY ESTIMATION VIA MULTIVARIATE ENTROPY MAXIMIZATION

Dissertação apresentada ao Curso de Mestrado Acadêmico em Engenharia de Teleinformática do Programa de Pós-Graduação em Engenharia de Teleinformática do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Engenharia de Teleinformática. Área de Concentração: Sinais e Sistemas

Orientador: Prof. Dr. Charles Casimiro Cavalcante

Coorientador: Prof. Dr. Zois Boukouvalas

FORTALEZA

2021

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

D162i Damasceno, Lucas de Paula.
Independent vector analysis using semi-parametric density estimation via multivariate entropy maximization / Lucas de Paula Damasceno. – 2019.
46 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Tecnologia, Programa de Pós-Graduação em Engenharia de Teleinformática, Fortaleza, 2019.
Orientação: Prof. Dr. Charles Casimiro Cavalcante.
Coorientação: Prof. Dr. Zois Boukouvalas.

1. Análise de vetores independentes. 2. Estimação de função densidade de probabilidade multivariada. 3. Princípio da máxima entropia. 4. Métodos de Monte Carlo. I. Título.

CDD 621.38

LUCAS DE PAULA DAMASCENO

INDEPENDENT VECTOR ANALYSIS USING SEMI-PARAMETRIC DENSITY ESTIMATION VIA MULTIVARIATE ENTROPY MAXIMIZATION

Dissertation defended at the Teleinformatics Engineering Master's Program at the Teleinformatics Engineering Post-Graduate Program of the Technology Center at the Federal University of Ceara, as a requirement to obtain the master degree in Teleinformatics Engineering. Concentration Area: Signals and Systems

Aproved on: May 3, 2021

EXAMINING COMMITTEE

Prof. Dr. Charles Casimiro Cavalcante (Advisor)
Federal University of Ceará

Prof. Dr. Zois Boukouvalas (Co-Advisor)
American University

Prof. Dr. Maria Barouti
American University

Prof. Dr. Yuri Levin-Schwartz
Icahn School of Medicine at Mount Sinai

Prof. Dr. Guilherme Barreto
Federal University of Ceará

To my parents, family and friends.

ACKNOWLEDGEMENTS

This study was financed by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (Procs. 132583/2019-4).

I would like to thank my parents and brothers, Damasceno, Rosangela, Robson and Tiago for all support. Also thank my girlfriend Ana Livia who always supported me from the beginning.

I am also very grateful to my advisor and co-advisor Prof. Dr. Charles Cavalcante and Prof. Dr. Zois Boukouvalas for their great guidance, talks, meetings, which provided the success of our research.

I would also like to thank my undergraduate advisor Prof. Dr. Alexandre Fernandes who encouraged me to start this journey.

Also, thank Prof. Dr. Maria Barouti, Prof. Dr. Yuri Levin-Schwartz, and Prof. Dr. Guilherme Barreto for accepting to be part of the evaluation committee.

My sincere acknowledgment to everyone!

"If you can't fly, then run, if you can't run, then walk, if you can't walk, then crawl, but whatever you do, you have to keep moving forward."

(Martin Luther King)

RESUMO

A separação cega de fontes (BSS) é uma ativa área de pesquisa em processamento estatístico de sinais devido às suas inúmeras aplicações, como análise de dados de imagens médicas, comunicações sem fio e processamento de imagens. Devido ao amplo uso da tecnologia de multi-sensores, a análise de múltiplos conjuntos de dados está no centro de muitos problemas desafiadores na engenharia. Isso motiva o desenvolvimento de modelos de separação cega de fontes para múltiplos conjunto de dados (JBSS) assumindo dependência estatística entre fontes latentes através de misturas. A análise de componentes independentes (ICA) é um método de BSS amplamente utilizado que pode alcançar a recuperação da fonte de forma exclusiva, sujeito apenas a ambigüidades de escala e permutação, por meio da suposição de independência estatística por parte das fontes latentes. Embora o ICA seja um dos algoritmos mais comumente usados, ele só pode decompor um único conjunto de dados. Isso tem impulsionado o desenvolvimento da análise de vetores independentes (IVA) uma generalização recente do ICA para múltiplos conjuntos de dados que pode alcançar um desempenho aprimorado em relação ao desempenho do ICA em cada conjunto separadamente, explorando dependências entre os conjuntos de dados. Embora os algoritmos ICA e IVA possam ser modelados com base na estrutura de máxima verossimilhança de modo que todos os tipos de diversidade disponíveis sejam levados em consideração simultaneamente por meio do uso de modelos de densidade geral para as fontes multivariadas latentes, eles frequentemente se desviam de suas propriedades de otimização devido à estimação inadequada da função densidade de probabilidade. Portanto, para garantir a eficiência dos algoritmos de IVA, é necessário um método de estimação de densidade eficiente. Nesta dissertação, apresentamos uma técnica de estimação de densidade multivariada com base no princípio da máxima entropia que utiliza conjuntamente funções de medição multidimensionais globais e locais para fornecer funções densidade de probabilidade flexíveis e além disso, integramos no algoritmo proposto uma técnica de integração multidimensional baseada no método de Monte Carlo. Então, derivamos um novo algoritmo de IVA, que aproveita a capacidade do método proposto de estimação de densidade para aprimorar o desempenho de separação de fontes em uma ampla gama de distribuições. Utilizamos experimentos numéricos para demonstrar o desempenho superior sobre algoritmos amplamente utilizados.

Palavras-chave: Análise de vetores independentes. Estimação de função densidade de probabilidade multivariada. Princípio da máxima entropia. Métodos de Monte Carlo.

ABSTRACT

Blind source separation (BSS) is an active area of research in statistical signal processing due to its numerous applications, such as analysis of medical imaging data, wireless communications, and image processing. Due to the wide use of multi-sensor technology, analysis of multiple datasets is at the heart of many challenging engineering problems. This motivates the development of the field of joint blind source separation (JBSS), which extends the classical BSS to simultaneously resolve several BSS problems by assuming statistical dependence between latent sources across mixtures. Independent component analysis (ICA) is a widely used BSS method that can uniquely achieve source recovery, subject to only scaling and permutation ambiguities, through the assumption of statistical independence on the part of the latent sources. Although ICA is one of the most commonly used, it can only decompose a single dataset. This has driven the development of independent vector analysis (IVA), a recent generalization of ICA to multiple datasets that can achieve improved performance over performing ICA on each dataset separately by exploiting dependencies across datasets. Though both ICA and IVA algorithms cast in the maximum likelihood (ML) framework such that all available types of diversity are taken into account simultaneously through the use of general density models for the latent multivariate sources, they often deviate from their theoretical optimality properties due to improper estimation of the probability density function (PDF). Therefore, in order to guarantee the effectiveness of IVA algorithms, an efficient density estimation method is required.

In this dissertation, we present a multivariate density estimation technique based on the maximum entropy principle (MEP) that jointly uses global and local multidimensional measuring functions to provide flexible PDFs while keeping the complexity low by integrating into the proposed algorithm a multidimensional Monte-Carlo (MC) integration technique. Finally, we derive a new IVA algorithm, which takes advantage of the accurate estimation capability of the proposed density estimation method to greatly improve separation performance from a wide range of distributions. We use numerical experiments to demonstrate the superior performance over widely used algorithms.

Keywords: Independent vector analysis. Multivariate probability density estimation. Maximum entropy distributions. Monte Carlo methods.

LIST OF FIGURES

Figure 1 – Analysis of brain activity using electroencephalogram (EEG) data collected from multiple subjects. 13
Figure 2 – JBSS structure. 17
Figure 3 – Illustration of SCV and source vector. 18
Figure 4 – Example of a generation of 16 samples in $[0, 1] \in \mathbb{R}$. The van der Corput sequence is obtained by reversing the bits in the binary decimal representation of the naive sequence. 26
Figure 5 – Estimation performance only using global constraints and using global and local constraints represented in the second and third line, respectively. . . . 28
Figure 6 – Comparison in terms of matching with the histogram. First line–histogram of the generated data, first column–parametric method (GMM), second column–non-parametric method (KDE), third column–semi-parametric method (M-EMK). 29
Figure 7 – Performance comparison in terms of CPU time. 30
Figure 8 – Estimation performance in terms of matching with the histogram. First column–histogram of the wind turbine data, second column–M-EMK. . . . 30
Figure 9 – Density estimation comparison. First column– parametric method (GMM), second column–non-parametric method (KDE), third column–semi-parametric method (M-EMK). 31
Figure 10 – Performance comparison of the first case in terms of Joint ISI 33
Figure 11 – Performance comparison of the second case in terms of Joint ISI. 34
Figure 12 – Performance comparison in terms average CPU time for different number of sample size. The first and second case represented in the first and second columns, respectively. 35
Figure 13 – Mean absolute error (MAE) as a function of number fused datasets. True fusion by IVA performs better than concatenating molecular featurization methods. 38

Figure 14 – Graphical multi-modal fusion example using IVA and its capabilities. Through the estimation of the source component vectors and their associated covariance matrices we identify correlation patterns among different multi-modal datasets and define common and distinct subspaces through the estimated feature embeddings. Common subspaces will enable the knowledge discovery and distinct subspaces will enable the detection of misinformation. 39

LIST OF ABBREVIATIONS AND ACRONYMS

BSS	Blind Source Separation
EEG	Electroencephalogram
GMM	Gaussian Mixture Model
HOS	Higher-order statistics
ICA	Independent Component Analysis
IVA	Independent Vector Analysis
IVA-M-EMK	IVA by Multivariate Entropy Maximization with Kernels
IVA-G	IVA-Gaussian
IVA-L	IVA-Laplacian
ISI	Inter-Symbol-Interference
JBSS	Joint Blind Source Separation
KDE	Kernel Density Estimation
MEP	Maximum Entropy Principle
ML	Maximum Likelihood
MC	Monte-Carlo
M-EMK	Multivariate Entropy Maximization with Kernels
MI	Mutual Information
PDF	Probability Density Function
QMC	Quasi-Monte Carlo
SCV	Source Component Vector

LIST OF SYMBOLS

$\mathbf{x}$	Dataset
K	Number of datasets
N	Number of statistically independent sources
$\mathbf{A}^{[k]}$	Mixing matrices
$\mathbf{s}^{[k]}$	Vector of latent sources
$\mathscr{L}$	ML function
$\mathbf{y}_n$	nth estimated random vector
$p(\mathbf{y}_n)$	Multidimensional PDF
$\mathbf{W}$	Demixing matrices
I	MI cost function
$f(\|\|)$	NKullback–Leibler distance
$\phi^{[k]}$	Score function
$\mathbf{w}_i^{[k]}$	Rows of $\mathbf{W}$
$\mathbf{H}_n$	Orthogonal projection onto the null space
p_n	Estimated PDF for the nth iteration
E_{p_n}	Expectation over the distribution p_n
$\mathbf{r}$	Vectors of global and local measuring functions
λ	Lagrangian multipliers
α	Sample averages
$\mathbf{J}$	Jacobian matrix
q	Gaussian kernel
$\boldsymbol{\mu}$	Mean vector
$\|\cdot\|$	Stands for the determinant of the argument
b_k	k-th prime number
Z_{b_k}	Least residue system mod b_k
w_z	Quasi-random number

Ω	Dimensional measure of the region of integration
g_i	Multivariate generalized Gaussian distribution

1 INTRODUCTION

1.1 Motivation

The problem of blind source separation (BSS) is one of the most relevant subjects in unsupervised signal processing and has been intensively studied by this research area because of its potential applications such as speech recognition systems, telecommunications, and medical signal processing. The first general framework to deal with BSS was introduced by Jeanny Hérault and Bernard Ans from 1984 and is founded on the concept of independent component analysis (ICA). Although ICA is one of the most efficient algorithms for the BSS problem, its framework is able to analyze one single set of data, which greatly limits its applicability because, with the growth of multi-sensor technology, the analysis of multiple data sets has been widely addressed in many current engineering problems. This has been the main reason for the development of the joint blind source separation (JBSS) area, a generalization of the BSS problem for joint analysis of multiple datasets (ROMANO *et al.*, 2010).

Figure 1 – Analysis of brain activity using electroencephalogram (EEG) data collected from multiple subjects.

An effective solution to the JBSS problem is independent vector analysis (IVA), a recent extension of independent component analysis (ICA) that makes full use of the statistical dependence across multiple datasets to achieve source separation. Real-world problems that IVA provides an effective solution include detection of a target in a given video sequence or multi-spectral remote sensing data, separation of music signals, stock prediction, and analysis of brain activity using medical image data collected from multiple subjects among many others. IVA can be formulated in a maximum likelihood (ML) framework such that all available types of diversity are taken into account simultaneously. The key issue that will enable the application of IVA to many problems with effective solutions is the development of effective models for the underlying source density and their estimation.

In this master thesis, we present a multivariate density estimation technique based on the MEP that jointly uses global and local multidimensional measuring functions to provide flexible PDFs while keeping the complexity low by integrating into the proposed algorithm a multidimensional Monte-Carlo (MC) integration technique. Next, using the proposed density estimator we derive an efficient IVA algorithm, in order to accurately separate sources from a wide range of multivariate PDFs. We use numerical experiments to demonstrate the superior performance over popular IVA algorithms.

1.2 Contributions

The framework as well as the contributions of this dissertation are organized as follows.

a) Chapter 3 - *Multivariate density estimation*

   In order to provide an optimal framework for the IVA model using the ML framework, the knowledge of the multidimensional PDF that best matches the underlying properties of the latent sources is necessary. The main goal regarding this estimation problem is to achieve a desirable balance between flexibility while maintaining a simple form that would enable generalization, and efficient implementation.

   – To provide a new multivariate density estimator based on entropy maximization with kernels by jointly using global and local measuring functions yielding flexible PDFs while keeping the complexity low;

   – Demonstrate the superior performance of M-EMK over competing density estimation algorithms using simulated as well as real-world data;

   – To propose a multidimensional Monte Carlo (MC) integration technique, that uses sequences of quasi-random numbers in order to achieve computational efficiency;

b) Chapter 4 - *Multivariate EMK IVA Algorithm*

Motivated by the efficient density estimation obtained by the proposed estimation method, we integrate our density estimator into the IVA model. Hence, the following contributions can be listed:

– To develop a new IVA algorithm, IVA by multivariate entropy maximization with kernels (IVA-M-EMK);

– To demonstrate superior source separation performance of IVA-M-EMK among widely used algorithms using simulated data.

1.3 Overview

This thesis is organized as follows. In chapter 2 we present all the theoretical formulation of the IVA algorithm under the ML umbrella. Chapter 3 provides a new flexible and efficient multivariate density estimator and verify its effectiveness. In chapter 4, we apply this density estimator to the development of an effective IVA algorithm that successfully matches multivariate latent sources from a wide range of distributions and demonstrate the superior performance of the new IVA algorithm numerically using simulated data. Finally, in chapter 5 we discuss the conclusions and future research topics of the proposed work.

2 INDEPENDENT VECTOR ANALYSIS

Independent vector analysis (IVA) is a recent generalization of independent component analysis (ICA) that enables the joint factorization of multiple sets of data. Similar to ICA, IVA can be formulated in a maximum likelihood (ML) framework such that all available types of diversity are taken into account simultaneously through the use of general density models for the latent multivariate sources. Originally IVA was formulated for solving the convolutive ICA problem in the frequency domain using multiple frequency bins (KIM, 2010). This led to the development of IVA-Laplacian (IVA-L) (KIM *et al.*, 2006; KIM *et al.*, 2007), an algorithm that takes only higher-order statistics (HOS) into account and assumes a Laplacian distribution for the underlying source component vectors. Conversely, IVA-Gaussian (IVA-G) (ANDERSON *et al.*, 2012; VIA *et al.*, 2011) exploits linear dependencies but does not take HOS into account. Finally, IVA based on the multivariate generalized Gaussian distribution (ANDERSON *et al.*, 2013b; ANDERSON *et al.*, 2014; BOUKOUVALAS *et al.*, 2015; BOUKOUVALAS *et al.*, 2015) are more general IVA implementations where both second and higher order statistics are taken into account. Although current IVA algorithms based on the underlying density model of the latent sources have shown great success in a number of applications. Therefore, the key issue that will enable the successful application of IVA to many problems where such multivariate density modeling is needed (BOUKOUVALAS *et al.*, 2018a; BHINGE *et al.*, 2016; ADALI *et al.*, 2018), is the development of flexible models for the underlying source density and their estimation.

In this chapter, we begin mathematically formulating the IVA algorithm based on a ML framework. Then, we derive the ML objective functions for the mutual information (MI) objective function, which provides a framework that enables the exploitation of multiple forms of diversity, while at the same time enjoying all the theoretical advantages of ML theory. Finally, we introduce the IVA model from an algorithmic point of view ending in an optimization problem that is motivated by unreasonable results and poor convergence of optimization problems with matrix parameters and present a decoupling procedure to overcome this problem.

2.1 IVA

As previously mentioned, IVA is a generalization of ICA. Due to this fact, IVA is mathematically formulated in a similar way, except that now we have K datasets $\mathbf{x}^{[k]}$, $k = 1, ..., K$ where each dataset is a linear mixture of N statistically independent sources. Using random vector notation under the assumption that samples are i.i.d, the noiseless IVA model is given by

$$\mathbf{x}^{[k]} = \mathbf{A}^{[k]}\mathbf{s}^{[k]}, \qquad k = 1,\ldots,K, \tag{2.1}$$

where $\mathbf{A}^{[k]} \in \mathbb{R}^{K\times K}$, $k = 1,\ldots,K$ are invertible mixing matrices and $\mathbf{s}^{[k]} = [s_1^{[1]},\ldots,s_N^{[k]}]^\top$ is the vector of latent sources for the kth dataset.

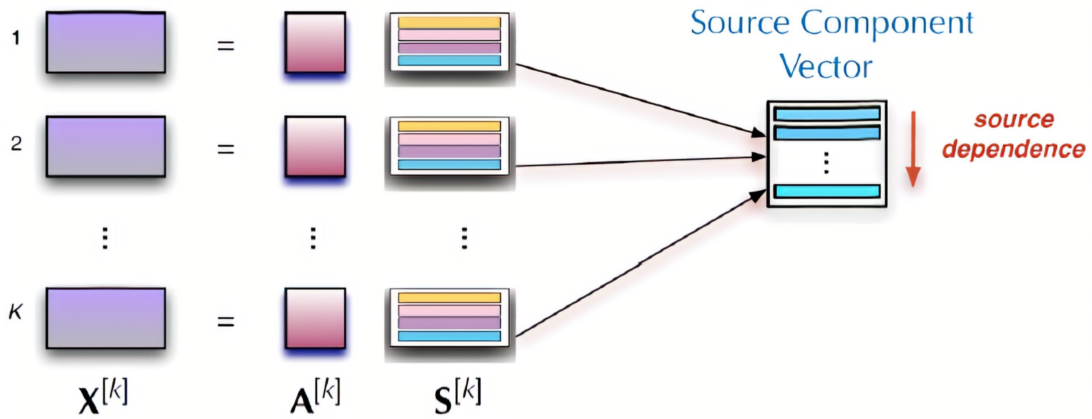

Figure 2 – JBSS structure.

In addiction, the components within each $\mathbf{s}^{[k]}$ are assumed to be independent, while at the same time, dependence across corresponding components of $\mathbf{s}^{[k]}$ in multiple datasets are allowed. This comes from the definition of the source component vector (SCV) that is defined by vertically concatenating the nth source from each of the K datasets and is denoted by

$$\mathbf{s}_n = \left[s_n^{[1]},\ldots,s_n^{[k]}\right]^\top, \tag{2.2}$$

where $\mathbf{s}_n$ is a K-dimensional random vector. An illustration of SCV is shown in Figure 2. In contrast to ICA, the goal in IVA is to estimate K demixing matrices to yield maximally independent source estimates $\mathbf{y}^{[k]} = \mathbf{W}^{[k]}\mathbf{x}^{[k]}$.

2.2 IVA cost function

IVA can be formulated in a ML framework. The ML objective function for IVA is given by (BOUKOUVALAS, 2018)

$$\mathscr{L}_{\text{IVA}} = \sum_{n=0}^{N} E\left\{\log p(\mathbf{y}_n)\right\} + \sum_{k=0}^{K} \log\left|\det\left(\mathbf{W}^{[k]}\right)\right|, \tag{2.3}$$

where $\mathbf{y}_n$ is the nth estimated random vector and $p(\mathbf{y}_n)$ denotes its multidimensional PDF. It has been shown by the asymptotic equipartition property, as sample size tends to infinity, the

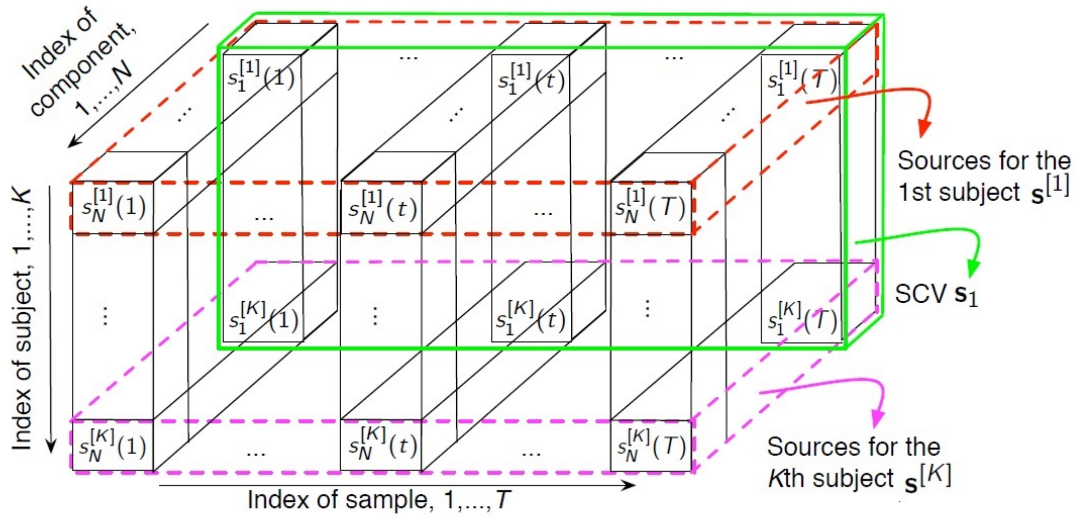

Figure 3 – Illustration of SCV and source vector.

minimization of mutual information (MI) cost function becomes equivalent to the maximization of the ML cost function (ADALI *et al.*, 2014a; BOUKOUVALAS, 2018; COVER; JOY, 2006), hence, making available all the theoretical advantages associated with the ML theory. Thus, the maximally independent sources can be achieved achieved by minimizing the MI cost function, which is given by (FU *et al.*, 2015)

$$I_{\text{IVA}} = \sum_{n=0}^{N} H(\mathbf{y}_n) - \sum_{k=0}^{K} \log \left| \det \left(\mathbf{W}^{[k]} \right) \right| - C, \tag{2.4}$$

where $H(\mathbf{y}_n)$ denotes the differential entropy of the nth SCV and C is the constant term $H(\mathbf{x}^{[1]}, \dots, \mathbf{x}^{[K]})$, and its gradient is given by (BOUKOUVALAS, 2018)

$$\frac{\partial I_{\text{IVA}}}{\partial \mathbf{W}^{[k]}} = -E\left\{ \frac{\partial \log p(y_n)}{\partial y_n^{[k]}} \frac{\partial y_n^{[k]}}{\partial \mathbf{W}^{[k]}} \right\} - \left(\mathbf{W}^{[k]} \right)^{-\top}, \tag{2.5}$$

where $p(y_n)$ denotes its probability density function (PDF). It has been shown that the minimization of (2.4) is equivalent to the maximization of the IVA ML cost function (ADALI *et al.*, 2014a; BOUKOUVALAS, 2018; COVER; JOY, 2006), hence, we may use the results from MI to handle IVA maximization. It is clear that minimizing (2.4) is not a straightforward task since there is no access to the true underlying PDF of each estimated SCV.

To mathematically demonstrate this, if $\hat{p}(\mathbf{y}_n)$ denotes the multivariate PDF of the nth estimated SCV then, its differential entropy can be expressed as

$$H(\mathbf{y}_n) = -f(p(\mathbf{y}_n) \,||\, \hat{p}(\mathbf{y}_n)) - E\{\log \hat{p}(\mathbf{y}_n)\}, \tag{2.6}$$

where $f(p(\mathbf{y}_n) \,||\, \hat{p}(\mathbf{y}_n))$ denotes the Kullback–Leibler (relative entropy) distance between the density of the nth estimated SCV and the true density of $\mathbf{y}_n$ (COVER; JOY, 2006). From (2.6), we can achieve perfect source recovery as long as the assumed model PDF matches the true latent multivariate density of the nth SCV, i.e., $f(p(\mathbf{y}_n), \hat{p}(\mathbf{y}_n)) = 0$.

2.3 IVA from an algorithmic point of view

Using the IVA MI objective function the derivative with respect to each of the demixing is given by (BOUKOUVALAS, 2018)

$$\frac{\partial I_{IVA}\left(\mathbf{W}^{[k]}\right)}{\partial \mathbf{W}^{[k]}} = E\left\{\phi^{[k]}(\mathbf{x}^{[k]})^{\top}\right\} - \left(\mathbf{W}^{[k]}\right)^{-\top}, \tag{2.7}$$

where $\phi^{[k]}$ is called score function and is given by

$$\phi^{[k]} = -\left[\frac{\partial \log p_{p_1}(y_1)}{\partial y_1^{[k]}}, \ldots, \frac{\partial \log p_{p_N}(y_N)}{\partial y_N^{[k]}}\right]. \tag{2.8}$$

Therefore, each of the K demixing matrices is updated using

$$\left(\mathbf{W}^{[k]}\right)^{\text{new}} \leftarrow \left(\mathbf{W}^{[k]}\right)^{\text{old}} - \gamma \frac{\partial I_{IVA}(\mathbf{W})}{\partial \mathbf{W}^{[k]}}, \tag{2.9}$$

where γ is the step size. However, optimization problems with matrix parameters arise in many BSS algorithms. In particular, for the IVA update rules, performing the optimization procedure on the space of all invertible matrices, may result in poor convergence due to inversion of $\mathbf{W}$ matrix at each iteration as shown in Equation (2.5). This motivates the division of the minimization of (2.4) into a series of subproblems such that instead of minimize (2.4) with respect to $\mathbf{W}^k$ we minimize the MI objective function with respect each of the row vectors $\mathbf{w}_1, \ldots, \mathbf{w}_K$ individually. This simplifies the density matching problem as the estimation of a given source will not affect the estimation of the others, improve the convergence characteristics of the algorithm, and simplifies the incorporation of constraints in the IVA framework.

A simple approach, introduced in (HYVARINEN, 1999), is to assume that $\mathbf{W}$ is orthogonal, i.e., $\mathbf{W}\mathbf{W}^{\top} = \mathbf{1}$. This assumption yields $|\det(\mathbf{W})| = 1$, and allows for the optimization of (2.4) with respect to each of the rows of $\mathbf{W}$. Although assuming $\mathbf{W}$ to be orthogonal simplifies the objective function and may improve the stability of the algorithm, the solution space is limited, which may significantly affect the overall separation performance. To avoid this issue we present a decoupling procedure that transforms the matrix optimization into a series of vector optimization problems without constraining $\mathbf{W}$ to be orthogonal.

2.3.1 Decoupling procedure

In order to achieve improved source separation performance, we use a decoupling procedure (LI; ZHANG, 2007; ANDERSON *et al.*, 2012), which instead of minimizing (2.4) with respect to $\mathbf{W}^{[k]}$, minimizes with respect to each row vector $\mathbf{w}_i^{[k]}$, $i = 1, ..., N$. In order to maintain a simple mathematical notation, we demonstrate the decoupling procedure approach within the ICA framework. It is worth mentioning that for IVA the procedure is straightforwardly formulated.

Let $\mathbf{W}_n = [\mathbf{w}_1, ..., \mathbf{w}_{n-1}, \mathbf{w}_{n+1}, ..., \mathbf{w}_N]^\top \in \mathbb{R}^{(N-1)\times N}$ denote the matrix that contains all rows of $\mathbf{W}$ except the nth term. As the determinant of a matrix is invariant under row permutation up to a sign ambiguity, the square of $\det(\mathbf{W})$ is written as

$$\begin{aligned}
\det(\mathbf{W})^2 &= \det(\mathbf{W}\mathbf{W}^\top) \\
&= \det\left(\begin{bmatrix}\mathbf{W}_n \\ \mathbf{W}_n^\top\end{bmatrix}\begin{bmatrix}\mathbf{W}_n \mathbf{w}_n^\top\end{bmatrix}\right) \\
&= \det\left(\begin{bmatrix}\mathbf{W}_n\mathbf{W}_n^\top & \mathbf{W}_n\mathbf{w}_n \\ \mathbf{w}_n^\top\mathbf{W}_n^\top & \mathbf{w}_n^\top\mathbf{w}_n\end{bmatrix}\right) \\
&= \det\left(\mathbf{W}_n\mathbf{W}_n^\top\right)\mathbf{w}_n^\top\left(\mathbf{I} - \mathbf{W}_n^\top\left(\mathbf{W}_n\mathbf{W}_n^\top\right)^{-1}\mathbf{W}_n\right)\mathbf{w}_n,
\end{aligned} \tag{2.10}$$

where $\mathbf{H}_n = \mathbf{I} - \mathbf{W}_n^\top\left(\mathbf{W}_n\mathbf{W}_n^\top\right)^{-1}\mathbf{W}_n$ is the orthogonal projection onto the null space of $\mathbf{W}_n$. Moreover, $\mathbf{H}_n = \mathbf{h}_n\mathbf{h}_n^\top$ since $\mathbf{H}_n$ is rank one by definition, where $\mathbf{h}_n$ is perpendicular to all row vectors of $\mathbf{W}_n$. Thus,

$$\begin{aligned}
|\det(\mathbf{W})| &= \sqrt{\det(\mathbf{W}_n\mathbf{W}_n^\top)^2\mathbf{w}_n^\top\mathbf{h}_n\mathbf{h}_n^\top\mathbf{w}_n} \\
&= \sqrt{\det(\mathbf{W}_n\mathbf{W}_n^\top)^2\left(\mathbf{h}_n^\top\mathbf{w}_n\right)^2} \\
&= \left|\det\left(\mathbf{W}_n\mathbf{W}_n^\top\right)\right|\left|\left(\mathbf{h}_n^\top\mathbf{w}_n\right)\right|.
\end{aligned} \tag{2.11}$$

By using (2.11), we can rewrite the MI cost function (2.4) as the following decoupled cost function

$$I_{\text{IVA}} = H\left(\mathbf{y}_n\right) - \log\left|\left(\mathbf{h}_n^{[k]}\right)^\top\mathbf{w}_n^{[k]}\right| - C_n^{[k]}, \tag{2.12}$$

and its gradient is given by

$$\frac{\partial I_{\text{IVA}}}{\partial \mathbf{w}_n^{[k]}} = E\left\{\phi_n^{[k]}\left(\mathbf{y}_n\right)\mathbf{x}^{[k]}\right\} - \frac{\mathbf{h}_n^{[k]}}{\left(\mathbf{h}_n^{[k]}\right)^\top\mathbf{w}_n^{[k]}}, \tag{2.13}$$

where the kth element of the multivariate score function can be written as

$$\phi\left(\mathbf{y}_n^{[k]}\right) = \frac{\partial \, \log\left(p_n\left(\mathbf{y}_n\right)\right)}{\partial y_n^{[k]}}. \tag{2.14}$$

The development of an efficient PDF estimator plays an important role on the IVA algorithms, since to achieve perfect source separation we need the term $f(p(\mathbf{y}_n), \hat{p}(\mathbf{y}_n))$ in (2.6) to become zero.

2.4 Summary

In this chapter, based on a maximum likelihood framework, we present a theoretic mathematical model that guides the mathematical formulation of the IVA algorithm, an efficient algorithm for the JBSS which is a generalization of the BSS problem to multiple datasets. Furthermore, as expected from multivariate approaches, the model becomes computationally challenging. To overcome this challenge, we use a decoupled procedure as a mathematical tool yielding several benefits to our approach and we guide the importance of developing an efficient PDF estimator which is provided in the next chapter.

3 MULTIVARIATE DENSITY ESTIMATION

The estimation of a multivariate probability density function (PDF), as observed in Chapter 2, is a key part of the IVA model. However, this information is usually not available in most real-world applications. Algorithms that use fixed or simple models for the underlying distribution of the latent sources can yield poor separation performance when the data deviates from the assumed model. Thus, effective characterization of the density is vital to the success of these approaches.

In this chapter, we present a multivariate density estimator that is based on the MEP and by jointly using global and local measuring functions we provide flexible PDFs. Moreover, in order to keep the complexity low we integrate into the proposed algorithm a multidimensional Monte-Carlo (MC) integration method. Finally, we demonstrate superior density estimation performance compared to widely used algorithms. We call the new estimator multivariate entropy maximization with kernels (M-EMK) and consider it as an extension of the univariate case (FU *et al.*, 2015).

3.1 Multivariate density estimation techniques

Multivariate density estimation approaches can be broadly classified as either parametric and non-parametric. Parametric methods, such as the multivariate Gaussian mixture model (GMM) (MCLACHLAN; D.PEEL, 2004), provide a simple form for the PDF and are computationally efficient, however they are limited when the underlying distribution of the data deviates from the assumed parametric form. On the other hand, non-parametric methods, such as multivariate kernel density estimation (KDE) (KUNG, 2014; ROJO-ÁLVAREZ *et al.*, 2018), can provide flexible density matching since they are not limited to any specific distribution. However, they are generally computationally demanding, especially when sample size is large, and they highly depend on the choice of tuning parameters. Semi-parametric methods combine the flexibility of the non-parametric techniques with the relatively simple density form of the parametric technique. Semi-parametric methods, such as those based on the maximum entropy principle (MEP) (BEHMARDI *et al.*, 2011; FU *et al.*, 2015; LI; ADALI, 2010), provide a desirable trade-off between non-parametric and parametric methods, yielding a global solution provided by the MEP (JAYNES, 1957a).

3.2 Maximum Entropy Principle

The maximum entropy principle states that the probability distribution which best represents the current state of knowledge is the one with the largest entropy (JAYNES, 1957a; JAYNES, 1957b). The maximum entropy density, subject to known constraints, can be written as the following optimization problem (COVER; JOY, 2006):

$$\begin{aligned} &\max_{p(\mathbf{y})} \; H(p(\mathbf{y})) = -\int_{\mathbb{R}^K} p(\mathbf{y}) \log p(\mathbf{y}) d\mathbf{y} \\ &\text{s.t.} \quad \int_{\mathbb{R}^K} r_i(\mathbf{y}) p(\mathbf{y}) d\mathbf{y} = \alpha_i, \text{ for } i = 0, ..., M, \end{aligned} \tag{3.1}$$

where $\mathbf{y} \in \mathbb{R}^K$, $p(\mathbf{y}) \geq 0$, and $r_i(\mathbf{y}) \in C(\mathbb{R}^K, \mathbb{R})$ for $i = 0, ..., M$. The space $C(\mathbb{R}^K, \mathbb{R})$, represents all measuring functions with domain $\mathbb{R}^K$ and co-domain $\mathbb{R}$, and $\alpha_i = \sum_{t=1}^{T} r_i(\mathbf{y}_t)/T$ for $i = 0, ..., M$ represent their corresponding sample averages, given observations $\mathbf{y}(t) \in \mathbb{R}^K$, $t = 1, ..., T$. We note that the first constraint needs to be $\int_{\mathbb{R}^K} p(\mathbf{y}) d\mathbf{y} = 1$, equivalently $r_0 = 1$ and $\alpha_0 = 1$, in order $p(\mathbf{y})$ to be a valid PDF. The optimization problem in (3.1) can be rewritten in a Lagrangian form and is given by

$$\begin{aligned} \mathscr{L}(p(\mathbf{y})) = &- \int_{\mathbb{R}^K} p(\mathbf{y}) \log p(\mathbf{y}) d\mathbf{y} \\ &+ \sum_{i=0}^{M} \lambda_i \int_{\mathbb{R}^K} (r_i(\mathbf{y}) - \alpha_i) p(\mathbf{y}) d\mathbf{y}, \end{aligned} \tag{3.2}$$

where λ_i, $i = 0, ..., M$, are the Lagrangian multipliers. Through the use of functional variation, we can “differentiate” (3.2) with respect to $p(\mathbf{y})$. By setting $\partial \mathscr{L} p(\mathbf{y}) / \partial p(\mathbf{y}) = 0$, we obtain the equation of maximum entropy distribution,

$$\hat{p}(\mathbf{y}) = \exp\left\{ -1 + \sum_{i=0}^{M} \lambda_i r_i(\mathbf{y}) \right\}, \tag{3.3}$$

where Lagrangian multipliers are chosen such that p satisfies the constraints in (3.1). By substituting (3.3) into the constraints in (3.1), we generate a nonlinear system of $M+1$ equations for the $M+1$ Lagrangian multipliers.

3.3 M-EMK

3.3.1 Mathematical formulations

We can evaluate the Lagrangian multipliers in (3.3) by the Newton iteration scheme, given by (FU *et al.*, 2015)

$$\boldsymbol{\lambda}_{n+1} = \boldsymbol{\lambda}_n - \mathbf{J}^{-1} E_{p_n}\{\boldsymbol{r} - \boldsymbol{\alpha}\}, \tag{3.4}$$

where p_n is the estimated PDF for the nth iteration, E_{p_n} is the expectation over the distribution p_n, and $\mathbf{r} = [r_0, ..., r_M]$, $\boldsymbol{\lambda} = [\lambda_0, ..., \lambda_M]$, $\boldsymbol{\alpha} = [\alpha_0, ..., \alpha_M] \in \mathbb{R}^{M+1}$ denote the vectors of global and local measuring functions, the Lagrangian multipliers and sample averages, respectively. By $\mathbf{J} \in \mathbb{R}^{M \times M}$ we denote the Jacobian matrix where the ij-th entry of $\mathbf{J}$ is given by

$$\mathbf{J}_{ij} = \int_{\mathbb{R}^K} r_i(\mathbf{y}) r_j(\mathbf{y}) p(\mathbf{y}) d\mathbf{y} = E_{p_n}\left\{r_i r_j\right\}. \tag{3.5}$$

The i-th entry of $E_{p_n}\{\boldsymbol{r} - \boldsymbol{\alpha}\}$ is given by

$$E_{p_n}\{\boldsymbol{r} - \boldsymbol{\alpha}\} = \int_{\mathbb{R}^K} r_i(\mathbf{y}) p(\mathbf{y}) d\mathbf{y} - \alpha_i. \tag{3.6}$$

As we can see from (3.6), accurate estimation of the Lagrange multipliers highly depends on the proper selection of the constraints both in terms of their number as well as the different types of measuring functions that provide information about the underlying statistical properties of the data. Failure to do so, may result in high complexity as well as poor data characterization.

3.3.2 Measuring functions

We jointly use global and local constraints to provide flexible multivariate density estimation while keeping the complexity low. Similar to (FU *et al.*, 2015; LI; ADALI, 2010) we use $\mathbf{1}, \mathbf{y}, \mathbf{y}^2, \mathbf{y}/(\mathbf{1}+\mathbf{y}^2)$ as the global constraints, since they provide computational efficiency and desirable performance for a wide range of distributions. Furthermore, these global constraints provide information on the PDF's overall statistics, such as the mean, variance, and higher order statistics (HOS). For the local constraint we use the following Gaussian kernel, given by

$$q(\mathbf{y}) = \frac{1}{\sqrt{|\Sigma|(2\pi)^K}} \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right), \tag{3.7}$$

where $\boldsymbol{\mu}$ denotes the mean vector, Σ denotes the covariance matrix and $|\cdot|$ stands for the determinant of the argument. The use of Gaussian kernels provides localized information about the PDF, This localized information plays an important role in the estimation procedure since we do not have access to the underlying true PDF local behavior only using global constraints. It is important to note that when we add the Gaussian kernel to the multidimensional integrals in (3.5) and (3.6), their computation becomes challenging due to the fact that the Gaussian kernel has infinite support.

3.3.3 Multidimensional integration

As noted in the previous section, the multidimensional integration is one of the main challenges in our estimation problem. To overcome this difficulty, we make use of an efficient multidimensional integration technique that is based on Quasi-Monte Carlo (QMC) methods. QMC methods are variants of the classical Monte Carlo (MC) methods and have shown to be efficient in terms of their rate of convergence becoming ideal for our proposed approach (O'LEARY, 2009). QMC methods achieve a convergence rate of order $O((\log T)^K/T)$ or faster if sufficient smoothness of the function is assumed (DICK *et al.*, 2013).

In order to introduce the QMC integration methods in our approach, we need to generate a sequence of quasi-random points which presents an advantage in terms of convergence rate compared to sequences of pseudo-random point (NIEDERREITER, 1992) that MC integration methods use. To achieve this, we use the van der Corput sequence, which is an example of a one-dimensional low-discrepancy sequence that uniformly covers the unit hypercube and can be constructed using a computational efficient procedure (NIEDERREITER, 1992). The z-th quasi-random number w_z, is constructed in the following way:

1) Let b_k denote the k-th prime number, for instance, when $k = 1$ then $b_1 = 2$, when $k = 2$ then $b_1 = 3$ and so forth, and $Z_{b_k} = \{0, 1, ..., b_k - 1\}$ denotes the least residue system mod

i	Naive Sequence	Binary	Reverse Binary	Van der Corput	Points in $[0,1]$
1	0	.0000	.0000	0	
2	1/16	.0001	.1000	1/2	
3	1/8	.0010	.0100	1/4	
4	3/16	.0011	.1100	3/4	
5	1/4	.0100	.0010	1/8	
6	5/16	.0101	.1010	5/8	
7	3/8	.0110	.0110	3/8	
8	7/16	.0111	.1110	7/8	
9	1/2	.1000	.0001	1/16	
10	9/16	.1001	.1001	9/16	
11	5/8	.1010	.0101	5/16	
12	11/16	.1011	.1101	13/16	
13	3/4	.1100	.0011	3/16	
14	13/16	.1101	.1011	11/16	
15	7/8	.1110	.0111	7/16	
16	15/16	.1111	.1111	15/16	

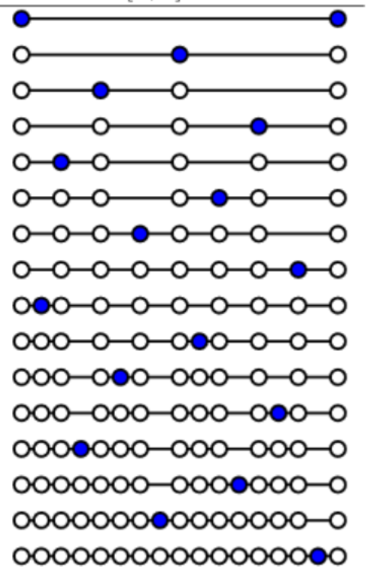

Figure 4 – Example of a generation of 16 samples in $[0,1] \in \mathbb{R}$. The van der Corput sequence is obtained by reversing the bits in the binary decimal representation of the naive sequence.

b_k. Every integer $t \geq 0$ has a unique digit expansion given by

$$t = \sum_{i=0}^{T-1} a_i(t) b_k^i \tag{3.8}$$

in base-b_k, where T is the sample size of quasi-random numbers and $a_i(t) \in Z_{b_k}$. The t-th quasi-random number is given by the following radical-inverse function in base-b_k,

$$w_t = \sum_{i=0}^{T-1} a_i(t) b_k^{-i-1}. \tag{3.9}$$

2) Using the numbers generated by (3.9), we approximate the multidimensional integrals in (3.5) and (3.6) in a similar manner as we do using traditional MC methods. Thus, each of the integrals is evaluated by

$$Q_{T,K}\left(p(\mathbf{y})\right) = \Omega \left(\frac{1}{T} \sum_{i=0}^{T-1} p(w_t) \right), \tag{3.10}$$

where Ω denotes the dimensional measure of the region of integration. For instance, length, area and volume for one, two and three-dimensional space, respectively.

3.4 Experimental results

In this section, by using simulated data, we first demonstrate how the performance of the density estimation is affected if we include only global measuring functions. Then, we demonstrate superior estimation performance, computational efficiency, and flexibility of M-EMK by comparing with KDE, which is a widely used non-parametric method, and GMM, which is classical parametric method. Finally, we show the estimation capability of M-EMK using several real datasets.

3.4.1 Simulated data

We generate data according to a mixture of generalized Gaussian distributions. The PDF is given by

$$p(\mathbf{x};\beta,\mu,\sigma_i) = \sum_{i=1}^{L} \pi_i g_i(\mathbf{x};\beta,\mu,\sigma_i), \quad \mathbf{x} \in \mathbb{R}^K$$

where each g_i is a multivariate generalized Gaussian distribution defined in (ANDERSON *et al.*, 2012). The shape and mean parameters for each of the components are chosen to be $\beta = 0.5$ and $\mu = 1$ respectively. The weight parameters π_1 and π_2 are chosen to be equal to 0.3 and 0.7, respectively. For all the experiments we select $K = 2$, also we set the mean of the Gaussian kernel equal to the zero vector and equal to the identity matrix.

3.4.2 Density estimation performance

For the first experiment, we demonstrate how the performance of the density estimation is affected if we include only global measuring functions and if we jointly include global and local measuring functions into the estimation phase. For this experiment the number of sample size is $T = 10000$. Figure 5, shows the histogram of the generated data as well as the estimated density by using only global measuring functions and the estimated density by jointly using global and local measuring functions.

We can see that by only including global measuring functions we are not able to capture the local behavior of the true underlying PDF and thus we achieve sub-optimal estimation performance in terms of matching with the histogram.

In the second experiment, we verify the effectiveness of our approach by comparing

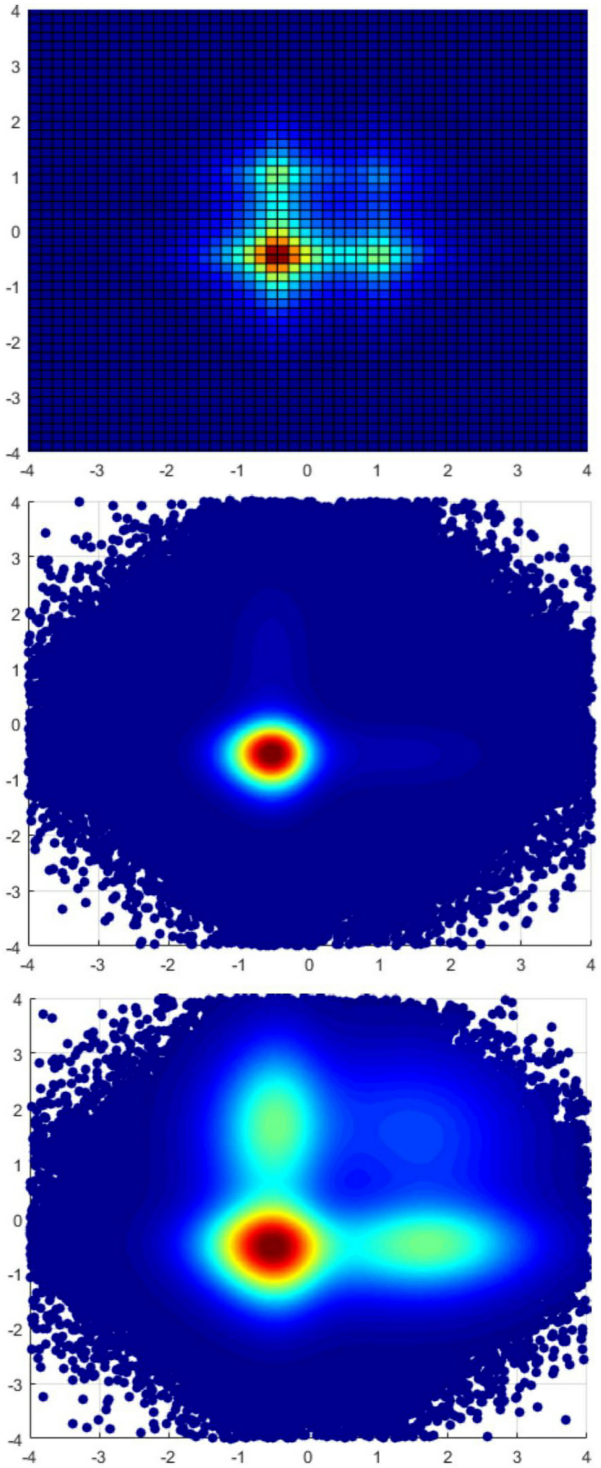

Figure 5 – Estimation performance only using global constraints and using global and local constraints represented in the second and third line, respectively.

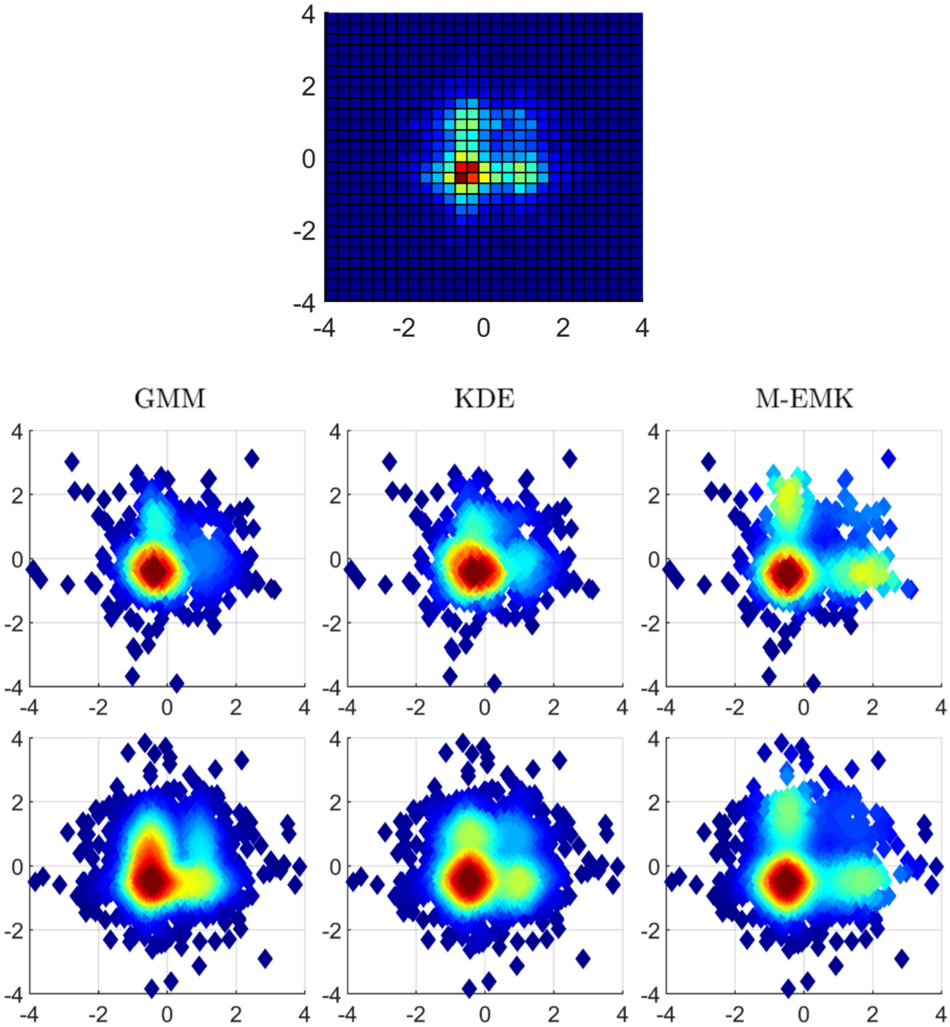

Figure 6 – Comparison in terms of matching with the histogram. First line–histogram of the generated data, first column–parametric method (GMM), second column–non-parametric method (KDE), third column–semi-parametric method (M-EMK).

M-EMK with KDE as well as GMM in terms of matching with the histogram of the generated data for different number of sample sizes. As we can see in Figure 6, when the number of sample size is $T = 1000$, M-EMK performs similarly to GMM and KDE in terms of matching with the histogram. However, when $T = 500$, GMM and KDE are not able to capture the details of the shape of all three peaks while, M-EMK is able to effectively estimate the three highest peaks.

In addition to the estimation capability, another important aspect that we examine is the computational efficiency in terms of the CPU time of M-EMK when compared to KDE, GMM, and M-EMK by using only global constraints. Data are generated in the same way as in the first experiment, by varying the number of sample size. From Figure 7, we can see that M-EMK provides the best performance in terms of CPU time when compared to KDE and GMM. As the number of sample size increases, KDE becomes computationally demanding making it

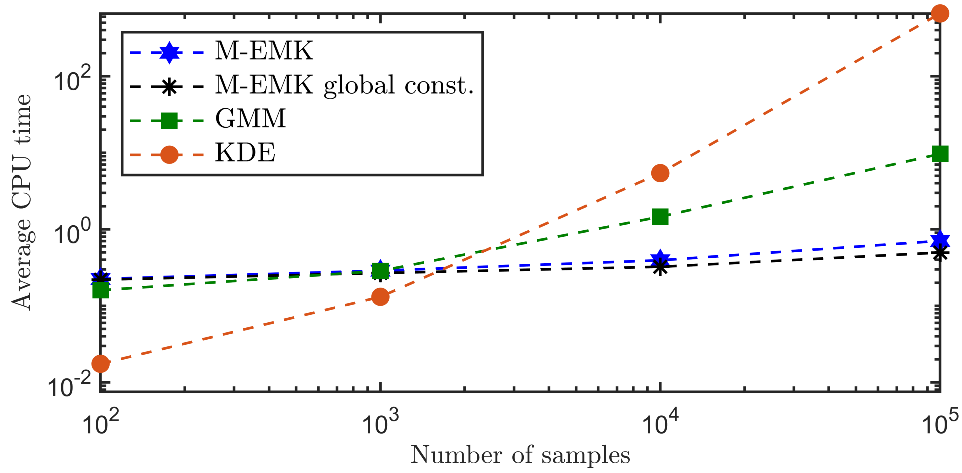

Figure 7 – Performance comparison in terms of CPU time.

impractical for high dimensional data applications. The low CPU time for M-EMK, has been observed due to the fast convergence rate introduced by the efficient integration technique. It is worth mentioning that adding local constraints to our approach does not have a significant impact to the average CPU time as number of sample size increases, verifying the low computational complexity of M-EMK.

Moreover, in order to show the flexibility of our approach, we submit our estimator to a real world dataset. The dataset represents the relation between the wind speed and power generated by a wind turbine. These collected data yield performance monitoring and an appropriate analysis of the energy generation. In Figure 8, we can see that M-EMK is able to capture the details of the dataset.

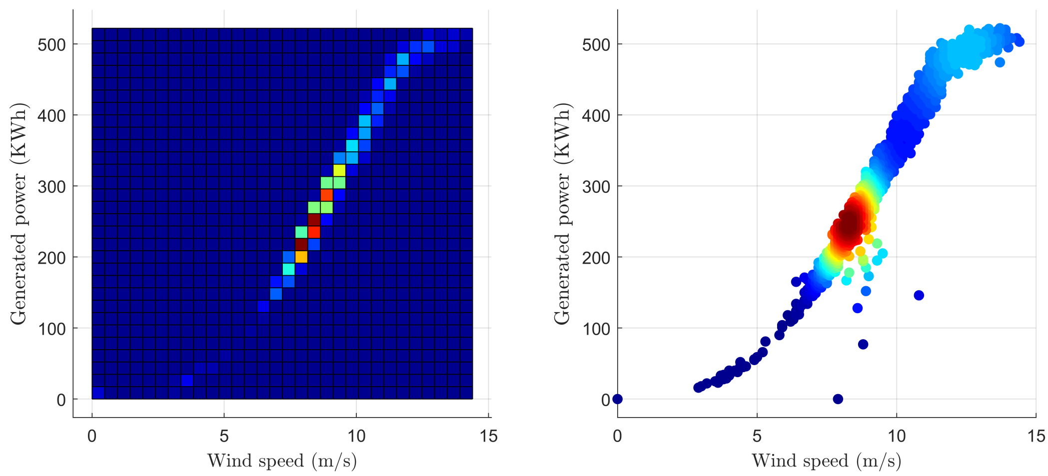

Figure 8 – Estimation performance in terms of matching with the histogram. First column–histogram of the wind turbine data, second column–M-EMK.

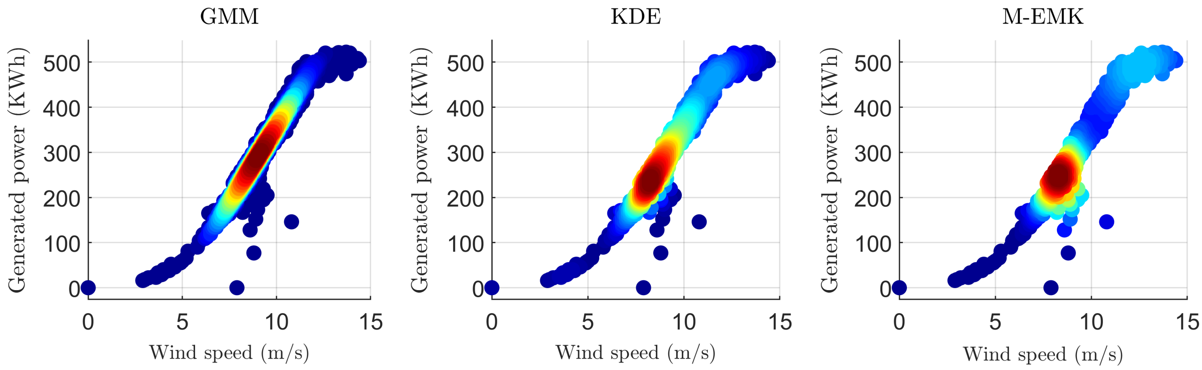

Figure 9 – Density estimation comparison. First column– parametric method (GMM), second column–non-parametric method (KDE), third column–semi-parametric method (M-EMK).

These details are clearly noticeable when we compare our method to the other estimation methods, GMM and KDE. We can see from Figure 9, that both the parametric and non-parametric methods are not able to capture the two peaks while, M-EMK is able to efficiently estimate the two peaks.

3.5 Summary

In this chapter, we introduced the estimator multivariate entropy maximization with kernels (M-EMK), a new multivariate PDF estimation technique using the maximum entropy principle. By jointly using global and local constraints functions, M-EMK enjoys a high level of flexibility providing a simple exponential form for multivariate PDFs while keeping the complexity low. Next, we present a efficient multidimensional integration technique which yields several computational benefits to the estimation problem. Finally, we show by experiments that M-EMK yields a very effective density estimation in terms of matching with the histogram. Therefore, in the next chapter, based on the success of the proposed multivariate density estimation technique, we derive a novel IVA algorithm using the M-EMK.

4 MULTIVARIATE EMK IVA ALGORITHM

In this chapter, we propose a new IVA algorithm based on the efficient multivariate density estimation technique presented in the previous chapter. We make use of M-EMK to derive a novel IVA algorithm, IVA by multivariate entropy maximization with kernels (IVA-M-EMK) through the use of the flexible and efficient estimation capability of M-EMK to greatly improve separation performance. Then, we demonstrate the superior performance of the new IVA algorithm numerically using simulated data.

4.1 IVA by entropy maximization

By using (3.3) as the maximum entropy PDF and the Lagrange multiplier estimates from using (3.4) provided by M-EMK, the kth element of the score function (2.14) can be rewritten as

$$\phi\left(\mathbf{y}_n^{[k]}\right) = -\sum_{i=0}^{M} \lambda_i \frac{\partial r_i\left(\mathbf{y}_n\right)}{\partial y_n^{[k]}}. \tag{4.1}$$

and its gradient update rule is given by

$$\frac{\partial I_{\mathrm{IVA}}}{\partial \mathbf{w}_n^{[k]}} = -\sum_{i=0}^{M} \lambda_i \frac{\partial r_i\left(\mathbf{y}_n\right)}{\partial y_n^{[k]}} E\left\{\mathbf{x}^{[k]}\right\} - \frac{\mathbf{h}_n^{[k]}}{\left(\mathbf{h}_n^{[k]}\right)^{\top} \mathbf{w}_n^{[k]}}. \tag{4.2}$$

Thus, with the mutual information function to be minimized by the gradient update rule and the efficient multivariate density estimation technique in our hands, we are able to proceed to the simulated experiments.

4.2 Experimental results

In this set of experiments, we show the effectiveness of the IVA-M-EMK algorithm by comparing its performance with six widely used IVA algorithms, These include, IVA-Gaussian (IVA-G) (ANDERSON *et al.*, 2012; VIA *et al.*, 2011) that uses second-order statistics, and does not constrain the demixing matrices to be orthogonal. The IVA-GGD (ANDERSON *et al.*, 2013a) and IVA-A-GGD (BOUKOUVALAS *et al.*, 2015) are based on the multivariate generalized Gaussian distribution model, and take all order statistical information into account. The IVA-Laplacian (IVA-L), and its extensions IVA-L-Decp, IVA-L-SOS (KIM *et al.*, 2006;

KIM *et al.*, 2007), use higher-order statistics to estimate the demixing matrices assuming that the sources are multivariate Laplacian distributed. We compare them in terms of the the CPU time and in terms of the joint inter-symbol-interference (ISI) as defined in (ANDERSON *et al.*, 2014). Joint ISI is a global metric for performance evaluation when the ground truth is available where zero ISI indicates perfect separation.

4.2.1 Source separation performance

For the following experiments, we consider two cases when generating the data for the SCVs. For the first case, we have $K = 2$ and generate three SCVs where each SCV is a mixture of MGGD sources where β, μ are chosen from the range $(0.5, 1)$ and $(0.5, 10)$ respectively.

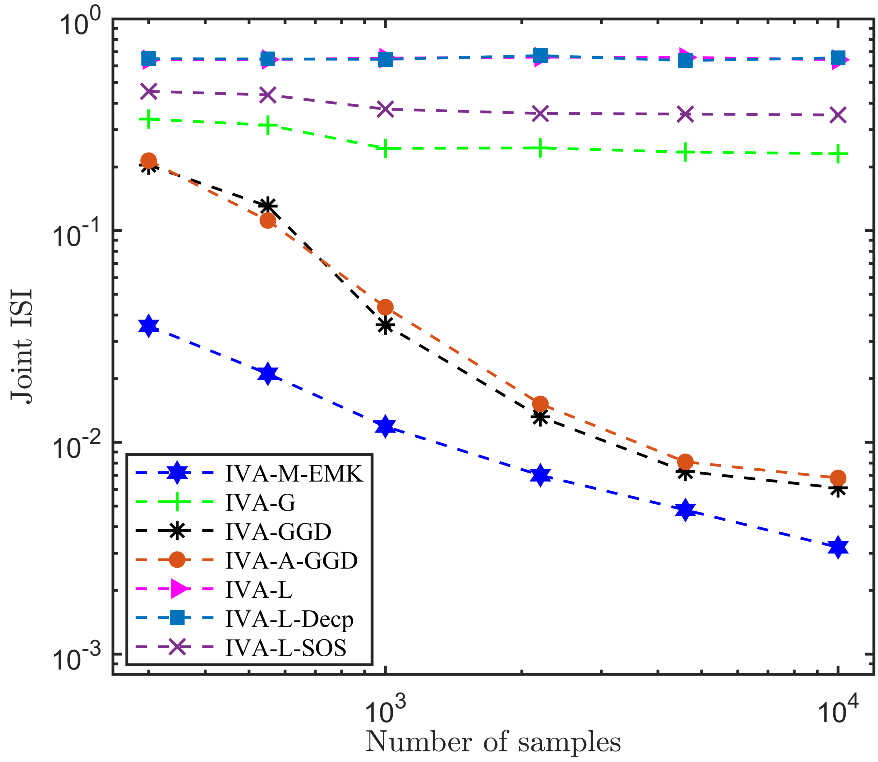

Figure 10 – Performance comparison of the first case in terms of Joint ISI

As we can see from Figure 10, IVA-M-EMK performs the best among the seven algorithms in terms of Joint ISI as function of sample size demonstrating its efficient applicability. For the second case, we have, $K = 3$, and generate one unimodal MGGD SCV where the shape parameter and the correlation within the SCV for each dataset are chosen to be $\beta = 3$ and $\mu = 0.6$, and a mixture of two MGGD sources where $\beta \in (0.6, 0.8)$ and $\mu \in (5, 10)$ respectively.

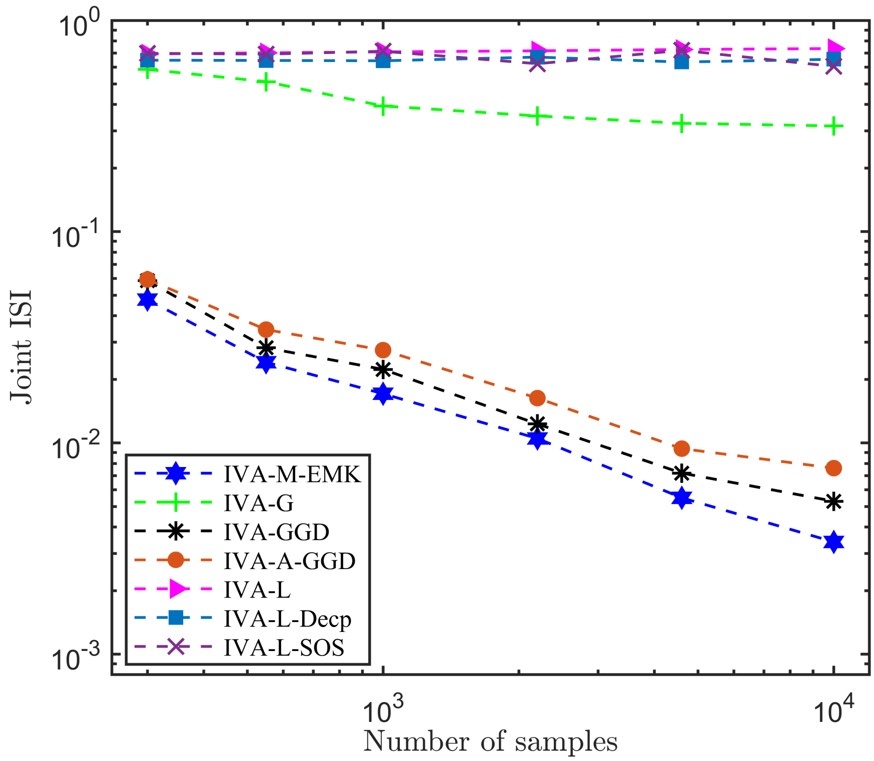

Figure 11 – Performance comparison of the second case in terms of Joint ISI.

From Figure 11, we see that IVA-GGD and IVA-A-GGD provide a desirable performance for the second case in terms of Joint ISI as function of sample size revealing the flexibility of their underlying density models. Conversely, the algorithms based on IVA-L and IVA-G do not provide a desirable performance due to their Laplacian and Gaussian distribution assumption for the underlying sources. Overall however, IVA-M-EMK performs the best in both cases among the seven IVA algorithms.

In addition to the source separation capability, another important aspect that we examine is the computational efficiency in terms of the CPU time. In Figure 12, among the algorithms that use a simple underlying density model, IVA-G provides the best performance for both cases. This is due to the assumption of Gaussian distribution for the underlying sources, which simplifies the gradient of the IVA objective function and makes the Hessian positive definite thus, enabling second-order algorithms to improve the quality of convergence. On the other hand, as it is expected, IVA-M-EMK is more computationally expensive, however, we see that as number of samples increase the increase in average CPU time is negligible.

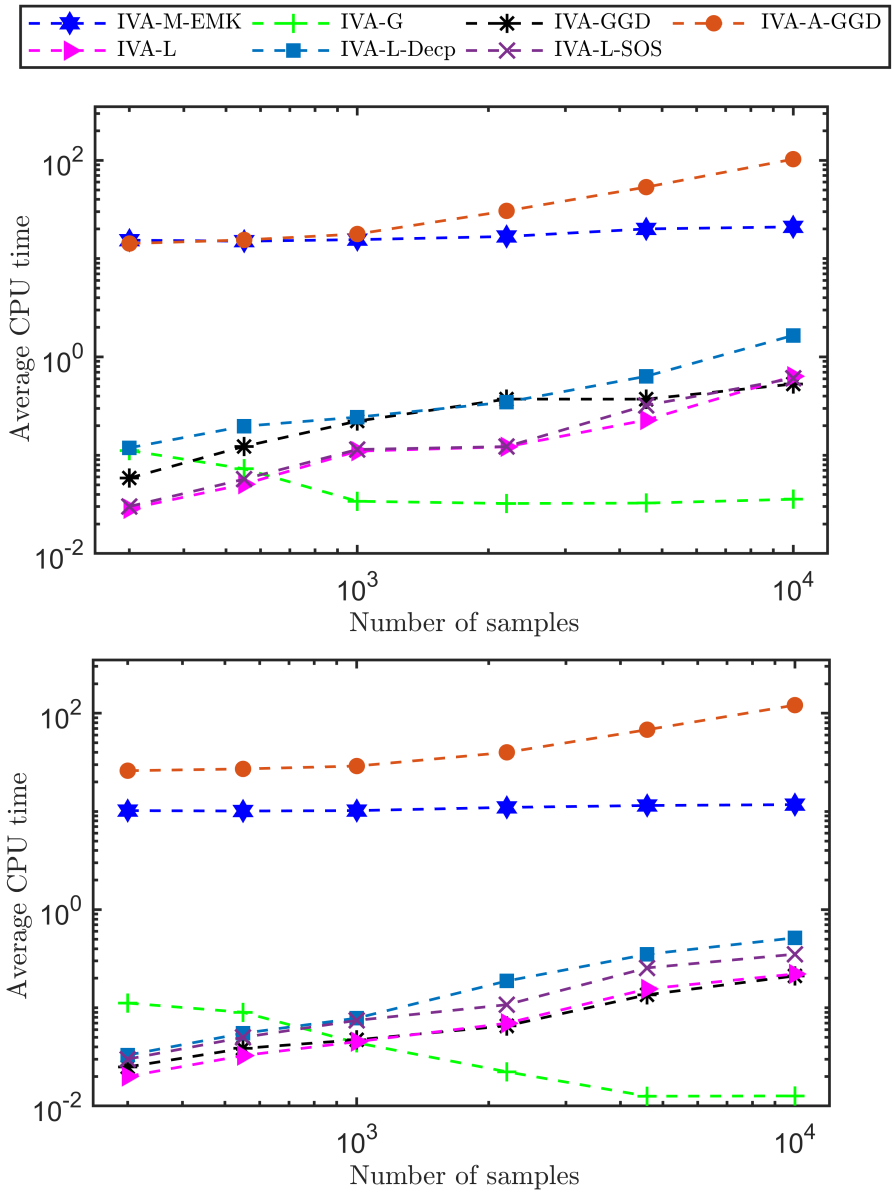

Figure 12 – Performance comparison in terms average CPU time for different number of sample size. The first and second case represented in the first and second columns, respectively.

4.3 Summary

In this chapter, we use the new multivariate density estimator presented in chapter 3 to derive an efficient IVA algorithm, IVA-M-EMK, that accurately separates sources from a wide range of multivariate PDFs, enabling the application of IVA to many practical applications where such multivariate density modeling is needed. We show by experiments that IVA-M-EMK performs the best among widely used algorithms in terms of separation performance and CPU time in different scenarios.

5 CONCLUSIONS AND FUTURE DIRECTIONS

In this chapter, we present the conclusions of this dissertation, as well as future directions for further research.

5.1 Summary

In this thesis, we introduced the multivariate entropy maximization with kernels (M-EMK) algorithm based on a new multivariate PDF estimation technique using the maximum entropy principle. By jointly using global and local constraints functions, M-EMK enjoys a high level of flexibility while providing a simple exponential form for PDFs in general.

Using M-EMK, we derived an efficient IVA algorithm that accurately separates sources from a wide range of multivariate PDFs, enabling the application of IVA to many practical applications where such multivariate density modeling is needed

Using simulated as well as real-world datasets we demonstrate how M-EMK yields a very effective density estimation in terms of matching with the histogram, and IVA-M-EMK performs better than several used algorithms.

5.2 Future directions

One of the main contributions of this dissertation is the development of an effective and efficient IVA algorithm. The success of IVA-M-EMK raises several interesting questions that can be explored in future work. Therefore, we plan to expand IVA and its capabilities into two major directions: algorithmic developments, as well as their use in novel natural language processing (NLP) tasks.

5.2.1 Misinformation detection

Due to the advancement of social media, the spread of information, as well as misinformation, plays an important role in society, particularly during high impact events, for instance, pandemics, natural disasters, and presidential election periods. This has been shown since the past year with Coronavirus Disease (COVID-19), where misinformation generates chaos due to the propagation of harmful health advice, fake vaccine schedules, and conspiracy theories, to name a few. Recent machine learning techniques have been widely used and show promising advances to detection of misinformation (JAIN *et al.*, 2016; WU *et al.*, 2019; YU *et al.*, 2017). In order to machine learning algorithms to be effective solutions for detection of misinformation, they must have explainability and interpretability, i.e., the ability to summarize in a reasonable way its decisions in an effective manner, which yields gain of trust for its users and understandable results for analysts. Furthermore, low computational cost and a good generalization model are desirable. Some of the most promising approaches are based on deep learning (GIRGIS *et al.*, 2018; S.SINGHANIA *et al.*, 2017). They have demonstrated superior performance in many machine learning tasks, however, the interpretation of their results is not direct or easily accessible, and their high computational cost becomes a watershed in the selection of an effective and flexible machine learning algorithm. In contrast to deep learning models, a recent algorithm based on the ICA model has shown promising results in detection of misinformation during COVID-19 (BOUKOUVALAS *et al.*, 2020). This motivates the formulation of the misinformation detection problem as a JBSS approach.

Therefore, based on the success of the IVA-M-EMK for the JBSS problem, as future work, we plan to develop an effective JBSS approach through the use of the IVA-M-EMK, in order to achieve detection of misinformation from multi-modal data, such as textual information, videos, images, hashtag topics, user references, comments, and repostings. Furthermore, under the JBSS and NLP umbrella, we will be able to develop algorithms for several multi-modal machine learning tasks yielding explainable and interpretable results, effective solutions at a reasonable computational cost, and good generalization ability.

5.2.2 Algorithmic development for misinformation detection

With its well-structured formulation, IVA provides an ideal starting point for developing a data fusion method that allows fusion of multi-modal data. Through the estimation of the underlying SCVs it can capture unique characteristics of multi-modal data that can be used to enhance the performance of a machine learning task. For instance, this idea has been demonstrated by the results presented in (BOUKOUVALAS *et al.*, 2018b) where true fusion was used to exploit the underlying complementary information contained in different molecular featurization methods. As seen from Figure 13, true fusion rather than simply concatenating different molecular featurization methods resulted in enhanced prediction performance of a regression model. In this development, the key steps to fully leverage the power of our true fusion framework for our application domain include (i) further development of efficient estimation

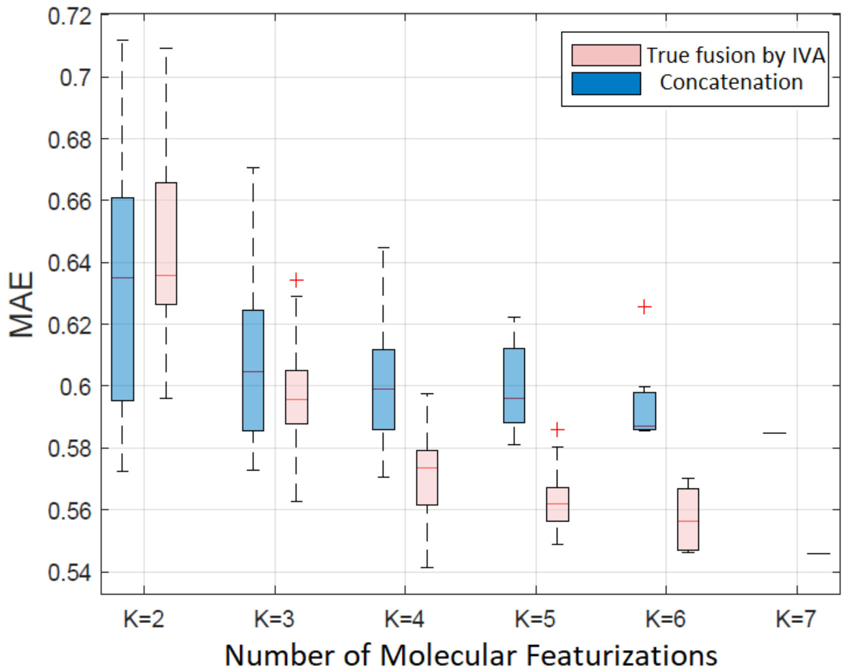

Figure 13 – Mean absolute error (MAE) as a function of number fused datasets. True fusion by IVA performs better than concatenating molecular featurization methods.

techniques for the underlying SCV densities, (ii) incorporation of prior knowledge into the IVA model through constraints, (iii) evaluation of the representation space spanned by the SCVs such that it becomes meaningful for our application domain.

5.2.2.1 Incorporation of prior knowledge and definition of common and distinct sub-spaces for misinformation detection

To motivate the incorporation of prior knowledge into the IVA model and to provide a physical interpretation of the *common* and *distinct* subspaces, one can think of the fusion example illustrated in Figure 14. For a collection of V tweets during a high impact event, we generate K multi-modal datasets. Example of multi-modal datasets includes textual features such as bag-of-words or n-grams (FÜRNKRANZ, 1998), references (@) and hashtags (#), as well as specific social content information such as number of times a particular times a tweet has been shared, number of followers and followees among others. IVA relies on the assumption of statistical independence of the latent variables. Although this might be a natural assumption in many problems, in our application, it might be too strong an assumption. Incorporation of reliable and meaningful prior information about the problem and the data can help relax the independence assumption, resulting in a better model match. This will result in better estimation of the SCV, and thus its corresponding estimated covariance matrix will better reveal associations

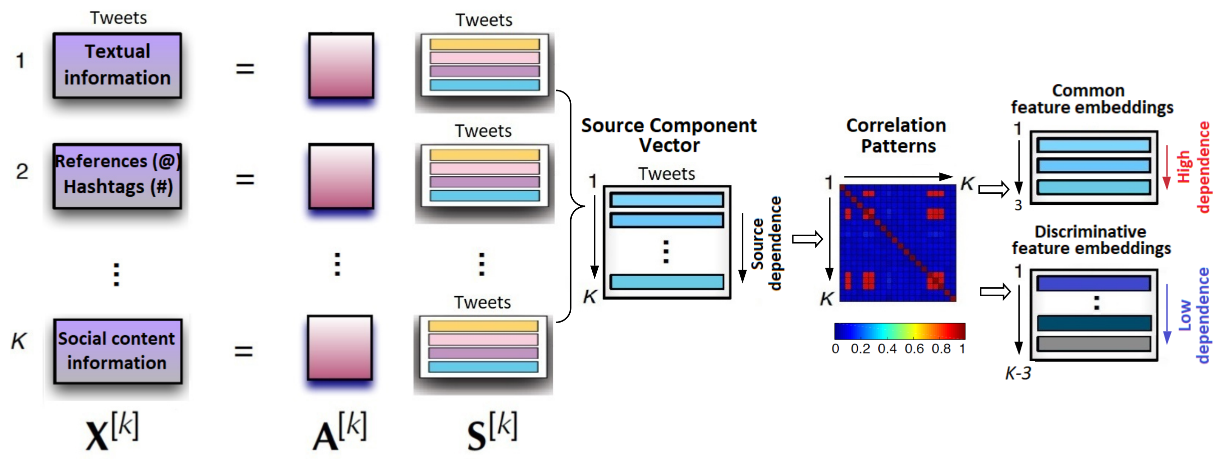

Figure 14 – Graphical multi-modal fusion example using IVA and its capabilities. Through the estimation of the source component vectors and their associated covariance matrices we identify correlation patterns among different multi-modal datasets and define common and distinct subspaces through the estimated feature embeddings. Common subspaces will enable the knowledge discovery and distinct subspaces will enable the detection of misinformation.

between extracted low dimensional feature embeddings across the modalities. This will enable the definition of common and distinct sub-spaces spanned by those feature embeddings. By accessing the multivariate correlation coefficients on the common sub-space for each estimated SCV and through a simple back-reconstruction stage (BOUKOUVALAS *et al.*, 2020), we can study which types of words, word pairs, references or hashtags, or social content characteristics are most significant to the joint representation space, and how these are associated with each other. This will enable us to discover how true and false information correlate during a high impact event. Moreover, features that span the distinct sub-spaces can be used to train a model for detection of misinformation. We plan to incorporate prior knowledge into the IVA model by using the following types of constraints.

1. Sparsity constraints Sparsity is a widely used form of prior information and typically implies that most of the energy of the data distribution is contained in only a few of the coefficients (HURLEY; RICKARD, 2009). We motivate the incorporation of sparsity into the IVA model through the fusion example illustrated in Figure 14. The data in each multi-modal dataset is sparse in nature since each term in the datasets is expected to have a value greater than zero in only few posts or tweets. Within these multi-modal sparse datasets, multiple highly correlated feature embeddings may appear, dependent on a set of feature embeddings in each of the other data sets. Hence, the estimated feature embeddings should be regrouped so as to form dependent subspaces. Motivated by (BOUKOUVALAS *et al.*, 2018), we propose to start with the IVA MI cost function

$$I_{\text{IVA}} = \sum_{n=0}^{N} H(\mathbf{y}_n) - \sum_{k=0}^{K} \log \left| \det \left(\mathbf{W}^{[k]} \right) \right| - C, \tag{5.1}$$

that accounts for statistical dependence across multiple datasets and incorporate prior information about the problem through regularization terms such as sparsity constraints or through the inclusion of sparse density models into (5.1).

A classical way to impose sparsity is by the penalization of (5.1) by the cardinality of the support of the weight vector that needs to be estimated. However, this leads to hard combinatorial problems. Motivated by (CANDES; TAO, 2005; SCHMIDT *et al.*, 2007; TIBSHIRANI, 1996; BOUKOUVALAS *et al.*, 2018), we propose to replace the cardinality of the support with the ℓ^1-norm and thus obtain estimators as solutions of convex programs. This approach has two optimization benefits. First, it leads to efficient estimation such as proximal algorithms (BOYD; VANDENBERGHE, 2004), which have drawn attention due to their desirable convergence properties and their ability to deal effectively with nonsmooth convex problems (NESTEROV, 2013). Second, it allows to answer questions related to estimation consistency and prediction efficiency (BICKEL *et al.*, 2009; NEGAHBAN *et al.*, 2012).

To identify common and distinct subspaces and to discover meaningful interactions among the variables within an SCV, we propose to use probabilistic graphical models that allow capturing of the conditional independence relationships between variables. A standard approach is to choose the sparsest network, i.e., the precision matrix—inverse covariance matrix. We will achieve this by solving the IVA regularized maximum likelihood problem with the ℓ^1-norm regularization term for the precision matrix.

In addition, we will incorporate sparsity through direct inclusion of sparse density models (SAITO, 2004; SUN *et al.*, 2016; BIEN; TIBSHIRANI, 2011) into (5.1). This enables us to stay within the IVA maximum likelihood solution hence preserving all the theoretical and practical advantages associated with the maximum likelihood theory (ADALI *et al.*, 2014b).

2. Constraints based on Fisher criterion to address the class imbalance problem in misinformation detection Prior information about the associations between the samples and their classes could enhance the discrimination power of the extracted feature embeddings spanned by the SCVs (JIN *et al.*, 2020; DHIR; LEE, 2011), especially in the case of the class imbalance problem present in misinformation detection. Fisher discriminant cost could be added to (5.1), encouraging the exploitation of the available class labels, thus, yielding feature embeddings with high discrimination power. In addition, by exploiting prior information about the multi-modal datasets, we can achieve commonly shared feature embeddings across different multi-modal sources as well as feature embeddings that can explain differences between the datasets. In both cases, we form a dual maximization problem under a Lagrangian framework, which jointly increases the mutual information within the feature embeddings of an SCV, and at the same time, maximizes the functional measure of discrimination of the different feature embeddings.

BIBLIOGRAPHY

ADALI, T.; AKHONDA, M.; CALHOUN, V. Ica and iva for data fusion: An overview and a new approach based on disjoint subspaces. **IEEE sensors letters**, IEEE, v. 3, n. 1, p. 1–4, 2018.

ADALI, T.; ANDERSON, M.; FU, G. Diversity in independent component and vector analyses: Identifiability, algorithms, and applications in medical imaging. **IEEE Signal Processing Magazine**, IEEE, v. 31, n. 3, p. 18–33, 2014.

ADALI, T.; ANDERSON, M.; FU, G.-S. Diversity in independent component and vector analyses: Identifiability, algorithms, and applications in medical imaging. **IEEE Signal Processing Magazine**, v. 31, n. 3, p. 18–33, May 2014. ISSN 1053-5888.

ANDERSON, M.; ADALI, T.; LI, L. X.-L. Joint Blind Source Separation With Multivariate Gaussian Model: Algorithmsand Performance Analysis. **Signal Processing, IEEE Transactions on**, v. 60, n. 4, p. 1672–1683, April 2012. ISSN 1053-587X.

ANDERSON, M.; FU, G.; PHLYPO, R.; ADALI, T. Independent Vector Analysis: Identification Conditions and Performance Bounds. **Signal Processing, IEEE Transactions on**, v. 62, 03 2013.

ANDERSON, M.; FU, G.; PHLYPO, R.; ADALI, T. Independent vector analysis, the kotz distribution, and performance bounds. In: **2013 IEEE International Conference on Acoustics, Speech and Signal Processing**. [S.l.: s.n.], 2013. p. 3243–3247.

ANDERSON, M.; FU, G.; PHLYPO, R.; ADALı, T. Independent Vector Analysis: Identification Conditions and Performance Bounds. **IEEE Transactions on Signal Processing**, v. 62, n. 17, p. 4399–4410, 2014.

ANDERSON, M.; LI, X.; RODRIGUEZ, P.; ADALI, T. An effective decoupling method for matrix optimization and its application to the ica problem. In: **2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. [S.l.: s.n.], 2012. p. 1885–1888.

BEHMARDI, B.; RAICH, R.; HERO, A. Entropy estimation using the principle of maximum entropy. In: IEEE. **2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. [S.l.], 2011. p. 2008–2011.

BHINGE, S.; BOUKOUVALAS, Z.; LEVIN-SCHWARTZ, Y.; ADALI, T. Iva for abandoned object detection: Exploiting dependence across color channels. In: IEEE. **2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. [S.l.], 2016. p. 2494–2498.

BICKEL, P.; RITOV, Y.; TSYBAKOV, A. Simultaneous analysis of Lasso and dantzig selector. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 37, n. 4, p. 1705–1732, 2009.

BIEN, J.; TIBSHIRANI, R. Sparse estimation of a covariance matrix. **Biometrika**, Oxford University Press, v. 98, n. 4, p. 807–820, 2011.

BOUKOUVALAS, Z. Development of ICA and IVA Algorithms with Application to Medical Image Analysis. **arXiv preprint arXiv:1801.08600**, 2018.

BOUKOUVALAS, Z.; ELTON, D.; CHUNG, P.; FUGE, M. **Independent Vector Analysis for Data Fusion Prior to Molecular Property Prediction with Machine Learning**. 2018.

BOUKOUVALAS, Z.; ELTON, D.; PETER, C. W.; FUGE, M. Independent vector analysis for data fusion prior to molecular property prediction with machine learning. **arXiv preprint arXiv:1811.00628**, 2018.

BOUKOUVALAS, Z.; FU, G.; ADALı, T. An efficient multivariate generalized Gaussian distribution estimator: Application to IVA. In: **2015 49th Annual Conference on Information Sciences and Systems (CISS)**. [S.l.: s.n.], 2015. p. 1–4.

BOUKOUVALAS, Z.; LEVIN-SCHWARTZ, Y.; CALHOUN, V.; ADALI, T. Sparsity and independence: Balancing two objectives in optimization for source separation with application to fMRI analysis. **Journal of the Franklin Institute**, Elsevier, v. 355, n. 4, p. 1873–1887, 2018.

BOUKOUVALAS, Z.; MALLINSON, C.; CROTHERS, E.; JAPKOWICZ, N.; PIPLAI, A.; MITTAL, S.; A.JOSHI; T.ADALı. **Independent Component Analysis for Trustworthy Cyberspace during High Impact Events: An Application to Covid-19**. 2020.

BOUKOUVALAS, Z.; SAID, S.; BOMBRUN, L.; BERTHOUMIEU, Y.; Adalı, T. A New Riemannian Averaged Fixed-Point Algorithm for MGGD Parameter Estimation. **IEEE Signal Processing Letters**, v. 22, n. 12, p. 2314–2318, 2015.

BOYD, S.; VANDENBERGHE, L. **Convex optimization**. [S.l.]: Cambridge university press, 2004.

CANDES, E.; TAO, T. Decoding by linear programming. **IEEE transactions on information theory**, IEEE, v. 51, n. 12, p. 4203–4215, 2005.

COVER, T.; JOY, T. **Elements of Information Theory**. 2nd. ed. [S.l.]: Wiley-Interscience, 2006. (Wiley Series in Telecommunications and Signal Processing). ISBN 0471241954.

DHIR, C.; LEE, S. Discriminant independent component analysis. **IEEE transactions on neural networks**, IEEE, v. 22, n. 6, p. 845–857, 2011.

DICK, J.; KUO, F. Y.; SLOAN, I. High-dimensional integration: The quasi-Monte Carlo way. **Acta Numerica**, Cambridge University Press, v. 22, p. 133–288, 2013.

FU, G.; BOUKOUVALAS, Z.; ADALı, T. Density estimation by entropy maximization with kernels. In: **2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. [S.l.: s.n.], 2015. p. 1896–1900. ISSN 2379-190X.

FÜRNKRANZ, J. A study using n-gram features for text categorization. **Austrian Research Institute for Artifical Intelligence**, v. 3, n. 1998, p. 1–10, 1998.

GIRGIS, S.; AMER, E.; GADALLAH, M. Deep learning algorithms for detecting fake news in online text. In: **2018 13th International Conference on Computer Engineering and Systems (ICCES)**. [S.l.: s.n.], 2018. p. 93–97.

HURLEY, N.; RICKARD, S. Comparing measures of sparsity. **IEEE Transactions on Information Theory**, IEEE, v. 55, n. 10, p. 4723–4741, 2009.

HYVARINEN, A. Fast and robust fixed-point algorithms for independent component analysis. **IEEE Transactions on Neural Networks**, v. 10, n. 3, p. 626–634, 1999.

JAIN, S.; SHARMA, V.; KAUSHAL, R. Towards automated real-time detection of misinformation on twitter. In: **2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)**. [S.l.: s.n.], 2016. p. 2015–2020.

JAYNES, E. T. Information Theory and Statistical Mechanics. **Phys. Rev.**, American Physical Society, v. 106, p. 620–630, May 1957.

JAYNES, E. T. Information Theory and Statistical Mechanics. II. **Phys. Rev.**, American Physical Society, v. 108, p. 171–190, Oct 1957.

JIN, R.; DONTARAJUA, K.; KIM, S.; AKHONDA, S.; ADALI, T. Dictionary learning-based fmri data analysis for capturing common and individual neural activation maps. **IEEE Journal of Selected Topics in Signal Processing**, IEEE, 2020.

KIM, T. Real-time independent vector analysis for convolutive blind source separation. **IEEE Transactions on Circuits and Systems I: Regular Papers**, IEEE, v. 57, n. 7, p. 1431–1438, 2010.

KIM, T.; ATTIAS, H. T.; LEE, S.; LEE, T. Blind source separation exploiting higher-order frequency dependencies. **Audio, Speech, and Language Processing, IEEE Transactions on**, IEEE, v. 15, n. 1, p. 70–79, 2007.

KIM, T.; ELTOFT, T.; LEE, T. Independent vector analysis: An extension of ICA to multivariate components. In: **Independent Component Analysis and Blind Signal Separation**. [S.l.]: Springer, 2006. p. 165–172.

KUNG, S. Y. **Kernel Methods and Machine Learning**. [S.l.]: Cambridge University Press, 2014. ISBN 978-1-107-02496-0.

LI, X.; ADALI, T. Independent Component Analysis by Entropy Bound Minimization. **IEEE Transactions on Signal Processing**, v. 58, n. 10, p. 5151–5164, Oct 2010. ISSN 1941-0476.

LI, X.; ZHANG, X. Nonorthogonal joint diagonalization free of degenerate solution. **IEEE Transactions on Signal Processing**, v. 55, n. 5, p. 1803–1814, 2007.

MCLACHLAN, G.; D.PEEL. **Finite Mixture Models**. [S.l.]: Wiley, 2004. (Wiley Series in Probability and Statistics). ISBN 9780471654063.

NEGAHBAN, S.; RAVIKUMAR, P.; WAINWRIGHT, M.; YU, B. A unified framework for high-dimensional analysis of *m*-estimators with decomposable regularizers. **Statistical Science**, Institute of Mathematical Statistics, v. 27, n. 4, p. 538–557, 2012.

NESTEROV, Y. Gradient methods for minimizing composite functions. **Mathematical Programming**, Springer, v. 140, n. 1, p. 125–161, 2013.

NIEDERREITER, H. **Random Number Generation and Quasi-Monte Carlo Methods**. USA: Society for Industrial and Applied Mathematics, 1992. ISBN 0898712955.

O'LEARY, D. **Scientific Computing with Case Studies**. [S.l.]: Society for Industrial and Applied Mathematics, 2009. (Other Titles in Applied Mathematics). ISBN 9780898716665.

ROJO-ÁLVAREZ, J. L.; MARTÍNEZ-RAMON, M.; MUÑOZ-MARI, J.; CAMPS-VALLS, G. **Digital Signal Processing with Kernel Methods**. [S.l.]: Wiley - IEEE Press, 2018. ISBN 978-1118611791.

ROMANO, J.; ATTUX, R.; CAVALCANTE, C.; SUYAMA, R. **Unsupervised Signal Processing: Channel Equalization and Source Separation**. [S.l.]: CRC Press, 2010. ISBN 9780849337512.

SAITO, N. The generalized spike process, sparsity, and statistical independence. **Modern signal processing**, Cambridge, UK: Cambridge Univ. Press, v. 46, p. 317–340, 2004.

SCHMIDT, M.; FUNG, G.; ROSALES, R. Fast optimization methods for ℓ^1 regularization: A comparative study and two new approaches. In: SPRINGER. **European Conference on Machine Learning**. [S.l.], 2007. p. 286–297.

S.SINGHANIA; FERNANDEZ, N.; RAO, S. 3han: A deep neural network for fake news detection. In: **Neural Information Processing**. Cham: Springer International Publishing, 2017. ISBN 978-3-319-70096-0".

SUN, Y.; BABU, P.; PALOMAR, D. Robust estimation of structured covariance matrix for heavy-tailed elliptical distributions. **IEEE Transactions on Signal Processing**, IEEE, v. 64, n. 14, p. 3576–3590, 2016.

TIBSHIRANI, R. Regression shrinkage and selection via the LASSO. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, p. 267–288, 1996.

VIA, J.; ANDERSON, M.; XI-LIN, L.; ADALI, T. A maximum likelihood approach for independent vector analysis of Gaussian data sets. In: **Machine Learning for Signal Processing (MLSP), 2011 IEEE International Workshop on**. [S.l.: s.n.], 2011. p. 1–6. ISSN 1551-2541.

WU, L.; MORSTATTER, F.; CARLEY, K.; LIU, H. Misinformation in social media: Definition, manipulation, and detection. Association for Computing Machinery, New York, NY, USA, v. 21, n. 2, p. 80–90, 2019. ISSN 1931-0145. Disponível em: <https://doi.org/10.1145/3373464.3373475>.

YU, F.; LIU, Q.; WU, S.; WANG, L.; TAN, T. A convolutional approach for misinformation identification. In: **Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17**. [s.n.], 2017. p. 3901–3907. Disponível em: <https://doi.org/10.24963/ijcai.2017/545>.