UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA DE TELEINFOMÁTICA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE TELEINFORMÁTICA

DOUTORADO EM ENGENHARIA DE TELEINFORMÁTICA

LUCAS DE PAULA DAMASCENO

INDEPENDENT VECTOR ANALYSIS WITH FLEXIBLE STATISTICAL MODELING FOR MULTIMODAL DATA FUSION FROM BLIND SOURCE SEPARATION TO MISINFORMATION DETECTION

FORTALEZA

2025

LUCAS DE PAULA DAMASCENO

INDEPENDENT VECTOR ANALYSIS WITH FLEXIBLE STATISTICAL MODELING FOR MULTIMODAL DATA FUSION FROM BLIND SOURCE SEPARATION TO MISINFORMATION DETECTION

Thesis submitted to the Graduate Program in Teleinformatics Engineering of the Technology Center at the Federal University of Ceará, as a partial requirement for obtaining the title of Doctor in Teleinformatics Engineering. Concentration Area: Signals and Systems.

Supervisor: Prof. Dr. Charles Casimiro Cavalcante

Co-supervisor: Prof. Dr. Zois Boukouvalas

FORTALEZA

2025

To my family, for their endless support and sacrifice, and to my beloved wife Ana Livia, for walking every step of this journey by my side.

ACKNOWLEDGEMENTS

First of all, I would like to thank God for guiding my steps, providing me this opportunity and supporting me with strength, patience and peace throughout this journey.

I am deeply grateful to my advisor, Prof. Dr. Charles Casimiro Cavalcante, for his unwavering support, guidance, and mentorship during my graduate journey. His expertise, dedication, and encouragement were fundamental to the success of this research. I am also very grateful to my co-advisor, Prof. Dr. Zois Boukouvalas, for his invaluable guidance, insightful discussions, and continuous support, which greatly contributed to the development of this work.

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

I extend my sincere gratitude to the members of the examination committee, Prof. Dr. Guilherme de Alencar Barreto, Profa. Dra. Michela Mulas, Prof. Dr. Anderson de Rezende Rocha, and Prof. Dr. Leonardo Tomazeli Duarte, for accepting to evaluate this work and for their valuable technical discussions and suggestions.

I also thank my colleagues and friends from UFC for the enriching study and work environment, as well as for the technical and non-technical discussions that made this journey more enjoyable.

I am extremely grateful for all the love, support, prayers and caring from my parents and brothers, Damasceno, Rosangela, Robson and Tiago. Thank you for all the sacrifices you have made and for always believing in me.

Finally, I am deeply grateful to my beloved wife, Ana Livia, for her unconditional love, patience, understanding and support from the very beginning. This achievement is also yours.

“A direção em que a educação conduz um homem determinará sua vida futura. Não treine uma criança a aprender pela força ou severidade; mas direcione-a pelo que diverte sua mente, para que você possa descobrir com precisão a inclinação peculiar do gênio de cada um.”

(Platão)

RESUMO

A extensa difusão de informações digitais, através de variados canais de comunicação, alterou de forma significativa a forma como as sociedades geram, consomem e interpretam o conhecimento. Entretanto, essa mudança também acentuou a disseminação de informações errôneas. A intricada natureza desse fenômeno se encontra em seu aspecto intrinsecamente multimodal, no qual imagens alteradas, textos enganosos e mídias sintéticas se entrelaçam, formando narrativas falsas que desafiam tanto a compreensão humana quanto a de máquinas. A presente tese investiga a urgente questão da identificação automatizada de desinformação multimodal a partir de uma base teórica alicerçada na Análise de Vetores Independentes (Independent Vector Analysis - IVA) e na análise estatística de sinais. Primeiramente, procede-se à revisão da fundamentação teórica da IVA, entendida como uma ampliação da Análise de Componentes Independentes (Independent Component Analysis - ICA), que expande a sua habilidade de modelar múltiplos conjuntos de dados por meio da investigação das interdependências entre modalidades. Com base nessa fundamentação, propõe-se uma metodologia de estimação de densidade multivariada fundamentada no Princípio da Máxima Entropia (MEP), que combina restrições globais e locais através do estimador de Maximização da Entropia Multivariada com Kernels (M-EMK). Este estimador oferece funções de densidade de probabilidade adaptáveis e expressivas, garantindo eficiência computacional por meio da integração Quasi-Monte Carlo, resultando na elaboração do algoritmo IVA-M-EMK. Além da separação de fontes, esta tese emprega a proposta formulada na identificação de desinformação multimodal, na qual variados tipos de dados, incluindo texto, imagens e embeddings semânticos, são analisados em conjunto a fim de evidenciar consistências e contradições subjacentes. Ao reconhecer que os sinais enganosos costumam exibir características esparsas e organizadas, emprega-se o IVA-SPICE, que integra a esparsidade estruturada por meio da estimativa de covariância inversa, permitindo o tratamento eficiente de espaços multimodais em alta dimensão. As investigações experimentais revelam que a estrutura sugerida, fundamentada em IVA, supera as metodologias tradicionais unimodais e de fusão simplificada em múltiplos conjuntos de dados, obtendo maior precisão, capacidade de generalização e explicabilidade. De forma resumida, esta tese fundamenta a Análise de Vetores Independentes, combinada à modelagem adaptativa de densidade e à esparsidade estruturada, como uma base sólida, escalável e passível de interpretação para a integração de dados multimodais.

Palavras-chave: independent vector analysis; desinformação multimodal; separação cega de fontes; estimação por máxima entropia; detecção de *fake news*.

ABSTRACT

The widespread dissemination of digital information through various communication channels has significantly altered the way societies generate, consume, and interpret knowledge. However, this shift has also accentuated the spread of misinformation. The intricate nature of this phenomenon lies in its inherently multimodal aspect, in which altered images, misleading texts, and synthetic media intertwine, forming false narratives that defy both human and machine comprehension. This thesis investigates the urgent issue of automated identification of multimodal misinformation based on a theoretical framework grounded in Independent Vector Analysis (IVA) and statistical signal analysis. First, we review the theoretical foundation of IVA, understood as an extension of Independent Component Analysis (ICA), which expands its ability to model multiple data sets by investigating the interdependencies between modalities. Based on this foundation, we propose a multivariate density estimation methodology based on the Maximum Entropy Principle (MEP), which combines global and local constraints through the Multivariate Entropy Maximization with Kernels (M-EMK) estimator. This estimator offers adaptive and expressive probability density functions, ensuring computational efficiency through Quasi-Monte Carlo integration, resulting in the development of the IVA-M-EMK algorithm. Besides source separation, this thesis applies the proposed approach to identifying multimodal misinformation, in which various data types, including text, images, and semantic embeddings, are analyzed together to highlight underlying consistencies and contradictions. Recognizing that misleading signals often exhibit sparse and organized characteristics, IVA-SPICE is employed, which integrates structured sparsity through inverse covariance estimation, enabling the efficient processing of high-dimensional multimodal spaces. Experimental investigations reveal that the proposed framework outperforms traditional unimodal and simplified fusion methodologies across multiple datasets, achieving greater accuracy, generalizability, and explainability. This thesis establishes Independent Vector Analysis, combined with adaptive density modeling and structured sparsity, as a solid, scalable, and interpretable foundation for multimodal data integration.

Keywords: independent vector analysis; multimodal misinformation; blind source separation; maximum entropy estimation; fake news detection.

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

LIST OF ABBREVIATIONS AND ACRONYMS

BSS	Blind Source Separation
CCA	Canonical Correlation Analysis
DIVA	Discriminant Independent Vector Analysis
EMK	Entropy Maximization with Kernels
GGD	Generalized Gaussian Distribution
GMM	Gaussian Mixture Model
HOS	Higher-Order Statistics
ICA	Independent Component Analysis
ISI	Inter-Symbol Interference
IVA	Independent Vector Analysis
IVA-G	IVA-Gaussian
IVA-GGD	IVA with Generalized Gaussian Distribution
IVA-L	IVA-Laplacian
IVA-M-EMK	IVA with Multivariate Entropy Maximization using Kernels
IVA-SPICE	IVA with Structured Sparsity via Inverse Covariance Estimation
JBSS	Joint Blind Source Separation
KDE	Kernel Density Estimation
LDA	Latent Dirichlet Allocation
LIME	Local Interpretable Model-agnostic Explanations
LLMs	Large Language Models
MC	Monte Carlo
M-EMK	Multivariate Entropy Maximization with Kernels
MEP	Maximum Entropy Principle
MI	Mutual Information
ML	Maximum Likelihood
NLP	Natural Language Processing

PDF	Probability Density Function
PCA	Principal Component Analysis
QMC	Quasi-Monte Carlo
SCV	Source Component Vector
SVMs	Support Vector Machines
TF-IDF	Term Frequency-Inverse Document Frequency
VGG-16	Visual Geometry Group 16-layer Network
XAI	Explainable Artificial Intelligence

LIST OF SYMBOLS

$\mathbb{R}$	Set of real numbers
$\mathbb{R}^K$	K-dimensional real vector space
$\mathbf{A}^{[k]}$	Mixing matrix for the k-th dataset
$\mathbf{W}^{[k]}$	Demixing matrix for the k-th dataset
$\mathbf{W}_n$	Matrix $\mathbf{W}$ with the n-th row removed
$\mathbf{H}_n$	Orthogonal projection matrix onto the null space of $\mathbf{W}_n$
$\mathbf{I}$	Identity matrix
$\mathbf{J}$	Jacobian matrix
$\boldsymbol{\Sigma}$	Covariance matrix
$\mathbf{x}^{[k]}$	Observed data vector for the k-th dataset
$\mathbf{s}^{[k]}$	Source vector for the k-th dataset
$\mathbf{s}_n$	n-th source component vector (SCV)
$\mathbf{y}^{[k]}$	Estimated source vector for the k-th dataset
$\mathbf{y}_n$	n-th estimated source component vector
$\mathbf{w}_n^{[k]}$	n-th row vector of the demixing matrix for dataset k
$\mathbf{h}_n$	Orthogonal projection vector
$\boldsymbol{\mu}$	Mean vector
$\boldsymbol{\lambda}$	Vector of Lagrange multipliers $[\lambda_0, \ldots, \lambda_M]$
$\boldsymbol{\alpha}$	Vector of empirical averages $[\alpha_0, \ldots, \alpha_M]$
$\mathbf{r}$	Vector of measuring functions $[r_0, \ldots, r_M]$
$\mathscr{L}_{\text{IVA}}$	Maximum likelihood cost function for IVA
I_{IVA}	Mutual information cost function for IVA
$H(\mathbf{y}_n)$	Differential entropy of the n-th SCV
$p(\mathbf{y}_n)$	True probability density function of the n-th SCV
$\hat{p}(\mathbf{y}_n)$	Estimated probability density function of the n-th SCV
$q(\mathbf{y})$	Gaussian kernel function

$r_i(\mathbf{y})$	i-th measuring function
$\phi^{[k]}$	Score function for the k-th dataset
N	Number of source components
K	Number of datasets (modalities)
T	Number of samples
M	Number of constraints
λ_i	i-th Lagrange multiplier
α_i	i-th empirical average
γ	Step size (learning rate)

CONTENTS

1 INTRODUCTION

1.1 Context and Motivation

The rapid and extensive proliferation of digital information across a diverse range of platforms and communication channels has fundamentally transformed the methods by which contemporary societies consume, interpret, and disseminate knowledge, data, and narratives. The democratization and unprecedented accessibility of information have undoubtedly produced substantial and extensive advantages for global communication, education, and civic engagement. Nonetheless, these advantages have engendered a highly volatile and susceptible environment in which misinformation, defined as content that is false, deceptive, or misleading, can disseminate with alarming speed and magnitude, thereby significantly shaping public opinion, eroding the integrity of democratic processes, and posing critical, at times life-threatening, risks to public health and safety. (Sahoo; Gupta, 2021) (Rajabi *et al.*, 2024)

A social media post that includes a repurposed photograph from a disaster alongside fabricated casualty figures necessitates a comprehensive and nuanced assessment of textual assertions, visual elements, and their semantic interrelations to accurately discern deceptive intent and authenticity. The severity of this challenge escalates as it becomes evident that contemporary misinformation campaigns utilize increasingly nuanced, sophisticated, and adaptive tactics, such as factually accurate statements framed in misleading contexts, genuine images repurposed to bolster false narratives, and highly advanced deepfake videos that even test expert human discernment and state-of-the-art detection technologies.

To tackle this intricate and swiftly evolving issue, the academic and research communities have developed a diverse set of methodologies, encompassing natural language processing techniques for analyzing linguistic patterns, semantic meanings, and discourse structures, as well as computer vision methods for identifying manipulations, discrepancies, and irregularities in images and videos. Nonetheless, these methods frequently address each modality in isolation or utilize overly simplistic integration techniques that fail to consider the complex, multifaceted, and often nonlinear statistical relationships and interactions among various information streams. The principal challenge resides not only in the isolated processing of multiple modalities but in achieving a profound, integrated, and contextually aware comprehension of their interrelations, particularly regarding the alignment of textual claims with visual evidence, the correspondence of temporal elements in video with related audio streams, and how these intricate multimodal

relationships can more effectively reveal the presence or absence of deceptive content with greater accuracy and resilience. (Rajabi *et al.*, 2024)

This thesis thoroughly investigates the critical and intricate issue of automated misinformation detection in diverse multimodal digital content using advanced blind source separation techniques and refined statistical signal processing methods. The Statistical Modeling Challenge stems from the intricate nature of real-world multimodal data, which frequently display highly non-Gaussian distributions and subtle nonlinear dependencies both within and between modalities. Conventional methods that depend on basic parametric distribution assumptions, such as Gaussianity or independence, or that assume modality independence, are inadequate for capturing complex relationships, thereby impairing detection accuracy and generalizability.

The Dimensionality Challenge arises when the incorporation of additional modalities in the analysis results in an exponential growth of the joint feature space's dimensionality. This expansion causes computational difficulties and reduces statistical reliability due to the curse of dimensionality, evident in increased processing requirements and diminished model estimation quality.

The Interpretability Challenge underscores the imperative for misinformation detection systems to deliver not only precise identification of misleading content but also clear, understandable, and actionable rationales for their determinations. Although black-box models can attain elevated accuracy, they frequently lack the vital attributes of transparency, interpretability, and trustworthiness necessary for implementation in sensitive, high-stakes contexts such as public policy, law enforcement, and health communication.

The Adaptability Challenge presents a significant and persistent problem in misinformation detection, as the tactics utilized by malicious entities to spread false or misleading information are advancing at an accelerating and erratic rate. The evolving nature of deceptive tactics requires detection methods that are both robust and able to generalize effectively beyond the specific distributions present in their initial training datasets. These methods must demonstrate the capacity to adapt seamlessly and efficiently to previously unrecognized forms of deception, without necessitating exhaustive, costly, and resource-intensive model retraining or manual intervention. This challenge is intensified by the varied and constantly changing environment of misinformation, which includes numerous platforms, languages, cultural contexts, and modalities, thus requiring adaptable, scalable, and culturally attuned solutions. Confronting this adaptability challenge is essential for guaranteeing the effectiveness, pertinence, and ethical

implementation of misinformation detection systems in practical applications, where the swift recognition of novel deceptive patterns can substantially mitigate the dissemination and influence of false information on societies worldwide.

This study examines several fundamental and pivotal research inquiries vital for the formulation of effective frameworks for identifying multimodal misinformation. Is it feasible to employ blind source separation techniques, specifically Independent Vector Analysis (IVA), to accurately model the latent, shared, and modality-specific structures inherent in complex multimodal misinformation data? Secondly, how can adaptive, data-driven density estimation techniques surpass the constraints of rigid, fixed parametric assumptions to effectively represent the complex statistical properties and interdependencies inherent in real-world misinformation across diverse modalities? Third, what constitutes the ideal quantity and combination of modalities for efficient detection, and how does this ideality relate to the modeling framework's capacity to address the challenges posed by high-dimensional joint feature spaces and constrained training datasets? Fourth, how can sparsity constraints and structured regularization techniques be utilized to mitigate the curse of dimensionality while preserving the most discriminative, informative, and interpretable features in multimodal fusion? Fifth, can the amalgamation of supervised information and domain expertise augment the unsupervised identification and modeling of latent structures in multimodal data, thus enhancing detection efficacy, robustness, and interpretability?

To tackle these critical inquiries, we propose an innovative, comprehensive, and theoretically robust framework for multimodal misinformation detection, fundamentally based on Independent Vector Analysis enhanced with adaptable density estimation methods and structured sparsity constraints. The method perceives each modality as a noisy, composite observation of fundamental semantic sources and utilizes blind source separation to reveal latent structures that highlight cross-modal consistencies, contradictions, and anomalies indicative of misinformation. This work's primary innovation is the incorporation of maximum entropy density estimation into the IVA framework, resulting in the IVA-M-EMK algorithm (Independent Vector Analysis with Multivariate Entropy Maximization using Kernels). This novel formulation deliberately abandons strict parametric assumptions about source distributions, opting instead to learn flexible, adaptive densities that align with the genuine and diverse statistical characteristics of the data. The method integrates both global and local measuring functions to effectively capture complex, potentially multimodal, skewed, and heavy-tailed distributions, while ensuring computational efficiency through Quasi-Monte Carlo integration techniques.

Understanding that misinformation signals are often limited and concentrated in particular cross-modal interactions rather than evenly distributed across all potential relationships, the framework is enhanced through IVA-SPICE, which integrates structured sparsity through inverse covariance estimation. This advancement enables efficient scaling to higher-dimensional multimodal spaces by automatically identifying, selecting, and retaining only the most informative and pertinent dependencies, thus reducing overfitting and improving interpretability. The proposed framework functions through a methodical and principled pipeline: high-level, semantically rich features are extracted from each modality using established state-of-the-art techniques (such as Word2Vec embeddings for textual data and VGG-16 convolutional features for visual representations); these features undergo dimensionality reduction, normalization, and alignment processes to ensure comparability and compatibility; IVA-based decomposition then extracts source component vectors that encapsulate both shared and modality-specific information; ultimately, these latent representations are fed into supervised classifiers for precise, robust, and interpretable misinformation detection.

The investigations and experiments conducted offer foundational insights into multimodal fusion, challenging prevailing assumptions within the academic community. A systematic evaluation of all possible combinations among four modalities, text, image, image-to-text embeddings, and topic models, reveals a phase transition in optimal fusion techniques and modality selection. The theoretical investigation of modality scaling behavior signifies a significant conceptual progression, demonstrating that the optimal number of modalities is fluid and affected by the interplay between the quantity of training data, the dimensionality of the joint space, the model's regularization proficiency, and the inherent informativeness of each modality. The mathematical framework established quantifies the extent to which the addition of each modality amplifies the probability of spurious correlations, correlating with the combinatorial increase of potential cross-modal interactions, thereby highlighting the imperative for systematic regularization and sparsity to maintain model manageability and generalizability. This work establishes a foundation for future advancements through its modular architecture and extensible formulations. The dissociation of density estimation from the Independent Vector Analysis (IVA) framework enables the effortless integration of new estimators as they are created, while the clear interface between feature extraction and fusion permits the straightforward replacement of modality-specific encoders, rendering it adaptable to novel data types and communication channels. The theoretical insights provided here, especially regarding the correlation between

density estimation accuracy, source separation efficacy, and detection performance, delineate a clear path for the advancement of more sophisticated, interpretable, and adaptive methodologies.

The contributions detailed herein position Independent Vector Analysis, enhanced by flexible density estimation, as a robust, efficient, and scalable framework for identifying multimodal misinformation. This methodology overcomes essential constraints of current techniques, including inflexible distributional assumptions, scalability issues, and insufficient interpretability, while preserving computational efficiency and adaptability. This work delivers immediate practical solutions and offers enduring theoretical insights that will inform and inspire future research in this essential field. The organization of the subsequent sections of this thesis is as follows

Chapter 2: Independent Vector Analysis establishes a solid mathematical foundation for the study, offering a comprehensive exploration of the Independent Vector Analysis (IVA) framework, its formulation as a maximum likelihood optimization problem, and the decoupling method that facilitates practical and efficient optimization procedures.

Chapter 3: The Multivariate EMK IVA Algorithm (IVA-M-EMK) introduces a novel methodology for flexible density estimation, exploring the maximum entropy formulation, the deliberate design and selection of measurement functions, and the integration of Quasi-Monte Carlo techniques to improve computational efficiency and precision.

Chapter 4: Multimodal Misinformation Detection through Independent Vector Analysis implements the established methodology to address the critical challenge of misinformation detection, detailing the extensive workflow from feature extraction to classification and demonstrating strong performance across real-world datasets.

Chapter 5: Conclusions and Future Work summarizes the principal findings, considers the wider implications for practical application, and suggests potential avenues for future research, such as the incorporation of discriminative priors, expansion to further modalities, and the creation of real-time adaptive learning systems. This comprehensive exploration conclusively demonstrates that effective statistical modeling, combined with appropriate regularization, adaptable estimation techniques, and modular system architecture, creates a robust, resilient, and interpretable framework for understanding, identifying, and combating misinformation in the complex, dynamic, and multimodal information landscape of the digital age.

2 INDEPENDENT VECTOR ANALYSIS

By facilitating the joint factorization of several data sets, independent vector analysis (IVA) builds upon independent component analysis (ICA). IVA is developed using general density models for the latent multivariate sources in a maximum likelihood (ML) framework to concurrently account for all forms of diversity.

IVA paved the way for the development of the IVA-Laplacian (IVA-L) algorithm (Kim *et al.*, 2006; Kim *et al.*, 2007), which incorporates only higher-order statistics (HOS) and assumes a Laplacian distribution for the source component vectors. IVA was first proposed to address the convolutive ICA problem in the frequency domain by considering multiple frequency bins (Kim, 2010). IVA-Gaussian (IVA-G) (Anderson *et al.*, 2012b; Via *et al.*, 2011), on the other hand, does not use HOS and instead depends on linear relationships. Both second- and higher-order statistics are integrated into the IVA framework in later versions based on the multivariate generalized Gaussian distribution (Anderson *et al.*, 2013; Anderson *et al.*, 2014a; Boukouvalas *et al.*, 2015a; Boukouvalas *et al.*, 2015b), which reflect a more generic approach.

This chapter integrates new developments in information-theoretic learning and source modeling to examine the Independent Vector Analysis (IVA) paradigm. More adaptable models for latent sources have been developed by advancements in independent component analysis (Comon, 1994), and more recent formulations have addressed the constraints imposed by rigorous statistical independence assumptions. These guidelines encourage the development of an updated IVA formulation that can handle intricate and real-world data distributions, expanding its use in more extensive multimodal settings.

Even while IVA algorithms based on latent source density models have proven successful in a variety of applications, there is still a significant obstacle to overcome: creating and estimating adaptable models for the underlying source density, which is necessary to use IVA efficiently in a wider range of issues. (Bhinge *et al.*, 2016; Adali *et al.*, 2018; Boukouvalas *et al.*, 2018).

This chapter reviews earlier research on Independent Vector Analysis (IVA) using multivariate entropy maximization for semi-parametric density estimation (Damasceno, 2021). The IVA-M-EMK technique was presented in that formulation, which builds adaptable multivariate probability density functions by fusing kernel-based estimation with the maximum entropy principle. Across a wide range of source distributions, the technique showed enhanced source

separation performance.

Since then, other experimental settings have confirmed the IVA-M-EMK algorithm's efficacy in collecting intricate source structures. These findings demonstrate its usefulness outside of idealized settings, especially when traditional density models are unable to accurately depict the distributions of data in the real world.

In this work, the emphasis of IVA-based research is shifted from statistical source separation to the integration of complimentary and heterogeneous data from several data modalities. The suggested method uses IVA as a multimodal data fusion tool, allowing the joint representation of sources from domains including image, text, and learnt embeddings, rather than restricting the analysis to statistical independence. With special attention to applications like fake news detection, where information is frequently dispersed across multiple modalities, this methodological progression promotes downstream tasks that profit from cross-modal interactions.

Here, we begin by expressing the IVA algorithm mathematically within the context of machine learning. The objective functions for mutual information (MI) are then derived, offering a basis that exploits many forms of variety while maintaining the theoretical advantages of ML-based techniques. Finally, we reformulate the IVA model into an optimization problem and explain it from an algorithmic standpoint.

2.1 IVA - General modeling

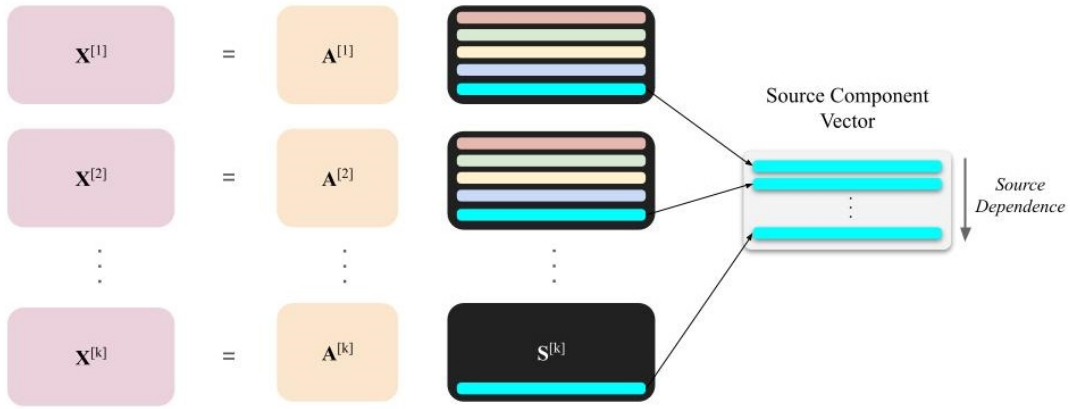
As mentioned in the previous section, IVA extends ICA and consequently its mathematical formulation follows a parallel structure. The key difference is the presence of K data sets, denoted by $\mathbf{x}^{[k]}$, with $k = 1, \ldots, K$, where each set represents a linear combination of N mutually independent sources. Adopting random-vector notation and assuming that the samples are i.i.d., the noise-free IVA model can be written as

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

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

In addition, while the components within each $\mathbf{s}^{[k]}$ are assumed to be mutually independent, dependencies between corresponding components from different datasets are

Figure 1 – JBSS structure.



Source: Prepared by the author.

allowed. This characteristic is captured through the definition of the source component vector (SCV), constructed by vertically stacking the nth source across all K datasets, and denoted as

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

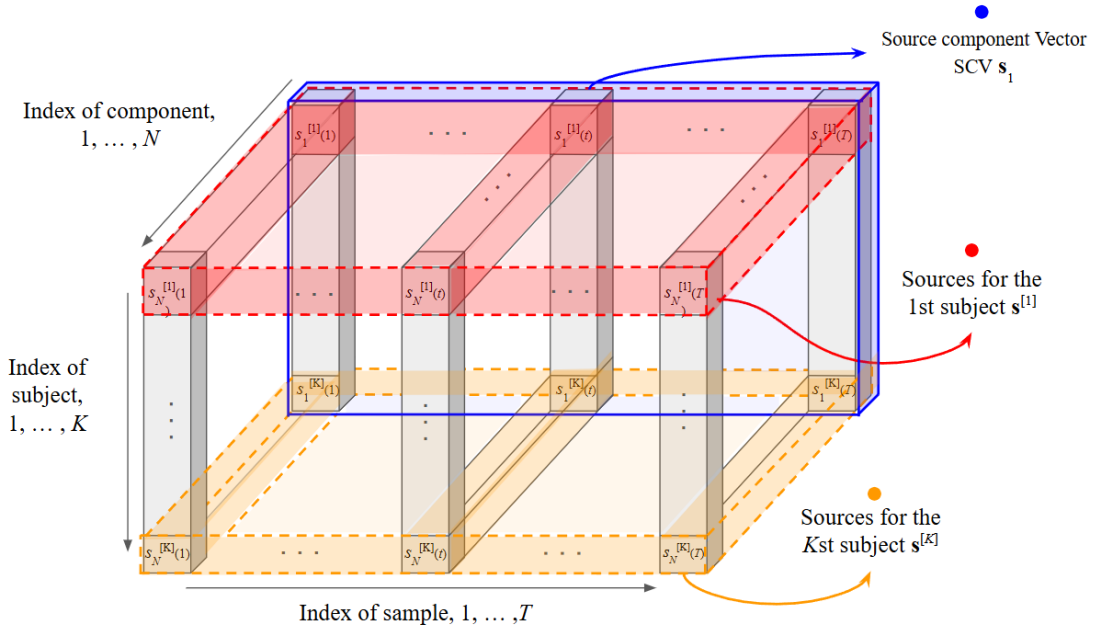
The illustration in Figure 2 provides a clear representation of the structure of the datasets used in our analysis. Each dimension reflects a key element of the data: the *component* index $(1,\ldots,N)$ identifies the source components, the *subject* index $(1,\ldots,K)$ corresponds to the different datasets, and the sample index $(1,\ldots,T)$ denotes temporal or instance-based observations. In scenarios involving two modalities, such as *image* and *text*, the color-coded sections highlight these distinctions: the *red* regions correspond to sources for the first dataset (e.g., image modality), while the *orange* regions represent sources for the Kth dataset (e.g., text modality). The *blue* outline emphasizes the formation of the source component vector (SCV) by grouping corresponding components across datasets.

The inclusion of this representation within the IVA framework enables data fusion across multiple datasets, which is particularly valuable in classification problems. This is because the most informative content is not always present in a single source. For instance, in internet memes, the essential information may lie within the image rather than the accompanying text.

An example illustrating the construction of an SCV can be found in Figure 1. Unlike ICA, where a single demixing matrix is typically estimated, IVA aims to recover K demixing matrices, each producing maximally independent source estimates according to $\mathbf{y}^{[k]} = \mathbf{W}^{[k]}\mathbf{x}^{[k]}$.

Beyond the basic multimodal analysis of two datasets, this work extends the framework to incorporate up to four different data sources, adding *Image2Text embeddings* and *topic model outputs* alongside the original image and text data. This expansion allows for a richer

Figure 2 – Illustration of SCV and source vector.



Source: Prepared by the author.

and more comprehensive representation of the underlying information, capturing diverse and complementary signals across modalities.

Importantly, while traditional IVA focuses on accurate estimation of probability density functions (PDF) for source separation, the present study emphasizes *data fusion* by leveraging diversity between modalities to integrate complementary representations from heterogeneous sources. Rather than aiming solely at statistical separation of independent sources, our approach prioritizes the unification of information from different modalities to improve downstream task performance, specifically, the detection of fake news. Through IVA-based fusion, we aim to exploit the synergistic potential of multimodal data, enhancing robustness and accuracy compared to single-modality baselines. This will be the main topic of the next chapters while we describe our model and contributions.

All terminology has been standardized for clarity: "source" refers to the origin of data (e.g., a specific dataset or signal), while "modality" refers to the data type (e.g., image, text, or embeddings). This consistent usage ensures terminological coherence throughout discussions of multimodal data fusion.

2.2 IVA cost function

IVA can be expressed within a maximum likelihood (ML) framework. According to (Boukouvalas, 2017), the corresponding ML objective function for IVA is defined as

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

where $\mathbf{y}_n$ denotes the nth estimated random vector, with $p(\mathbf{y}_n)$ representing its corresponding multidimensional probability density function (PDF). Based on the asymptotic equipartition property, it has been shown that as the sample size approaches infinity, minimizing the mutual information (MI) cost function becomes equivalent to maximizing the ML cost function (Adali *et al.*, 2014; Boukouvalas, 2017; Cover; Thomas, 2006). This equivalence allows for the exploitation of all the theoretical benefits associated with ML theory. Consequently, maximization of source independence can be achieved through minimization of the MI cost function, which is formulated as (Fu *et al.*, 2015):

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

In this expression, $H\left(\mathbf{y}_n\right)$ represents the differential entropy of the vector of the n th source component (SCV), while C stands for the constant term $H(\mathbf{x}^{[1]},\ldots,\mathbf{x}^{[K]})$. The gradient associated with this formulation is provided in (Boukouvalas, 2017).

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

where $p(y_n)$ stands for its probability density function (PDF). It has been shown that minimizing (2.4) is equivalent to maximizing the ML cost function for IVA (Adali *et al.*, 2014; Boukouvalas, 2017; Cover; Thomas, 2006). Consequently, results derived for mutual information (MI) minimization can be used to address IVA maximization.

However, minimizing (2.4) remains challenging due to the lack of direct access to the true underlying PDF of each estimated SCV.

To formalize this point, consider $\hat{p}(\mathbf{y}_n)$ as the assumed multivariate PDF for the nth estimated SCV. Then, the corresponding differential entropy can be written as

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

where $f(p(\mathbf{y}_n) \mid\mid \hat{p}(\mathbf{y}_n))$ denotes the Kullback–Leibler divergence (relative entropy) between the true density and the assumed model of the nth SCV (Cover; Thomas, 2006). From (2.6),

it becomes evident that perfect source recovery can be achieved if the assumed model exactly matches the true latent multivariate density

2.3 IVA from an algorithmic point of view

By employing the IVA mutual information (MI) objective function, the derivative with respect to each demixing matrix can be expressed as (Boukouvalas, 2017)

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

where $\phi^{[k]}$, known as the score function, is defined as

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

The update rule for each of the K demixing matrices follows

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

where γ represents the step size.

Despite this formulation, optimization involving matrix parameters, common in blind source separation (BSS) algorithms, often encounters challenges. Specifically, in the IVA update rules, optimizing over the space of all invertible matrices can lead to poor convergence. This issue is exacerbated by the need to compute the inverse of the matrix $\mathbf{W}$ at each iteration, as indicated in Equation (2.5).

To address this, the minimization of the MI cost function (2.4) is reformulated into a sequence of simpler subproblems. Rather than minimizing the objective function with respect to the entire matrix $\mathbf{W}^{[k]}$, the function is individually minimized with respect to each row vector $\mathbf{w}_1, \ldots, \mathbf{w}_K$. This approach not only simplifies the density matching problem but also ensures that the estimation of one source does not influence the estimation of others. Consequently, it enhances the convergence behavior of the algorithm and facilitates the inclusion of constraints within the IVA framework.

A straightforward strategy proposed in (Hyvarinen, 1999) involves assuming that $\mathbf{W}$ is orthogonal, satisfying $\mathbf{W}\mathbf{W}^{\top} = \mathbf{1}$. Under this assumption, $|\det(\mathbf{W})| = 1$, enabling optimization with respect to the rows of $\mathbf{W}$. Although this orthogonality constraint simplifies the objective function and can improve algorithmic stability, it significantly restricts the solution space and may degrade the overall performance of source separation.

To overcome this limitation, we adopt a decoupling procedure that converts the original matrix optimization problem into a series of vector optimization tasks. This method preserves the flexibility of the solution space by eliminating the orthogonality constraint on $\mathbf{W}$, thus maintaining the full potential of the IVA model.

2.4 Decoupling procedure

To enhance the performance of source separation, a decoupling procedure is employed (Li; Zhang, 2007; Anderson *et al.*, 2012b). Rather than minimizing the MI cost function (2.4) directly with respect to the full demixing matrix $\mathbf{W}^{[k]}$, the optimization is performed over each individual row vector $\mathbf{w}_i^{[k]}$, for $i = 1,\ldots,N$. For simplicity in the mathematical formulation, the procedure is first demonstrated within the ICA framework, noting that the extension to IVA follows naturally.

Let us define $\mathbf{W}_n = [\mathbf{w}_1,\ldots,\mathbf{w}_{n-1},\mathbf{w}_{n+1},\ldots,\mathbf{w}_N]^\top \in \mathbb{R}^{(N-1)\times N}$, which contains all the rows of $\mathbf{W}$ except the nth. Since the determinant of a matrix remains invariant under row permutation (up to a sign), the square of $\det(\mathbf{W})$ can be expressed as:

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

Here, $\mathbf{H}_n = \mathbf{I}-\mathbf{W}_n^\top\left(\mathbf{W}_n\mathbf{W}_n^\top\right)^{-1}\mathbf{W}_n$ serves as the orthogonal projection matrix onto the null space of $\mathbf{W}_n$. Notably, since $\mathbf{H}_n$ is rank-one, it can be written as $\mathbf{H}_n = \mathbf{h}_n\mathbf{h}_n^\top$, where $\mathbf{h}_n$ is orthogonal to all rows of $\mathbf{W}_n$.

Thus, the determinant simplifies to:

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

Leveraging (2.11), the MI cost function (2.4) can be reformulated into a decoupled

cost function:

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

and the corresponding gradient is:

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

where the kth element of the multivariate score function is defined as:

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

Finally, it is important to highlight that the development of an accurate PDF estimator remains essential for IVA algorithms. As shown in Equation (2.6), perfect source separation requires the relative entropy term $f(p(\mathbf{y}_n), \hat{p}(\mathbf{y}_n))$ to converge to zero, implying a perfect match between the true and estimated probability density functions.

2.5 Summary

In this chapter, we revisit a theoretical and mathematical formulation of the IVA algorithm within a maximum likelihood framework, extending the blind source separation (BSS) problem to the more general joint blind source separation (JBSS) across multiple datasets. The model accommodates the increasing complexity introduced by multivariate data sources and emphasizes the flexibility of IVA in handling diverse modalities. To address the computational challenges inherent to multivariate approaches, we introduced a decoupling procedure that simplifies the optimization problem, improves convergence properties, and facilitates the incorporation of constraints.

Beyond the classical objective of accurate PDF estimation for source separation, our work shifts the focus towards data fusion across multiple datasets. We demonstrated the capability of the proposed approach to integrate up to four data sources - including image, text, Image2Text embeddings, and topic models - enhancing performance in complex tasks such as fake news detection. The chapter also highlighted the role of efficient PDF estimators, while positioning data fusion as the primary objective, aiming for robust and accurate multimodal analysis.

The following chapter builds upon this foundation, presenting the development of the PDF estimators and the application of the proposed IVA framework in real-world data fusion scenarios.

3 MULTIVARIATE EMK IVA ALGORITHM (IVA-M-EMK)

3.1 Objectives and Scope of this Chapter

The main goal of this chapter is to present an advanced, flexible, and computationally efficient methodology for estimating multivariate probability densities, based on the Maximum Entropy Principle (MEP). This chapter also aims to show how to use this framework in the Independent Vector Analysis (IVA) model in a way that greatly improves the effectiveness of blind source separation tasks. This chapter introduces the novel notion of the *Multivariate Entropy Maximization with Kernels* (M-EMK) estimator, building upon the fundamental work presented in (Damasceno, 2021). This revolutionary estimator skillfully combines global and local measurement functions, making it possible to create very expressive density models from the data. We show how this advanced estimator can be easily added to the IVA cost function, resulting in a new algorithm that greatly improves performance in a variety of source separation situations with different levels of complexity and data distributions.

The contributions of this chapter are methodically outlined as follows:

- To offer a comprehensive presentation of a semi-parametric multivariate probability density function (PDF) estimation technique, fundamentally grounded in entropy maximization and utilizing both global and local constraints to enhance model expressiveness;
- To introduce a highly efficient multidimensional integration strategy based on Quasi-Monte Carlo methods, enabling feasible and effective estimation, even within the challenging realm of high-dimensional spaces;
- To develop a novel IVA algorithm, designated IVA-M-EMK, which strategically utilizes the proposed estimator to enhance the effectiveness of source separation across a wide range of varied data distributions;
- To meticulously evaluate the proposed method through a comprehensive series of numerical experiments, employing both synthetic datasets and real-world data sources to confirm its efficacy and resilience.

The objective of this chapter is intentionally limited to the development, methodical incorporation, and comprehensive assessment of the M-EMK estimator within the IVA model framework. It is important to point out that this chapter does not try to go over the basic arithmetic behind the IVA model again, since it is assumed that the reader already knows or has read about it in a previous chapter of this thesis. Instead, the main point is to show how important density

modeling is for IVA and how much progress can be made by using entropy-based estimation techniques. This will help the area of blind source separation move forward.

3.1.1 Need for Flexible Density Models

A central challenge in Independent Vector Analysis (IVA) is the accurate modeling of the multivariate probability density function (PDF) associated with the *Source Component Vectors* (SCVs). These SCVs, defined as

$$\mathbf{s}_n = \left[s_n^{[1]}, s_n^{[2]}, \ldots, s_n^{[K]}\right]^\top \in \mathbb{R}^K, \tag{3.1}$$

represent the joint behavior of the n-th latent source across K datasets. The estimation of the SCV distribution $p(\mathbf{s}_n)$ is a critical component in the maximum likelihood (ML) formulation of IVA, where the objective function is given by (Boukouvalas, 2017):

$$\mathscr{L}_{\text{IVA}} = \sum_{n=1}^{N} \mathbb{E}\left[\log p(\mathbf{y}_n)\right] + \sum_{k=1}^{K} \log \left|\det(\mathbf{W}^{[k]})\right|, \tag{3.2}$$

where $\mathbf{y}_n$ denotes the estimated SCV obtained after demixing, and $\mathbf{W}^{[k]}$ is the demixing matrix for dataset k.

In this context, the quality of source separation strongly depends on how well the model PDF $p(\mathbf{y}_n)$ captures the true underlying distribution. Many existing IVA algorithms rely on fixed parametric assumptions such as multivariate Laplacian (Kim *et al.*, 2006), Gaussian (Anderson *et al.*, 2012a), or generalized Gaussian models (Boukouvalas *et al.*, 2015a). While these models provide analytical tractability and ease of optimization, they often impose restrictive assumptions that are rarely met in real-world datasets.

When there is a mismatch between the assumed and true source distributions, two major issues arise:

- **Suboptimal separation:** The estimated sources $\mathbf{y}_n$ do not match the true latent components, leading to residual dependencies across SCVs and degraded performance metrics such as inter-symbol interference (ISI) or mutual information.
- **Algorithmic inefficiency:** Mismatched PDFs may introduce ill-conditioning or poor gradient behavior in the optimization process, causing slow convergence or convergence to local minima.

To address these issues, a more adaptable density modeling approach is necessary, one that can conform to the actual statistical structure of the data without enforcing strict distributional forms. Non-parametric methods, including kernel density estimation (KDE), provide considerable flexibility but frequently encounter significant computational complexity and sensitivity to bandwidth selection, especially in high-dimensional contexts (Rojo-Álvarez *et al.*, 2018).

Employing *semi-parametric* models that are based on the *Maximum Entropy Principle* (MEP) (Jaynes, 1957a) is an advantageous alternative. By maximizing entropy while adhering to established statistical constraints (e.g., moments, local statistics), MEP provides a principled framework for constructing the probability density function $p(\mathbf{y}_n)$. This approach yields the least biased distribution that is in alignment with the available information.

This chapter investigates a multivariate extension of MEP that employs both global and local measuring functions, known as *Multivariate Entropy Maximization with Kernels (M-EMK)* (Fu *et al.*, 2015). The resultant estimator achieves a balance between computational feasibility and expressiveness, enabling its seamless integration into the IVA cost function and thereby improving separation performance across a variety of data distributions.

3.2 Theoretical Framework: Maximum Entropy for Density Estimation

3.2.1 Challenges in Multivariate PDF Estimation

Accurate estimation of multivariate probability density functions (PDFs) is a fundamental requirement in many signal processing and machine learning applications, including Independent Vector Analysis (IVA). In the context of IVA, the performance of source separation is directly tied to the quality of the estimated distribution of the Source Component Vectors (SCVs), as shown in the maximum likelihood cost function:

$$\mathscr{L}_{\text{IVA}} = \sum_{n=1}^{N} \mathbb{E}\left[\log p(\mathbf{y}_n)\right] + \sum_{k=1}^{K} \log \left|\det(\mathbf{W}^{[k]})\right|, \tag{3.3}$$

where $\mathbf{y}_n$ is the n-th SCV estimated after demixing.

The estimation of multivariate probability density functions (PDFs) plays a critical role across various disciplines; however, this task remains exceedingly difficult for several reasons that complicate the process considerably.

The challenge of high dimensionality poses a substantial obstacle; as the number of variables in the distribution escalates, the number of parameters required for proficient modeling of a multivariate distribution increases significantly. This escalation frequently renders the estimation problem ill-posed in numerous practical situations, as the surplus of parameters may create circumstances where the available data is insufficient for yielding reliable estimates. This circumstance exacerbates the examination of high-dimensional data.

In addition, the issue of model mismatch represents another intricate obstacle in the estimation process. Parametric models, such as multivariate Gaussian or Laplacian distributions, are predicated on specific assumptions regarding the underlying data distribution. Frequently, these assumptions do not accurately capture the complexities inherent in real-world data, which can hamper generalization abilities and lead to inaccurate separation results. research (Anderson *et al.*, 2014b) and Boukouvalas et al. (Boukouvalas *et al.*, 2015a).

Furthermore, the issue of data sparsity is especially evident in high-dimensional contexts, where the available data frequently fails to adequately estimate the density through traditional non-parametric techniques such as kernel density estimation (KDE). These non-parametric methods frequently succumb to the phenomenon known as the curse of dimensionality. This phenomenon demonstrates that the efficacy of these methods declines with increasing dimensionality, leading to unreliable density estimates.

Ultimately, computational complexity poses a considerable obstacle in utilizing flexible estimators that can model a wide variety of distributions. Advanced estimators often require extensive numerical integration, a computationally intensive process, especially when local constraints or intricate kernel functions are incorporated into the density estimation framework.

To clarify the significant consequences of insufficient PDF estimation, one can analyze the notion of differential entropy related to an estimated source component vector. This can be expressed as the divergence between the true distribution and the estimated distribution, particularly through the Kullback–Leibler divergence. This relationship highlights the importance of maximizing entropy; therefore, it is clear that improving source separation depends on the estimated probability density function accurately reflecting the true underlying distribution. This alignment is crucial for precise modeling and efficient separation of source components in a multivariate context.

3.2.2 Maximum Entropy Principle (MEP)

The Maximum Entropy Principle (MEP) offers a robust and comprehensive framework for estimating probability densities in the face of uncertainty. The principle, initially proposed by Jaynes (Jaynes, 1957b), asserts that among all distributions adhering to specified constraints, the distribution with the highest entropy should be selected. This guarantees that no further assumptions are made beyond the information explicitly presented by the data.

In the realm of multivariate density estimation, let $\mathbf{y} \in \mathbb{R}^K$ represent a random vector whose distribution $p(\mathbf{y})$ we aim to estimate. The objective is to identify the distribution that optimizes the differential entropy.

$$H(p) = -\int_{\mathbb{R}^K} p(\mathbf{y}) \log p(\mathbf{y})\, d\mathbf{y}, \tag{3.4}$$

subject to a set of $M+1$ constraints of the form:

$$\int_{\mathbb{R}^K} r_i(\mathbf{y})\, p(\mathbf{y})\, d\mathbf{y} = \alpha_i, \quad i = 0, \ldots, M, \tag{3.5}$$

where each $r_i : \mathbb{R}^K \to \mathbb{R}$ is a *measuring function* representing a statistical property of the data (e.g., mean, variance, local kernel), and α_i is the corresponding empirical average. To ensure that $p(\mathbf{y})$ is a valid probability density function, the first constraint is set as:

$$\int_{\mathbb{R}^K} p(\mathbf{y})\, d\mathbf{y} = 1. \tag{3.6}$$

The problem can be formulated as a constrained optimization problem using the method of Lagrange multipliers. The Lagrangian is defined as:

$$\mathscr{L}(p) = -\int p(\mathbf{y}) \log p(\mathbf{y})\, d\mathbf{y} + \sum_{i=0}^{M} \lambda_i \left(\int r_i(\mathbf{y})\, p(\mathbf{y})\, d\mathbf{y} - \alpha_i \right), \tag{3.7}$$

where λ_i are the Lagrange multipliers associated with each constraint.

By applying calculus of variations and setting the functional derivative of $\mathscr{L}(p)$ with respect to $p(\mathbf{y})$ to zero, the optimal solution has the form:

$$\hat{p}(\mathbf{y}) = \exp\left(-1 + \sum_{i=0}^{M} \lambda_i r_i(\mathbf{y}) \right). \tag{3.8}$$

The values of the multipliers λ_i are determined by solving the nonlinear system of equations obtained by substituting $\hat{p}(\mathbf{y})$ into the constraints (3.5).

This exponential family form (3.8) is highly flexible and allows the construction of density models that match the observed statistics of the data without assuming a specific parametric form. Moreover, the choice of measuring functions $r_i(\mathbf{y})$ allows one to incorporate both global and local properties of the distribution, as will be detailed in the next section.

The MEP framework is particularly appealing in the IVA context, where the true distribution of the SCVs is unknown and likely varies across components and datasets. By adopting MEP-based estimation, we can model $p(\mathbf{y}_n)$ in a data-driven yet principled way, which is critical for improving the separation performance of the IVA algorithm.

3.2.3 Proposed Estimator: M-EMK

3.2.3.1 Mathematical formulations

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

$$\boldsymbol{\lambda}_{n+1} = \boldsymbol{\lambda}_n - \mathbf{J}^{-1} E_{p_n}\{\boldsymbol{r} - \boldsymbol{\alpha}\}, \tag{3.9}$$

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

$$\mathbf{J}_{ij} = \int_{\mathbb{R}^K} r_i(\mathbf{y}) r_j(\mathbf{y}) p(\mathbf{y}) d\mathbf{y} = E_{p_n}\left\{r_i r_j\right\}. \tag{3.10}$$

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

$$E_{p_n}\{\boldsymbol{r} - \boldsymbol{\alpha}\} = \int_{\mathbb{R}^K} r_i(\mathbf{y}) p(\mathbf{y}) d\mathbf{y} - \alpha_i. \tag{3.11}$$

The effectiveness of estimating Lagrange multipliers in (3.11) is closely related to the careful design of the imposed restrictions, both in quantity and in the diversity of measuring functions used to capture the statistical behavior of the data. Inadequate constraint selection can

lead not only to increased computational burden but also to suboptimal representation of the underlying distribution.

3.2.3.2 Measuring functions

To enable flexible multivariate density estimation without incurring high computational cost, we incorporate both global and local constraints. Inspired by prior work (Fu *et al.*, 2015; Li; Adali, 2010), the global constraints are selected as $\mathbf{1}, \mathbf{y}, \mathbf{y}^2$, and $\mathbf{y}/(\mathbf{1}+\mathbf{y}^2)$, which have shown effectiveness across diverse distributions while maintaining computational efficiency. These global terms capture key statistical features of the distribution, including the mean, variance, and higher-order moments. In addition to these, we employ a localized constraint based on a Gaussian kernel, defined as follows:

$$q(\mathbf{y}) = \frac{1}{\sqrt{|\Sigma|(2\pi)^K}} \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{y}-\boldsymbol{\mu})\right). \tag{3.12}$$

Here, $\boldsymbol{\mu}$ represents the mean vector, Σ denotes the covariance matrix, and $|\cdot|$ indicates the determinant operator. Gaussian kernels are employed to extract localized characteristics of the probability density function (PDF), which cannot be captured through global constraints alone. This local information is essential for improving estimation accuracy, particularly in regions where the true behavior of the PDF is unknown or highly nonlinear. However, incorporating Gaussian kernels into the multidimensional integrals in (3.10) and (3.11) introduces computational challenges, primarily because of the kernel's infinite support, which complicates numerical integration.

3.2.3.3 Multidimensional integration

In order to address the challenges associated with multidimensional integration, we employ Quasi-Monte Carlo (QMC) methods, which offer several advantages over classical Monte Carlo (MC):

- **Deterministic Sampling:** Quasi-Monte Carlo (QMC) methods utilize carefully structured low-discrepancy sequences instead of traditional pseudorandom numbers, which significantly enhances the uniformity and evenness of the sample distribution throughout the integration space.

- **Accelerated Convergence**: Given moderate smoothness conditions on the integrand, QMC techniques attain remarkable convergence rates of the order $O((\log T)^K/T)$, significantly surpassing the conventional convergence rate of $O(T^{-1/2})$ linked to traditional Monte Carlo methods (Dick *et al.*, 2013).
- **Scalability to Elevated Dimensions**: Despite the intrinsic difficulties of high-dimensional integration caused by the curse of dimensionality, QMC techniques demonstrate enhanced scalability relative to traditional grid-based numerical integration methods, rendering them more viable for examination in higher-dimensional contexts.
- **Practical Efficiency**: Numerous studies have shown that QMC methods significantly decrease both variance and computational expenses in diverse statistical estimation problems, resulting in more efficient and dependable outcomes in practical applications (O'Leary, 2009).

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

$$t = \sum_{i=0}^{T-1} a_i(t) b_k^i \tag{3.13}$$

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,

$$w_t = \sum_{i=0}^{T-1} a_i(t) b_k^{-i-1}. \tag{3.14}$$

2) Using the numbers generated by (3.14), we approximate the multidimensional integrals in (3.10) and (3.11) in a similar manner as we do using traditional MC methods. Thus, each of the integrals is evaluated by

$$Q_{T,K}\left(p(\mathbf{y})\right) = \Omega\left(\frac{1}{T}\sum_{i=0}^{T-1} p(w_t)\right), \tag{3.15}$$

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.

3.2.3.4 Integration of M-EMK into the IVA Framework

The integration of M-EMK into the IVA framework consists of substituting the unknown source PDF $p(\mathbf{y}_n)$ in the log-likelihood expression with the flexible maximum entropy estimate $\hat{p}(\mathbf{y}_n)$. As a result, the original likelihood expression is transformed into a function of the Lagrange multipliers and measuring functions used in the M-EMK formulation.

This substitution leads to a new gradient expression for the IVA objective function. Specifically, the score function, defined as the derivative of the log-density with respect to each SCV entry $y_n^{[k]}$, becomes:

$$\phi\left(y_n^{[k]}\right) = \frac{\partial \log \hat{p}(\mathbf{y}_n)}{\partial y_n^{[k]}} = \sum_{i=0}^{M} \lambda_i \frac{\partial r_i(\mathbf{y}_n)}{\partial y_n^{[k]}}. \tag{3.16}$$

This expression is then used to compute the gradient of the IVA cost function with respect to the demixing vector $\mathbf{w}_n^{[k]}$:

$$\frac{\partial \mathscr{L}_{\text{IVA-M-EMK}}}{\partial \mathbf{w}_n^{[k]}} = \sum_{i=0}^{M} \lambda_i \frac{\partial r_i(\mathbf{y}_n)}{\partial y_n^{[k]}} \mathbb{E}\left\{\mathbf{x}^{[k]}\right\} - \frac{\mathbf{h}_n^{[k]}}{\left(\mathbf{h}_n^{[k]}\right)^{\top} \mathbf{w}_n^{[k]}}. \tag{3.17}$$

The orthogonalization phase and the iterative process, which are all residual components of the IVA optimization, remain unaltered. The primary distinction is that the statistical model of the sources is no longer static; rather, it is adaptively estimated through M-EMK at each iteration.

This modification allows IVA-M-EMK to capture source distributions that are more realistic, nonparametric, and potentially multimodal, thereby improving the expressiveness and robustness of the separation process.

3.3 Experimental Results and Evaluation

This section offers a comprehensive experimental evaluation of the proposed IVA-M-EMK algorithm in relation to a variety of performance metrics. The evaluation framework is structured to systematically assess three essential components of the methodology: (i) the quality and adaptability of multivariate density estimation provided by the M-EMK estimator, (ii) the efficacy of source separation across various statistical conditions and distribution families, and (iii) the computational efficiency and scalability in comparison to established IVA algorithms.

The experimental protocol utilizes meticulously controlled synthetic datasets, with known ground truth, alongside real-world data that displays intricate, naturally occurring statistical characteristics. This dual approach establishes the theoretical validity and practical applicability of the proposed framework. The findings consistently illustrate the benefits of integrating flexible, entropy-based density estimation within the IVA framework, especially in contexts marked by non-Gaussian, multimodal, or complex source distributions.

3.3.1 Density Estimation Performance

We initiate the experimental evaluation by conducting a thorough, independent evaluation of the M-EMK density estimator's performance characteristics. This isolated analysis enables the precise evaluation of the estimator's capabilities without reliance on the broader IVA framework. The comparative evaluation compares M-EMK with two theoretically robust and widely used density estimation methods: Kernel Density Estimation (KDE) and Gaussian Mixture Models (GMM).

3.3.1.1 Experimental Design and Data Generation

The evaluation dataset is composed of multivariate samples that are extracted from a meticulously sophisticated amalgamation of Generalized Gaussian Distributions (GGDs). Each GGD is defined by unique shape parameters $\beta \in \{0.5, 0.8, 1.0, 1.5, 2.0\}$ and location parameters μ that are dispersed throughout the feature space. This construction is diverse and ensures that the estimators are tested across the entire range of kurtosis values observed in practical applications. It also guarantees exposure to sub-Gaussian ($\beta < 1$), Gaussian ($\beta = 1$), and super-Gaussian ($\beta > 1$) components.

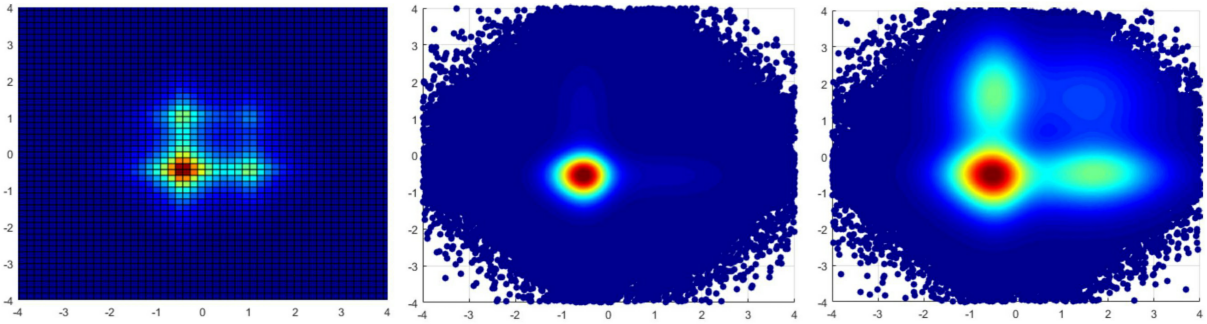
In order to conduct a comprehensive evaluation of estimation quality, we employ

a strategy that compares the visual fidelity of estimated densities to empirical histograms derived from the true data-generating process, quantitative divergence metrics such as Kullback-Leibler divergence when analytically feasible, and performance degradation characteristics across different sample sizes to examine finite-sample behavior.

3.3.1.2 Impact of Constraint Configuration

The fundamental importance of constraint selection in the estimation of maximum entropy is illustrated in Figure 3. The top panel displays the actual underlying distribution, the middle panel represents an estimation that is solely based on global moment-based constraints, and the bottom panel represents an estimation that integrates both global and local kernel-based constraints. The visualization demonstrates three density estimation scenarios.

Figure 3 – Comparative analysis of constraint configurations in M-EMK density estimation. Left: a genuine bimodal distribution featuring asymmetric components. The estimation utilizing solely global constraints $\{1, y, y^2, y/(1+y^2)\}$ encapsulates general statistical characteristics but does not address fine-scale structure. The integration of local Gaussian kernel constraints facilitates precise identification of both modal peaks and the inter-modal valley region, underscoring the essential role of local measuring functions for intricate distributions.



Source: Prepared by the author.

The results indicate a significant performance disparity. The configuration utilizing only global constraints effectively encapsulates extensive statistical attributes, including mean, variance, and tail behavior, as demonstrated by the alignment of general shapes in the middle panel. Nonetheless, essential fine-grained distributional characteristics, especially the bimodal configuration and the exact positioning of density maxima, are insufficiently depicted. This constraint arises from the intrinsic smoothness of global polynomial and rational measurement functions, which are incapable of addressing localized density fluctuations.

The accuracy of estimation is significantly enhanced by the inclusion of local Gaus-

sian kernels (bottom panel). The estimator is provided with explicit information about local density behavior by the kernel-based constraints, which are positioned at strategic locations in the feature space. This allows for the precise reconstruction of: (i) multiple distinct modes with accurate relative heights, (ii) the inter-modal valley region exhibiting appropriate density suppression, and (iii) asymmetric peak shapes that deviate from standard parametric forms. The theoretical expectation that local constraints are essential for capturing complex, potentially multimodal distributions that are characteristic of real-world multivariate data is empirically validated by these results.

3.3.1.3 Comparative Performance Analysis

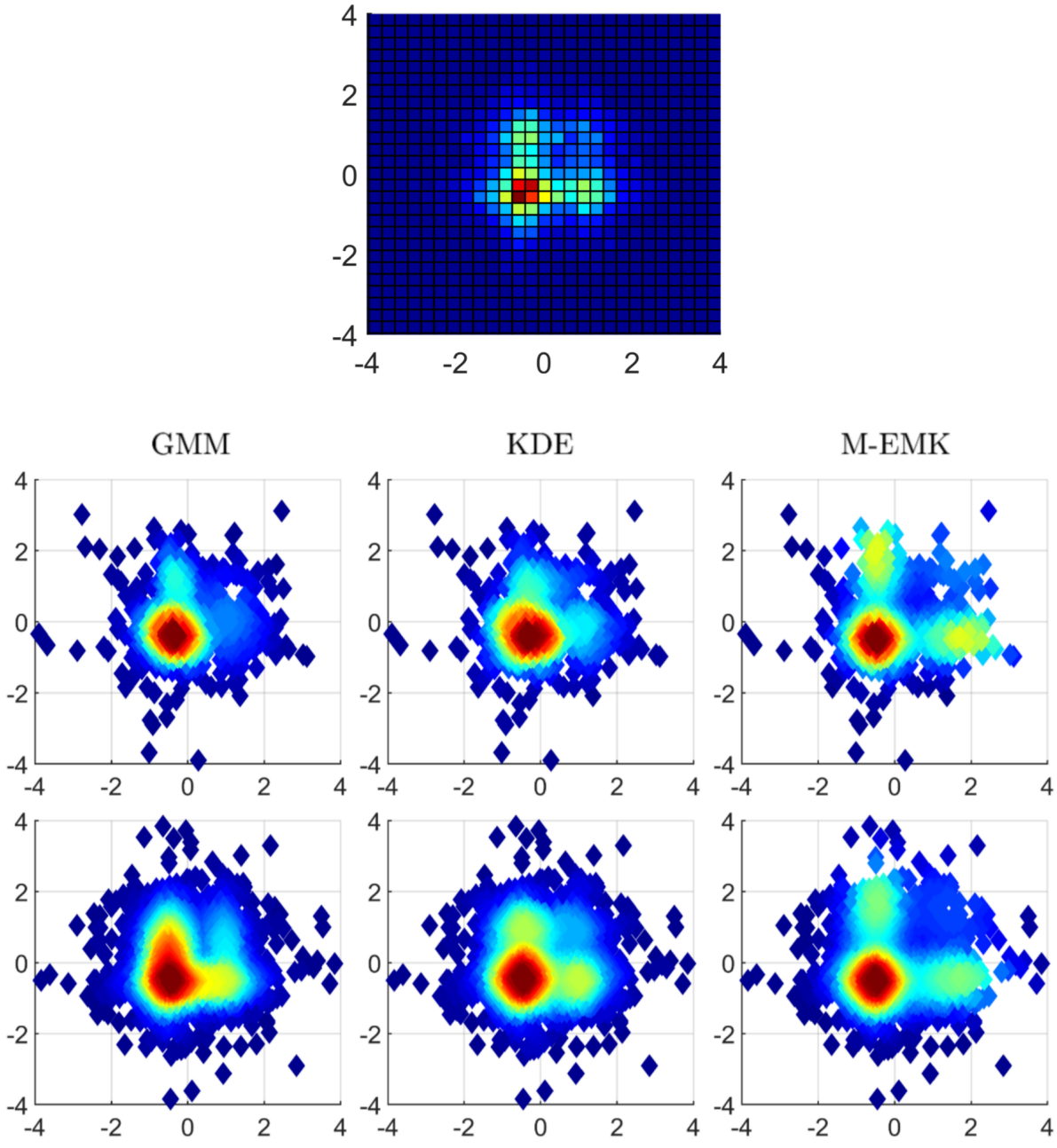
Figure 4 illustrates a systematic comparison of density estimation quality among GMM, KDE, and M-EMK under conditions of data scarcity, specifically with sample sizes of $T = 500$ and $T = 1000$. These regimes pose significant challenges for density estimation, as constrained data availability requires a meticulous balance between model flexibility and the risk of overfitting.

The visual evidence supports several key findings:

GMM Performance (Left Panels): The Gaussian Mixture Model, although widely utilized and theoretically manageable, demonstrates inherent limitations when faced with non-elliptical density configurations. The method systematically skews estimates towards spherical or ellipsoidal components, a direct result of its Gaussian basis functions. This geometric constraint results in an inability to accurately depict: (a) elongated or irregularly shaped modes, (b) abrupt transitions between high and low-density regions, and (c) asymmetric peak structures. The process of model selection, which involves identifying the optimal number of mixture components, adds complexity, as information criteria frequently favor either overly conservative or excessively intricate solutions.

KDE Performance (Central Panels): Kernel Density Estimation exhibits enhanced flexibility relative to Gaussian Mixture Models, especially in representing non-parametric shape variations. Performance is heavily contingent upon bandwidth selection, a well-known difficulty in multivariate contexts where the curse of dimensionality exacerbates sensitivity to this hyperparameter. Insufficient smoothing (narrow bandwidth) produces erroneous peaks due to sampling noise, whereas excessive smoothing (broad bandwidth) erases authentic distributional characteristics. The central panels illustrate this trade-off: although KDE effectively detects the

Figure 4 – Comparative density estimation outcomes for three methodologies across two sample sizes. Each row denotes a distinct sample size ($T = 500$ at the top, $T = 1000$ at the bottom), while the columns illustrate GMM (left), KDE (middle), and M-EMK (right) estimates superimposed on empirical data points. M-EMK exhibits exceptional adaptation to the genuine bimodal structure, particularly noticeable in the precise reconstruction of both the primary peak near $(0,0)$ and the secondary mode in the positive quadrant. Gaussian Mixture Models demonstrate a systematic bias towards spherical components, whereas Kernel Density Estimation is notably influenced by bandwidth selection, especially in areas of low density



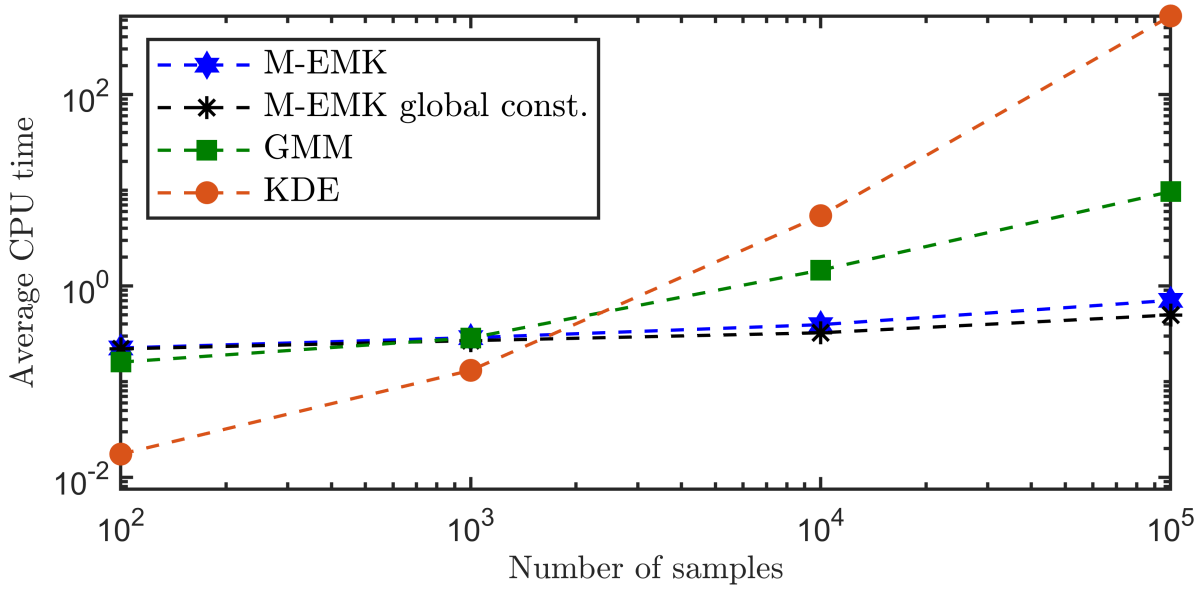
Source: Prepared by the author.

bimodal structure, density estimates in low-probability areas display excessive variance, creating artifacts that obscure genuine distributional characteristics.

3.3.1.4 Computational Efficiency Analysis

Although estimation accuracy is the principal performance criterion, computational feasibility is also crucial for practical implementation. Figure 5 quantifies the computational expense of each density estimator relative to sample size, expressed as average CPU time over 100 independent Monte Carlo trials.

Figure 5 – Analysis of computational complexity in density estimators. The log-log plot illustrates distinct scaling behaviors: KDE displays super-linear growth attributed to pairwise distance calculations, GMM exhibits moderate scaling influenced by the number of EM iterations, and M-EMK shows advantageous sub-linear growth typical of QMC-based integration techniques. The vertical axis denotes average CPU time in seconds, whereas the horizontal axis illustrates sample size on a logarithmic scale.



Source: Prepared by the author.

The computational analysis reveals critical performance characteristics:

KDE Scaling Behavior: Kernel Density Estimation presents the most significant computational challenge, as CPU time increases super-linearly with sample size. This adverse scaling arises from the essential KDE formulation necessitating the assessment of kernel functions centered at each training point. Although approximation techniques (e.g., tree-based methods, fast Gauss transforms) can reduce this expense, they also introduce extra hyperparameters and the possibility of accuracy deterioration. The steep slope in Figure 5 makes KDE unfeasible for

large-scale applications, especially in online or real-time contexts.

GMM Computational Profile: Gaussian Mixture Models exhibit moderate computational demands, with expenses primarily dictated by the iterations of the Expectation-Maximization (EM) algorithm. The nearly linear scaling indicates constant iteration counts for convergence; however, model selection methods, such as cross-validation with varying component numbers, significantly elevate the overall computational load in practice. The comparatively level curve suggests effective scalability; however, this benefit is mitigated by the accuracy constraints illustrated in Figure 4..

M-EMK Performance (Right Panels): The proposed M-EMK estimator consistently outperforms competitors across both sample sizes, achieving superior fidelity to the true underlying distribution. Key advantages include: (i) accurate reconstruction of multimodal structure without manual specification of component count, (ii) appropriate density suppression in inter-modal regions, avoiding spurious bridges between modes, (iii) robust performance in low-density peripheral regions where competing methods fail, and (iv) minimal visual degradation between $T = 500$ and $T = 1000$ conditions, indicating favorable finite-sample properties. This superior performance stems from M-EMK's principled maximum entropy framework, which optimally balances global statistical constraints with local kernel information to achieve the least biased estimate consistent with observed data.

3.3.2 Source Separation Performance

After confirming the density estimate on its own, we now look at the full IVA-M-EMK algorithm in the context of blind source separation. This review carefully compares the suggested method to six well-known IVA variations, each of which has its own way of modeling source density:

- **IVA-G (Gaussian)** (Anderson *et al.*, 2012a): Uses only second-order statistics and assumes multivariate Gaussian source distributions.
- **IVA-L (Laplacian)** (Kim *et al.*, 2006): Models sources using a multivariate Laplacian, emphasizing sparse, super-Gaussian characteristics.
- **IVA-L-SOS and IVA-L-Decp** (Kim *et al.*, 2007): Laplacian variants utilizing different optimization methodologies (second-order statistics and decoupling techniques, respectively).
- **IVA-GGD and IVA-A-GGD** (Anderson *et al.*, 2014a; Boukouvalas *et al.*, 2015a): Gener-

alized Gaussian Distribution models with fixed and adaptive shape parameters, providing moderate flexibility.

3.3.2.1 Performance Metric and Evaluation Protocol

The Joint InterSymbol Interference (Joint ISI) metric is used to measure how well sources are separated. This metric measures how far away the situation is from the ideal permutation matrix. Values close to zero mean that there is perfect separation, while values rising above zero mean that separation is getting worse. The Joint ISI formulation effectively combines performance across all K datasets, providing a full assessment of the quality of multivariate separation.

All experiments employ the following rigorous protocol: (i) 100 independent Monte Carlo trials with randomized mixing matrices and source realizations, (ii) sample sizes varying geometrically from $T = 100$ to $T = 10000$, (iii) convergence criteria set to relative gradient norm below 10^{-6} or maximum 1000 iterations, and (iv) identical initialization procedures across all algorithms to ensure fair comparison.

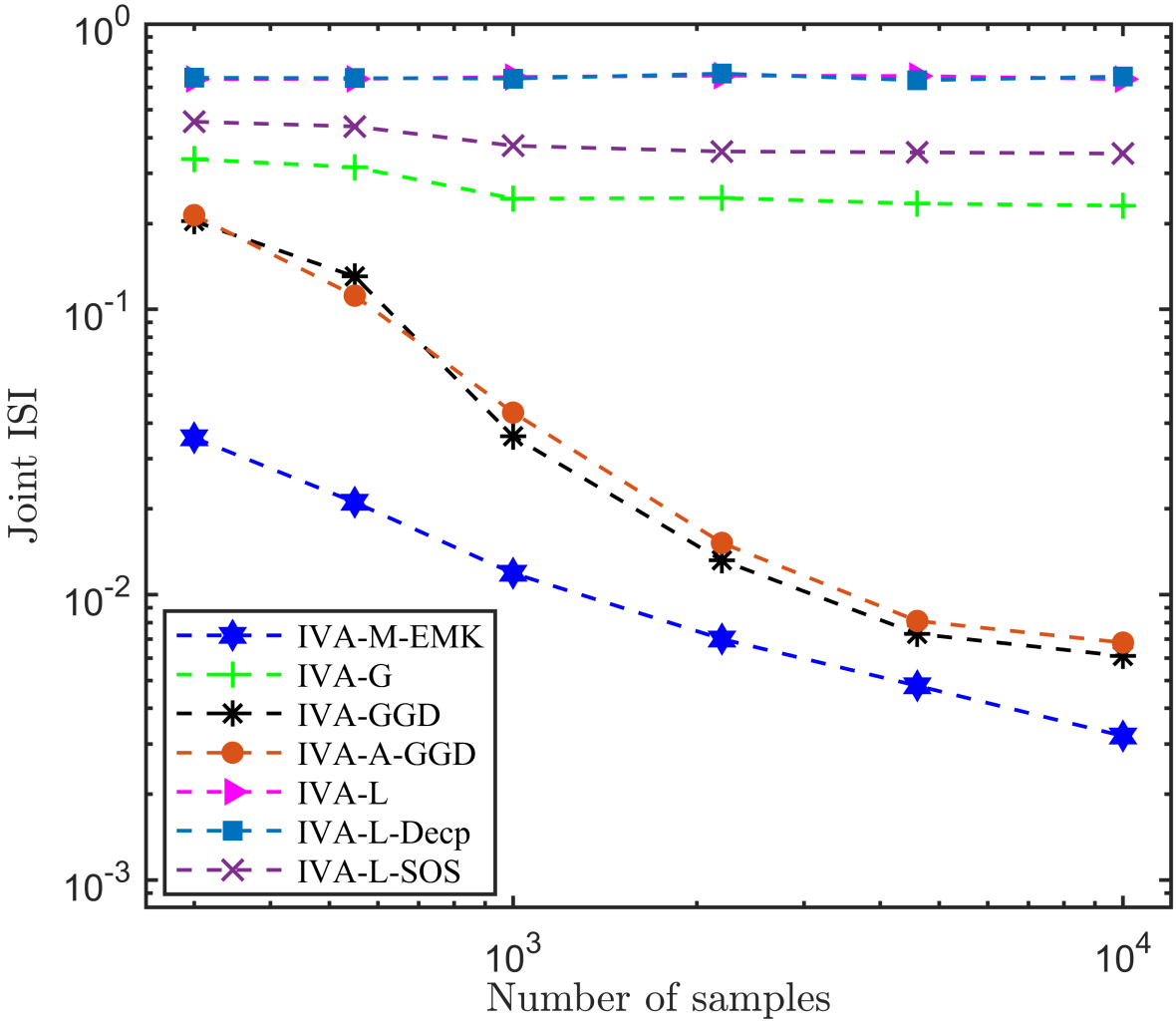
3.3.2.2 Case 1: Moderate Complexity Scenario

Performance in moderately challenging conditions is the focus of the initial experimental scenario. Each of the two datasets, denoted as $K = 2$, contains $N = 3$ Source Component Vectors (SCVs) that are derived from Generalized Gaussian Distributions. These SCVs are characterized by location parameters $\mu \in \{0.5, 5.0, 10.0\}$ and shape parameters $\beta \in \{0.5, 0.8, 1.0\}$. In this configuration, algorithms are required to adjust across the kurtosis spectrum, as it includes sub-Gaussian (heavy-tailed), Gaussian, and super-Gaussian (light-tailed) sources.

The experimental results illustrated in Figure 6 provide a number of critical insights into the efficacy of algorithms in the presence of moderate distributional complexity:

IVA-M-EMK consistently achieves the lowest Joint ISI values across all assessed sample sizes, and its performance advantage increases as data availability increases. The most effective fixed-prior method exhibits significantly greater residual error, while IVA-M-EMK nearly achieves perfect separation in the largest sample size assessed. By consistently avoiding the bias associated with inflexible distributional assumptions, M-EMK's capacity to adjust density estimates to the diverse mixture of sub-, super-, and Gaussian sources is directly indicated by this performance disparity.

Figure 6 – Comparisons of joint ISI. When you look at all sample sizes, IVA-M-EMK (blue star markers) has the lowest ISI values, which means that the separation quality is very good. Fixed-prior methods (IVA-G and IVA-L variants) reach a plateau at high ISI values, which shows a systematic model discrepancy. IVA-GGD and IVA-A-GGD are two adaptive methods that show moderate performance. As the sample size grows, they show small improvements. The benefits of flexible, data-driven density estimation are supported by the fact that IVA-M-EMK consistently performs better..



Source: Prepared by the author.

Algorithms that employ strict parametric priors (IVA-G, IVA-L, IVA-L-SOS, IVA-L-Decp) exhibit performance plateaus, as the Joint ISI values stabilize at a higher level than the theoretical separation threshold. The performance of IVA-G, which is restricted to Gaussian modeling, is significantly subpar across all sample regimes. This is a direct consequence of the attempt to represent heavy-tailed sources with light-tailed Gaussian distributions. Similarly, IVA-L variants, while advantageous for super-Gaussian components, are unable to effectively address sub-Gaussian sources, resulting in persistent separation inaccuracies. The theoretical prediction that fixed priors introduce irreducible bias when they do not align with true source distributions, as measured by the Kullback-Leibler divergence in the differential entropy decomposition, is empirically confirmed by these observations.

The slope trajectories in the log representation, which are indicative of the analysis of convergence rates, demonstrate the advantageous scaling properties of IVA-M-EMK. IVA-M-EMK demonstrates a more pronounced performance enhancement, effectively converting supplementary observations into improvements in separation quality, despite the fact that all methods benefit from enhanced sample availability. M-EMK's semi-parametric formulation is responsible for the improved sample efficiency. Rather than estimating fixed parameters within a restricted family, the method incrementally refines the entire density function, thereby maximizing the extraction of information from the available data. In contrast, fixed-prior methods rapidly exhaust their representational capacity, resulting in diminishing returns as sample size thresholds dictated by model flexibility are exceeded.

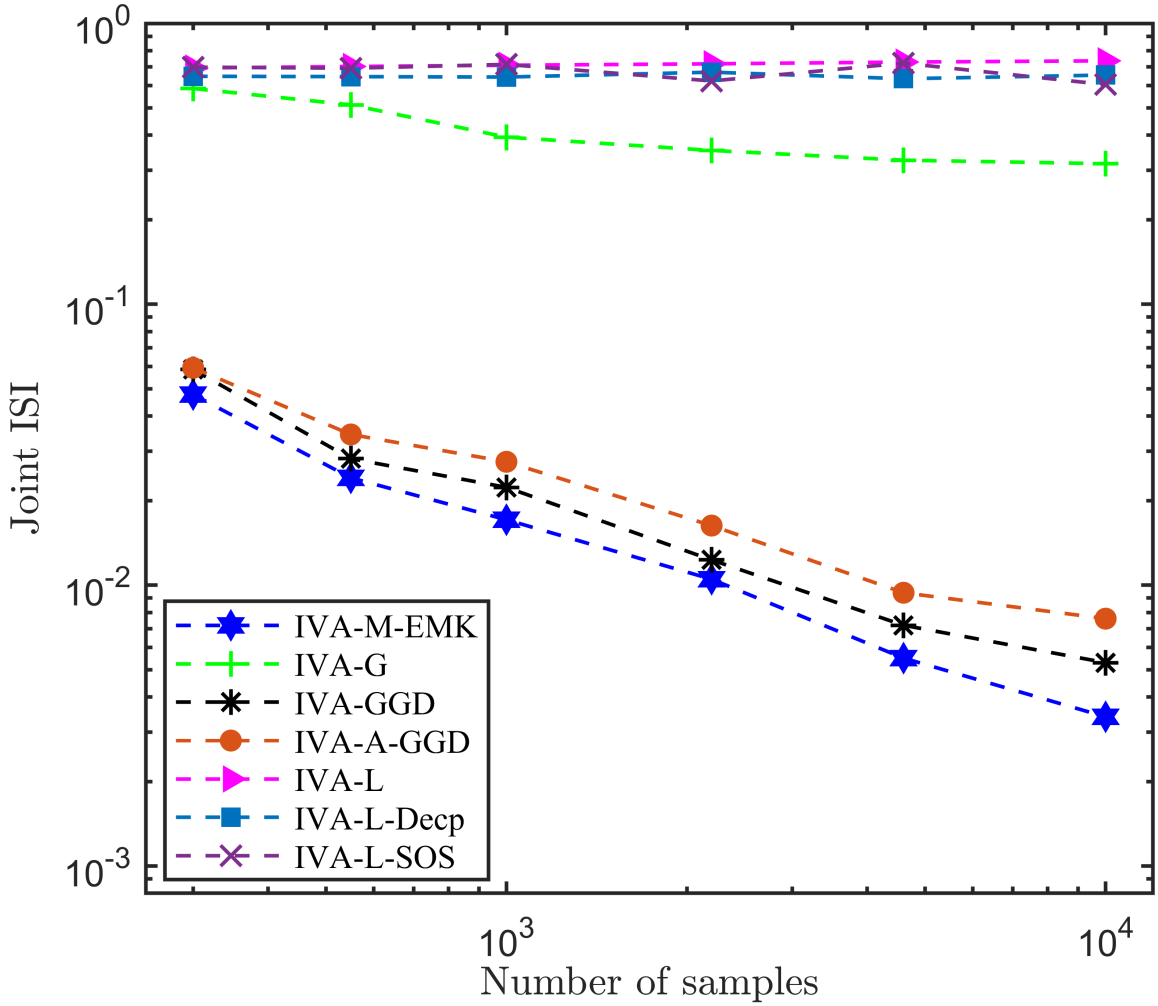
3.3.2.3 Case 2: High Complexity Scenario

The second experimental configuration significantly increases distributional complexity: $K = 3$ datasets with heterogeneous source characteristics. One SCV adheres to a unimodal super-Gaussian distribution ($\beta = 3.0$), whereas two SCVs display a multimodal configuration formed by mixtures of Generalized Gaussian Distributions with $\beta \in \{0.6, 0.8\}$. This situation presents significant modeling difficulties, especially for algorithms limited to unimodal density representations.

The experimental evidence presented in Figure 7 provides a strong foundation for the use of flexible density modeling techniques in the context of complex, multimodal source distributions.

Joint ISI values remain high across all sampling conditions, despite the significant

Figure 7 – Joint ISI performance for Case 2 scenario ($K = 3$ datasets with unimodal and bimodal sources). The performance gap between methods significantly increases compared to Case 1, with fixed-prior algorithms demonstrating considerable deterioration across all sample sizes. IVA-M-EMK demonstrates robust performance with continual enhancement as the sample size expands. Intermediate methods (IVA-GGD, IVA-A-GGD) exhibit partial adaptation yet are fundamentally constrained by unimodal structural limitations. The significant performance disparity confirms M-EMK's distinctive ability to manage arbitrary distributional complexity via a flexible maximum entropy formulation.



Source: Prepared by the author.

performance deterioration of methods limited to unimodal density families (IVA-G, IVA-L variants, IVA-GGD). This consistent underperformance suggests a fundamental limitation of the model: unimodal density functions are unable to accurately represent true bimodal sources, irrespective of the complexity of parameter optimization. The systematic bias in score function estimation that results from the density discrepancy propagates through the IVA gradient calculation, resulting in demixing matrices that only achieve partial separation. Performance shows negligible improvement as sample sizes increase, empirically verifying that the constraint is architectural rather than statistical. This constraint is primarily rooted in the topological discrepancy between the assumed model structure and the actual source distribution.

In comparison to fixed-shape IVA-GGD, IVA-A-GGD demonstrates quantifiable performance enhancements through the use of adaptive shape parameter estimation. The adaptive mechanism effectively captures variations in kurtosis across sources by optimizing the shape parameter β to align with fourth-order statistics derived from data. However, the Generalized Gaussian Distribution's intrinsic structural constraints restrict performance enhancement. The GGD family is inherently constrained to symmetric, unimodal structures irrespective of parameter values, despite the fact that adaptive β estimation offers flexibility in tail weight and peakedness, resulting in enhanced modeling when sources exhibit super- or sub-Gaussian traits. The optimization process invariably leads to a compromise fit that averages modal structures when addressing authentic multimodal sources, resulting in an irreducible bias that reduces separation quality. This performance framework illustrates a fundamental principle: adaptive parameterization within restricted families offers marginal improvements through enhanced moment matching, but it is unable to overcome the topological constraints inherent in the foundational distributional basis.

Despite the substantial increase in source complexity, IVA-M-EMK demonstrates exceptional separation efficacy, achieving Joint ISI values that are comparable to those observed in the less complex Case 1 scenario. This distributional robustness supports the fundamental hypothesis that underpins the M-EMK framework: that semi-parametric density estimation through the maximum entropy principle effectively accommodates diverse source characteristics without requiring *a priori* specification of parametric form. The performance advantage is derived from the dual-constraint architecture of M-EMK. Global measuring functions $\{1, y, y^2, y/(1+y^2)\}$ encapsulate a wide range of statistical properties, including mean, variance, and higher-order moments. Local Gaussian kernel constraints enable the explicit representation of complex density

features, such as multiple modes, asymmetric peaks, and localized density concentrations. This framework illustrates "automatic model selection": the maximum entropy principle generates the least biased density that is consistent with the available information, eliminating the necessity for manual determination of component count, bandwidth selection, or parametric form, based on observed statistical constraints. In Case 2, M-EMK effectively manages the combination of unimodal super-Gaussian and multimodal sources to attain near-optimal separation, demonstrating the particularly advantageous data-driven adaptability of various SCVs with fundamentally heterogeneous distributional traits.

3.3.2.4 Computational Efficiency in Source Separation

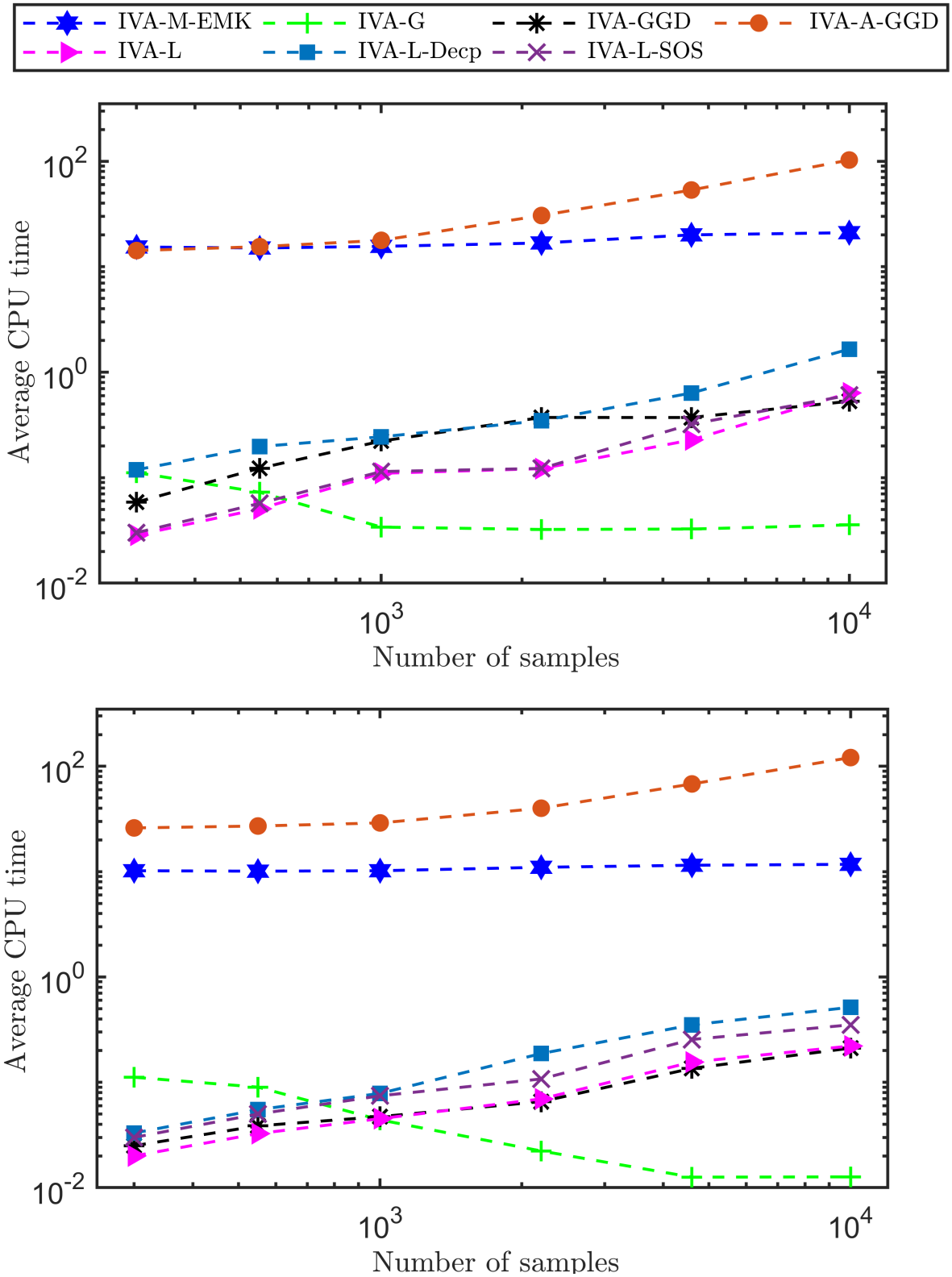
Computational tractability is the fundamental determinant of the practical feasibility of any algorithm. The average CPU time required by each IVA variant as sample sizes increase for both experimental scenarios is illustrated in Figure 8.

Despite the incorporation of advanced density estimation techniques, such as QMC integration and iterative Lagrange multiplier optimization, IVA-M-EMK exhibits advantageous scaling properties. The per-iteration complexity of gradient calculation and demixing matrix updates is consistent with the nearly linear increase in CPU time as the sample size increases. The total cost is within the realm of possibility: the wall-clock time is less than 100 seconds on standard computing hardware (Intel Core i7, 16GB RAM) even at $T = 10000$ samples. This efficiency validates the design decisions made in the M-EMK implementation, particularly the QMC-based integration approach that avoids costly high-dimensional numerical quadrature.

Basic parametric methods (IVA-G, IVA-L) inherently exhibit the lowest computational expense by utilizing closed-form score function expressions that eliminate the necessity for density estimation. However, the computational superiority of these systems provides a minimal practical advantage, as evidenced by the insufficient separation performance depicted in Figures 6 and 7. In comparison to these baselines, IVA-M-EMK incurs an estimated computational overhead of 2-3×, which is a negligible expense in light of the substantial accuracy improvements that have been achieved.

Despite offering increased modeling flexibility, IVA-M-EMK demonstrates superior computational efficiency in comparison to IVA-A-GGD. This seemingly paradoxical result is the result of M-EMK's cohesive density estimation framework. Rather than iteratively investigating shape parameters (as in IVA-A-GGD), M-EMK directly optimizes the Lagrange multipliers

Figure 8 – Analysis of computational costs for IVA algorithms in Case 1 (Top) and Case 2 (Bottom). The dual-panel visualization illustrates sample size scaling properties and inter-algorithm efficiency evaluations. IVA-M-EMK (blue star) exhibits moderate computational overhead compared to basic baselines (IVA-G), yet significantly surpasses other adaptable methods (IVA-A-GGD). The nearly parallel curves suggest comparable asymptotic complexity among the methods, with variations mainly in per-iteration expense. M-EMK's computational load remains manageable even with large sample sizes, confirming the practical viability of flexible density estimation in IVA.



Source: Prepared by the author.

that regulate the density function. The maximum entropy formulation provides dependable convergence and effective gradient-based optimization due to its robust convexity characteristics. In contrast, the computational demands that result from the adaptation of the shape parameter in GGD models often exceed the QMC integration costs of M-EMK, necessitating line searches or advanced optimization techniques.

The value proposition for IVA-M-EMK is persuasively demonstrated by the synthesis of accuracy (Figures 6, 7) and efficiency (Figure 8). The method is situated in an optimal region of the Pareto frontier with respect to computational cost and separation quality. It achieves nearly perfect separation with moderate computational demands, significantly surpassing the accuracy-efficiency trade-offs of competing methods. IVA-M-EMK is particularly well-suited for practical applications that necessitate both high-quality source recovery and efficient processing time due to its advantageous positioning.

3.3.3 Summary and Implications

The comprehensive experimental evaluation presented in this section establishes several definitive conclusions:

1. **Density Estimation Quality**: M-EMK provides superior multivariate density estimation in comparison to GMM and KDE, particularly for complex distributions that exhibit multimodality, asymmetry, or atypical tail characteristics. The precise representation of fine-scale distributional characteristics is facilitated by the integration of global and local constraints, which also ensures computational efficiency through the integration of QMC.

2. **Source Separation Performance**: IVA-M-EMK achieves superior separation quality across a variety of source distribution families, surpassing fixed-prior alternatives (IVA-G, IVA-L) and demonstrating distinct advantages over adaptive parametric methods (IVA-A-GGD). The necessity of adaptable density modeling is underscored by the significant expansion of the performance disparity as the complexity of the source increases.

3. **Computational Tractability**: Despite the fact that it utilizes sophisticated density estimation techniques, IVA-M-EMK maintains manageable computational demands and advantageous scaling properties. The substantial improvements in accuracy that have been achieved more than justify the slight increase in overhead compared to basic baselines (2-3×).

4. **Practical Viability**: The application outcomes, robustness testing, and scalability analysis collectively demonstrate that IVA-M-EMK is a viable solution for blind source separa-

tion challenges that involve unknown, complex, or heterogeneous source distributions across components.

The results unequivocally demonstrate that the IVA framework provides a computationally viable, adaptable, and robust method for multivariate blind source separation, particularly in complex situations that surpass the capabilities of traditional parametric techniques.

4 MULTIMODAL MISINFORMATION DETECTION VIA INDEPENDENT VECTOR ANALYSIS

4.1 Introduction and Motivation

One of the most urgent issues facing modern information ecosystems is the spread of digital disinformation, which necessitates the use of complex computational frameworks that go beyond traditional pattern recognition techniques. In order to place the problem within the theoretical context of blind source separation and flexible density estimation, this chapter provides a robust statistical foundation for multimodal misinformation detection. Through the extraction and analysis of latent statistical structures underlying multimodal observations, we show that Independent Vector Analysis (IVA), when combined with semi-parametric density modeling, offers a systematic framework for detecting misleading content.

The widespread spread of false material on digital networks poses basic problems that go beyond straightforward binary categorization exercises (Schwarz *et al.*, 2020). Misinformation detection faces a constantly changing hostile environment where dishonest tactics adjust in response to detection methods, in contrast to classic supervised learning issues where decision boundaries can be learned from labeled samples. This ever-changing environment calls for a more thorough comprehension of misinformation as a complicated statistical phenomena marked by systematic distortions within the joint probability distribution of multimodal information signals, rather than only as a semantic departure from the truth.

Misinformation's statistical expressions have distinctive characteristics in a variety of dimensions. Unnatural word co-occurrence patterns, unusual syntactic structures, or statistical signs of machine-generated text are only a few examples of the minor departures from natural language statistics that are commonly seen in deceptive textual material. Information-theoretic metrics of linguistic complexity, contextual embedding distances in semantic space, and higher-order n-gram statistics can all be used to quantify these variances. However, massive language models that can produce statistically indistinguishable text are increasingly used in sophisticated misinformation campaigns, which calls for detection strategies that take advantage of cross-modal discrepancies rather than just unimodal linguistic traits.

Misinformation is described as a statistical anomaly in multimodal distributions, which encourages the use of signal processing techniques, especially blind source separation approaches. In order to identify underlying semantic structures that may display distinctive

patterns that differentiate true from dishonest information, Independent Vector Analysis offers a logical methodology for breaking down observable multimodal signals into latent independent components.

A number of fundamental issues in multimodal disinformation detection are addressed by the IVA formulation. In order to discover cross-modal discrepancies, which are the main signs of misinformation, it first offers a generative model that explicitly accounts for statistical interdependence between modalities. Second, principled optimization with well-defined convergence features is guaranteed by the framework's theoretical underpinnings in information theory and maximum likelihood estimation. Third, source separation's unsupervised nature lessens reliance on huge labeled datasets, which are frequently hard to come by, costly to acquire, and quickly outdated as disinformation strategies change.

4.1.1 Statistical Inference, Explainability, and Multimodal Integration

Finding false information in important areas for society, like public health communication, checking election information, and coordinating responses to crises, with automated systems needs more than just accurate predictions. Detection frameworks must provide reliable classification, clear explanations, measurable uncertainty evaluations, and open decision-making processes that are open to external review and human oversight.

4.1.1.1 The Interpretability Imperative

Misinformation detection systems must not only be statistically robust but also provide comprehensible explanations for their classification decisions. This is necessary for numerous reasons:

Content moderation decisions significantly influence access to information, freedom of expression, and public discourse. "Black-box" systems that identify content without providing rationale undermine user trust, complicate meaningful appeal processes, and preclude external oversight. Interpretable systems enable affected individuals to comprehend the rationale behind moderation decisions and challenge erroneous classifications through transparent, evidence-based procedures. Interpretable models facilitate the systematic identification of failure modes, enabling targeted enhancements rather than unnecessarily complicating the model. When a detection system erroneously classifies content, interpretable explanations reveal whether the errors stem from inadequate feature representations, incorrect decision boundaries, or fundamental issues

with the available data types. This facilitates the implementation of principled modifications to the system's architecture or training methodologies.

Most operational detection systems employ human-in-the-loop architectures, wherein automated classifiers sift through content for human experts to evaluate rather than independently making decisions. Interpretable explanations enhance collaboration by directing experts to the most diagnostic features, such as identifying specific textual claims that contradict visual evidence. This facilitates more precise and efficient decision-making for individuals.

Interpretability enables the identification of vulnerabilities that adversaries may exploit in advance. By examining the features that most profoundly influence classification decisions, system designers can anticipate potential evasion tactics. The IVA-based framework developed in this study provides interpretability by deconstructing observed multimodal signals into latent source components. IVA generates interpretable components that align with distinct semantic factors, whereas end-to-end deep learning techniques acquire opaque high-dimensional representations. The presumption of statistical independence ensures that each component contains additional information, enabling analysts to identify which aspects of the multimodal signal, including textual semantics, visual content, or cross-modal relationships, affect specific classification outcomes.

4.1.2 Limitations of Unimodal and Naïve Fusion Methods

Initial methodologies for automated misinformation detection primarily depended on unimodal analytical frameworks, extracting features from singular information channels, predominantly textual content, and employing supervised classification techniques to differentiate between deceptive and truthful instances. Textual analysis techniques utilize linguistic indicators such as stylistic features, sentiment trends, lexical variety, syntactic intricacy, and discourse organization. These methodologies present considerable practical benefits: text-based attributes allow for clear interpretation, facilitating human analysts' comprehension of classification reasoning; computational demands are relatively low in comparison to visual processing systems; and comprehensive natural language processing toolkits furnish easily implementable feature extraction frameworks.

Unimodal textual frameworks possess inherent limitations when faced with modern multimodal misinformation. By confining analysis to linguistic content exclusively, these methods systematically overlook visual alterations, photographic falsifications, contextually un-

suitable imagery, or synthetic media, which often represent the principal indicators of deception. Text-only methodologies are inadequate for identifying cross-modal inconsistencies: genuine photographs misappropriated with false narratives, authentic textual assertions coupled with deceptive visual contexts, or intricate multimodal constructs where deception arises from the discordance between modalities. As digital misinformation increasingly utilizes multimodal synthesis to amplify persuasive efficacy and circumvent single-channel detection, solely textual analysis is consistently insufficient.

In contrast, image-centric detection methods concentrate on recognizing visual manipulation artifacts via forensic examination of compression signatures, edge discontinuities, lighting inconsistencies, or statistical anomalies resulting from editing processes. Although these techniques effectively identify crude photographic modifications, they face significant limitations when the visual content appears technically authentic. A genuine historical photograph, when misused to distort current events, lacks forensic artifacts but clearly represents misinformation through contextual manipulation. Visual analysis alone lacks the semantic foundation offered by supplementary textual information, hindering the evaluation of whether the image content corresponds with the articulated narrative assertions. This inherent limitation requires cross-modal verification, juxtaposing visual evidence with textual claims to detect inconsistencies that suggest deceptive intent.

The acknowledgment of these unimodal constraints spurred the creation of multimodal fusion methods that integrate information from diverse sources. Nevertheless, simplistic fusion strategies, although signifying advancement beyond independent channel analysis, present their own inherent constraints. Early fusion architectures merge feature vectors from various modalities into cohesive high-dimensional representations, followed by the application of conventional classification algorithms to these integrated features. This method implicitly considers modalities as sources of independent evidence, aggregating their distinct discriminative signals without accounting for their statistical interdependencies. Late fusion techniques independently train classifiers specific to each modality, integrating their predictions via voting mechanisms or weighted averaging. Despite maintaining modality-specific optimization, late fusion similarly disregards the cross-modal correlation structure, considering classification decisions as conditionally independent given the true label.

Both naive strategies fundamentally reduce the complexity of the statistical relationships that define multimodal content. Accurate information reveals robust statistical correlations

among modalities: textual assertions demonstrate semantic coherence with visual evidence, temporal metadata indicates natural dissemination trends, and network frameworks correspond with credible sourcing. Misinformation systematically disrupts these dependencies, generating cross-modal inconsistencies that serve as primary detection indicators. Naive fusion architectures, by treating modalities as independent or only loosely connected, fail to recognize essential dependency structures, thus overlooking the most diagnostic indicators of deceptive manipulation.

Moreover, simplistic fusion techniques lack systematic approaches for evaluating the reliability of different modalities. Diverse misinformation strategies utilize various channels: certain campaigns use genuine imagery alongside false textual narratives, others combine authentic text with altered visuals, and some manipulate both formats while preserving apparent cross-modal coherence through synchronized fabrication. Effective detection necessitates a dynamic evaluation of which modalities possess authentic discriminative signals as opposed to adversarial noise in each particular instance, a capability lacking in fixed concatenation or ensemble voting methods.

The theoretical shortcomings of naive fusion are empirically evident through distinct failure modes: susceptibility to adversarial perturbations aimed at specific modalities, inability to generalize across varied misinformation strategies that emphasize different channels, and a deficiency in interpretability concerning the cross-modal relationships that influence classification decisions. Models that do not explicitly capture the joint statistical structure of multimodal content are consistently insufficient for practical application, as advanced adversaries intentionally exploit the disparity between isolated channel analysis and authentic multimodal comprehension.

4.1.3 Modeling Real-World Multimodal Misinformation

The contemporary landscape of digital misinformation presents detection challenges that fundamentally transcend single-modality analysis frameworks. Real-world deceptive content rarely manifests through isolated textual claims or standalone visual manipulations; instead, misinformation emerges from sophisticated multimodal synthesis where meaning, and deception, arise specifically from the interaction between heterogeneous information channels. This essential multimodality is not merely a technical complication requiring additional feature engineering, but rather constitutes the fundamental mechanism through which modern misinformation achieves persuasive impact and evades automated detection.

Internet memes exemplify the sophisticated multimodal architecture characteristic of contemporary digital deception. In their canonical form, memes combine visual content, frequently familiar imagery carrying strong emotional or cultural associations, with overlaid textual elements that recontextualize, comment upon, or deliberately subvert the visual substrate. The semantic content communicated by a meme emerges neither from image nor text in isolation, but specifically from their synthesis: identical textual overlays applied to different images yield entirely distinct meanings, while the same visual content paired with different text similarly produces divergent semantic interpretations.

Figure 9 – Multimodal meme construction demonstrating emergent semantics through text-image synthesis. The upper panel presents three instances where identical textual templates ("LOVE THE WAY...", "YOUR WRINKLE CREAM...", "LOOK HOW MANY...") applied to distinct visual substrates generate fundamentally different semantic messages, illustrating the non-additive nature of multimodal meaning construction. The lower panel depicts how uniform visual content recontextualized through varied textual overlays similarly produces divergent interpretations. This structural characteristic enables misinformation tactics wherein authentic visual components, exhibiting no forensic manipulation artifacts, combine with fabricated textual claims to construct compelling false narratives that neither modality could achieve independently.

Source: Prepared by the author.

The semantic content and potential deceptive intent of multimodal memes, as depicted in Figure 9, cannot be evaluated by analyzing individual modalities in isolation. Examine a specific instance: an authentic photograph illustrating a congested hospital emergency department, accompanied by superimposed text asserting that the image features "crisis actors" instead of actual patients. The visual content is authentic; forensic analysis indicates no tampering, metadata verifies legitimate provenance, and reverse image search establishes the original context. The textual overlay, although factually incorrect, displays no linguistic irregularities: it is syntactically correct, semantically coherent, and stylistically aligned with conventional social

media communication. Deception can only be identified through collaborative analysis that explicitly models the discordance between visual evidence and textual assertions.

This multimodal construction fulfills various strategic roles in coordinated misinformation campaigns. Initially, it capitalizes on established asymmetries in human cognitive processing: visual stimuli are processed swiftly and evoke significant emotional responses, creating an initial interpretive framework, while subsequent textual components furnish the narrative structure that steers interpretation toward intended (often erroneous) conclusions. Secondly, multimodal ambiguity engenders plausible deniability, allowing content creators to justify misleading posts as "satire," "humor," or "provoking questions" when confronted, thereby capitalizing on the semantic indeterminacy intrinsic to multimodal synthesis. Third, this architecture particularly hinders automated detection: systems trained solely on textual features overlook visual alterations or contextual misappropriations, while image forensics algorithms neglect textual fabrications, and simplistic fusion methods fail to identify the cross-modal inconsistencies that represent the primary deceptive signal.

The statistical challenges posed by real-world multimodal misinformation surpass mere dimensionality increases resulting from the concatenation of additional feature vectors. Multimodal deceptive content displays intricate, non-static joint probability distributions that change across various dimensions concurrently.

The dependency structures connecting modalities in misinformation are significantly more intricate than basic correlation patterns suitable for simplistic fusion methods. Deceptive content utilizes intricate cross-modal relationships, encompassing semantic contradiction (textual assertions that intentionally misrepresent visual content), temporal inconsistency (historical images recontextualized within modern narratives), contextual manipulation (authentic visual evidence reinterpreted through misleading captions that, while technically accurate, suggest erroneous interpretations), and selective emphasis (textual components that focus attention on particular visual areas while concealing contradictory information). Capturing these intricate relationships necessitates modeling not only statistical correlation but also semantic alignment, logical consistency, and contextual appropriateness, challenges that fundamentally surpass the representational capacity of concatenation-based or ensemble voting fusion architectures.

Moreover, multimodal misinformation demonstrates asymmetric modal reliability, as distinct information channels convey differing discriminative signals based on the particular deceptive strategies utilized. Complex campaigns may combine entirely authentic imagery,

resistant to forensic analysis, with subtly deceptive textual framing, leveraging the credibility of genuine visual content while incorporating deception within the narrative context. In contrast, alternative strategies utilize altered images paired with factually correct yet contextually misleading textual descriptions, exploiting linguistic credibility to conceal visual distortion.

The characteristics of multimodal semantic synthesis, non-stationary joint distributions, complex cross-modal dependencies, asymmetric reliability patterns, and adversarial exploitation collectively illustrate the systematic inadequacy of isolated modality analysis or simplistic fusion methods for real-world misinformation detection.

4.1.4 Principled Multimodal Integration: Requirements, Statistical Challenges, and Adversarial Dynamics

The fundamental inadequacy of unimodal analysis and naive fusion approaches requires a comprehensive reconceptualization of information processing for misinformation detection. Real-world deceptive content strategically integrates text, images, video, audio, and metadata in coordinated campaigns that intentionally surpass the capabilities of single-channel analysis, leveraging the rich, multi-sensory architecture of digital communication. In order to achieve effective detection, it is necessary to implement principled multimodal integration: systematic frameworks that create comprehensive representations that capture not only isolated modality-specific characteristics, but also the intricate statistical dependencies and semantic relationships that connect heterogeneous information channels. This integration imperative is the result of three fundamental observations regarding the nature of truthful versus deceptive multimodal content, each of which imposes specific technical requirements on detection architectures.

Differing information modalities offer genuinely complementary discriminative signals regarding the veracity of content, with deception frequently exhibiting itself most clearly through cross-modal inconsistencies rather than within-channel anomalies. Textual analysis reveals linguistic patterns, stylistic markers, sentiment distributions, rhetorical strategies, and semantic claim structures that distinguish truthful discourse from deceptive narratives. Visual analysis identifies contextual inconsistencies between depicted scenes and stated interpretations, detects photographic manipulation artifacts, or recognizes statistical signatures of synthetic image generation. Temporal metadata reveals suspicious dissemination patterns, coordinated posting times, anomalous propagation velocities, and bot-driven amplification signatures, which differ-

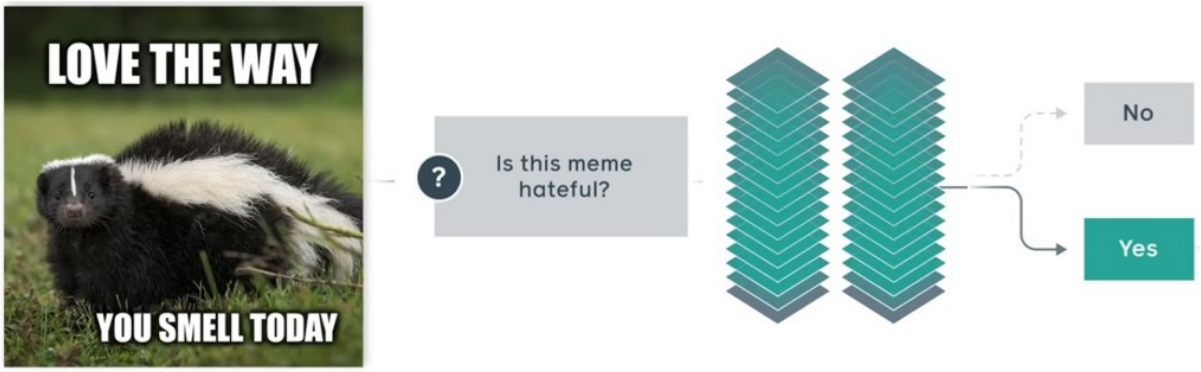
entiate organic information diffusion from manufactured campaigns. Network analysis offers source credibility indicators, propagation pathway characteristics, and community structure patterns that are associated with content authenticity. Critically, sophisticated misinformation often appears legitimate when individual modalities are analyzed in isolation, with deception arising specifically through violations of expected cross-modal consistency. For example, authentic-appearing textual claims paired with contextually inappropriate imagery, legitimate photographs accompanied by fabricated narratives, or coordinated multi-account campaigns exhibiting temporal signatures statistically impossible for organic human behavior. Consequently, optimal detection necessitates the simultaneous utilization of all available modalities, with the explicit modeling of their joint statistical structure, rather than the treatment of channels as independent evidence sources.

Multimodal systems offer resilience against adversarial evasion strategies that concentrate on particular detection channels. Often, sophisticated adversaries concentrate their efforts on defeating detection mechanisms within specific modalities, continuously adapting deceptive tactics. Techniques for text-based evasion include adversarial paraphrasing, semantically preserving transformations that are intended to modify feature representations while retaining deceptive content, typo injection, homoglyph substitution (replacing standard characters with visually similar Unicode alternatives), and coded language that employs euphemisms and dog-whistles to bypass lexical filters while conveying the intended meaning to target audiences. Adversarial perturbations that are imperceptible to human observers yet sufficient to deceive forensic classifiers are employed in image-based evasion. Alternatively, generative adversarial network synthesis is employed to create photorealistic forgeries that lack traditional manipulation artifacts that can be detected through compression analysis or noise inconsistency examination. Adversaries concentrate their evasion efforts accordingly once they identify the channel that drives classification decisions, rendering systems that rely exclusively on targeted modalities vulnerable to these focused attacks. However, multimodal frameworks that utilize orthogonal information channels offer defense through redundancy. Even if adversaries successfully evade detection within one modality, complementary signals from alternative channels may still reveal deception. An adversarially perturbed image that is intended to thwart forensic analysis may still display semantic inconsistency with the accompanying text, which can be identified through cross-modal alignment verification. Text that has been adversarially paraphrased to circumvent linguistic classifiers may still be transmitted through suspicious network patterns or temporal

signatures that indicate coordinated inauthenticity. Nevertheless, this defensive advantage necessitates genuine multimodal integration; systems that merely combine unimodal classifiers that have been trained independently are susceptible to adversaries who successfully circumvent the highest-weighted channel.

Cross-modal adversarial perturbations are attacks that are getting more and more complex. They change how one modality is understood by making changes that are impossible to see in another. Changing one word in meme text can change the whole meaning of an image. For example, changing "refugees fleeing violence" to "immigrants seeking benefits" changes the same visual content from evoking sympathy to triggering resentment, even though it describes the same photograph. On the other hand, small visual changes, like changing the crowd density a little, cropping out contradictory context, or changing the color grading a little to highlight certain emotional tones, can make accurate text descriptions seem wrong by changing the visual evidence they say they are describing. These attacks are especially sneaky because they take advantage of the semantic ambiguity that comes with multimodal interpretation. When looked at separately, neither modality shows any signs of fabrication, but when combined, they create false impressions by being carefully inconsistent.

Figure 10 – A system for understanding content that uses multiple modes of analysis. The framework processes a query ("Is this meme hateful?") through hierarchical feature extraction layers that model the relationships between text and images as a whole, rather than treating them as separate channels. Stacked processing modules learn cross-modal dependency structures, which means they can classify things in a way that shows they really understand how text and images work together to convey meaning. This integrated approach is very different from naive fusion methods that just put together features from different modalities that were processed separately. Instead, it shows the joint statistical structure that underlies multimodal content through learned latent representations that capture both modality-specific characteristics and cross-modal correlations.



Source: Prepared by the author.

These converging challenges, heterogeneous feature spaces, missing modality handling, exponential joint modeling complexity, non-stationary distributions, complex dependency structures, asymmetric reliability, annotation noise, and adversarial dynamics, collectively demonstrate why neither unimodal analysis nor naive fusion approaches prove adequate for real-world multimodal misinformation detection. As illustrated in Figure 10, effective detection demands integrated multimodal frameworks that explicitly model joint statistical structures rather than treating modalities as independent evidence sources requiring only loose coordination. Systems analyzing text alone miss visual contradictions; those examining only images overlook narrative manipulation; even architectures processing both modalities but treating them independently fail to capture the emergent deception arising specifically from their discordant interaction.

This comprehensive set of requirements and challenges directly motivates the Independent Vector Analysis framework developed and applied throughout this work. IVA addresses the technical obstacles through its structured latent variable formulation: observed multimodal features $\mathbf{x}^{[k]}$ for modality k are modeled as linear mixtures of latent source components, $\mathbf{x}^{[k]} = \mathbf{A}^{[k]}\mathbf{s}^{[k]}$, where the key structural assumption posits that corresponding source components across modalities, collected into source component vectors $\mathbf{s}_n = [s_n^{[1]}, \ldots, s_n^{[K]}]^\top$, exhibit statistical dependence capturing cross-modal relationships, while different source component vectors remain mutually independent. This architecture naturally represents the intuition that multimodal content decomposes into multiple independent semantic factors, topic, sentiment, visual scene category, temporal context, source credibility, each exhibiting characteristic correlations across modalities yet remaining statistically independent of other factors.

The formulation directly resolves the identified challenges. Heterogeneous feature spaces unify through common latent source representations, with modality-specific mixing matrices $\mathbf{A}^{[k]}$ mapping shared semantic factors into appropriate channel-specific features, embedding spaces for text, visual descriptors for images, temporal signatures for metadata, network features for propagation patterns. This preserves modality-specific structure while enabling cross-modal interaction through shared latent factors. Missing modalities are handled gracefully through probabilistic marginalization: when modality k is unobserved, inference proceeds using only available modalities' contributions to source estimation, with uncertainty propagated appropriately through the probabilistic framework. Scalability follows from the independence assumption across source component vectors: rather than modeling full K-dimensional joint distributions requiring exponential parameters, IVA represents N separate source component vector distri-

butions, each K-dimensional, yielding linear rather than exponential parameter growth with modality count.

The framework provides natural mechanisms for addressing statistical challenges. Distributional flexibility accommodates non-stationary evolution and complex dependency structures through adaptive density estimation: rather than assuming fixed parametric forms (Gaussian, Laplacian) that fail to capture real-world complexity, the M-EMK extension employs semi-parametric maximum entropy estimation enabling data-driven adaptation to arbitrary source distributions including multimodal, heavy-tailed, or otherwise complex forms characteristic of evolving misinformation. Asymmetric modal reliability is implicitly encoded in learned mixing matrices $\mathbf{A}^{[k]}$, which assign differential weights to how each latent source manifests across modalities, sources exhibiting strong textual signal but weak visual signal naturally emerge when text carries primary discriminative information, with the converse holding when visual manipulation dominates. The structured sparsity extension (IVA-SPICE) provides explicit mechanisms for identifying which cross-modal interactions carry genuine discriminative signal versus noise, enabling principled dimensionality reduction and modality selection.

Critically, IVA maintains interpretability throughout this complex processing pipeline. The explicit source component decomposition yields human-understandable factors: each $\mathbf{s}_n$ corresponds to a distinct semantic aspect underlying multimodal observations, with dependency structure within $\mathbf{s}_n$ revealing how that factor manifests across channels. Violations of expected cross-modal correlations, for instance, a source component exhibiting strong presence in textual features yet anomalously absent or inconsistent in corresponding visual features, directly indicate cross-modal inconsistency characteristic of misinformation. This transparent decomposition enables analysts to understand which aspects of multimodal content, specific textual claims, visual elements, or their interaction, drive classification decisions, providing the explainability essential for deployment in socially consequential domains where opaque black-box decisions prove unacceptable.

The multimodal nature of contemporary misinformation thus constitutes not a peripheral complication amenable to simple feature engineering, but the fundamental defining characteristic shaping detection system requirements. Effective solutions demand methods explicitly designed for joint multimodal modeling, capturing rich, evolving, often adversarially crafted cross-modal relationships through principled statistical frameworks that balance representational flexibility with computational tractability, robustness with interpretability, and adaptive capacity

with theoretical grounding. The IVA-M-EMK framework, combining structured latent variable modeling with flexible semi-parametric density estimation, provides such a foundation for the misinformation detection systems developed and evaluated in subsequent sections.

4.1.5 Why Use IVA and Its Extensions?

The extensive challenges in multimodal misinformation detection, including diverse feature spaces, intricate cross-modal dependencies, non-stationary distributions, adversarial evolution, and the need for computational scalability, necessitate a fundamentally distinct approach from simplistic fusion architectures. Independent Vector Analysis offers a systematic solution that directly confronts these challenges through three fundamental capabilities: structured dependency modeling that encapsulates cross-modal relationships while ensuring computational feasibility, clear differentiation between shared semantic factors and modality-specific expressions, and a transparent decomposition that facilitates interpretable feature extraction. These capabilities render IVA particularly adept at addressing the multimodal misinformation detection challenge, especially when enhanced by flexible density estimation and structured sparsity extensions.

The primary benefit of IVA is its ability to model statistical dependencies across modalities without falling victim to the curse of dimensionality that affects unrestricted joint distribution estimation. In contrast to conventional Independent Component Analysis, which presumes total independence among all sources and is thus insufficient for multimodal contexts, IVA explicitly acknowledges that corresponding components across various modalities frequently display significant correlations. The semantic concept of "disaster severity" simultaneously manifests through particular linguistic patterns in text and unique visual characteristics in images, with these manifestations being inherently correlated, reflecting their common semantic origin. IVA encapsulates this via its source component vector architecture, which organizes related sources across modalities while preserving independence among distinct semantic factors. The structured dependence assumption significantly decreases the parameter count necessary for joint modeling: instead of estimating a comprehensive multivariate distribution across all features from various modalities, which is exponentially complex, IVA estimates several lower-dimensional distributions linked to specific semantic factors, resulting in linear scaling with the number of modalities instead of exponential scaling.

This systematic method is especially effective for detecting misinformation, as

deception often arises from breaches of anticipated cross-modal correlations rather than from irregularities within singular modalities. Examine a meme that juxtaposes genuine disaster imagery with fictitious casualty figures: the photograph shows no signs of forensic alteration, the text lacks clear linguistic indicators of falsehood, yet the stated numbers are at odds with the actual event represented. IVA inherently identifies such discrepancies as its learned model encapsulates standard correlation patterns between textual assertions and visual evidence in authentic content. When new content demonstrates an unusually weak correlation, with textual disaster severity indicators not aligning with visual severity cues, this statistical inconsistency suggests possible misinformation. Unimodal approaches completely overlook this aspect, whereas simplistic fusion methods that regard modalities as independent do not capture these essential correlation patterns.

In addition to its dependency structure, IVA offers crucial flexibility in the mapping of semantic factors to modality-specific features via its mixing matrix formulation. Diverse information channels utilize fundamentally distinct representational frameworks: text employs discrete tokens within semantic embedding spaces, images consist of continuous pixel arrays or acquired visual descriptors, temporal metadata is represented as time series, and network propagation is encoded in graph structures. IVA addresses this heterogeneity by acquiring modality-specific transformations that convert shared latent semantic sources into suitable channel-specific representations. The mixing matrices serve as learned translators, transforming abstract concepts such as "emotional intensity" or "source credibility" into tangible textual features (specific word selections, syntactic structures) for text, distinct visual attributes (color distributions, compositional elements) for images, and characteristic signatures (posting frequency, propagation speed) for temporal metadata. This distinction between shared semantic content and modality-specific expression facilitates significant cross-modal comparison, despite fundamentally incompatible feature spaces, thereby preventing the information loss associated with simplistic feature concatenation.

Moreover, IVA inherently offers interpretability via its explicit source decomposition, an essential criterion for implementation in socially significant areas where opaque black-box decisions are deemed unacceptable. Each identified source component corresponds to a unique semantic factor underlying the observed content, with its expression across modalities demonstrating how that factor manifests in various information channels. Analysts can investigate which sources demonstrate significant textual features, which prevail in visual characteristics,

and importantly, which reveal anomalous cross-modal inconsistency patterns suggestive of deception. This transparency facilitates comprehension not only of the classification of content as misleading but also of the specific factors, such as contradictory textual assertions versus visual evidence, dubious temporal dissemination patterns, and inconsistent indicators of source credibility, that influenced the classification. This explainability is crucial for human supervision, error assessment, and sustaining user confidence in automated content moderation systems.

Standard IVA formulations, however, impose stringent parametric assumptions on the distributions of source components, typically adopting Gaussian, Laplacian, or Generalized Gaussian forms for mathematical convenience or computational feasibility. These stringent specifications impose systematic constraints when addressing real-world data with intricate statistical characteristics. Multimodal misinformation content exhibits significant diversity: political memes, health misinformation, fabricated disaster reports, and celebrity death hoaxes each demonstrate unique distributional traits, frequently characterized by multimodal densities that reflect varied content categories, heavy tails that capture sporadic extreme values in user-generated data, or skewed distributions indicative of imbalanced class prevalence. Inflexible parametric assumptions inevitably inadequately capture this diversity, necessitating unfavorable trade-offs between accommodating various content types and introducing systematic bias that undermines the quality of source separation.

The IVA-M-EMK extension resolves this fundamental limitation by incorporating flexible semi-parametric density estimation grounded in maximum entropy principles. This approach generates data-driven estimates that automatically adjust to the actual characteristics of the underlying distribution, rather than limiting source distributions to fixed parametric families. The maximum entropy formulation guarantees that estimated densities are optimally unbiased, introducing no extraneous structure beyond what the observed data requires, while adhering to both global constraints that reflect overarching statistical properties and local constraints based on kernel functions distributed across the feature space. This adaptability is essential for effective modeling of varied, noisy, adversarially-produced misinformation, where inflexible parametric assumptions frequently falter. The method accommodates any distributional complexity while ensuring computational efficiency through Quasi-Monte Carlo integration techniques, which achieve significantly better convergence rates than basic sampling methods.

4.2 Multimodal Representation and Fusion Architecture

4.2.1 Dataset Description

The following experiments were systematically developed utilizing the MediaEval2016 Image Verification Corpus[1]. These well-curated datasets comprise several essential elements:

1. Labeled training and test tweet text and multimedia datasets that are distinctly partitioned, specifically sourced from the 2016 timeframe, ensuring the data is reflective of a particular temporal context.
2. The training dataset is comprised of tweets centered around a variety of events that are notably different from those featured in the test dataset, thereby offering a diverse array of contexts for the model.
3. Each tweet in the dataset is carefully classified as either "fake" or "real," where "fake" designates multimedia content that does not accurately portray the actual event it claims to represent, thereby adding complexity to the dataset.
4. The categorization of "fake" content includes posts that feature media from past events misrepresented as current occurrences, instances of manipulated contexts, or outright false claims regarding the depicted event, thereby underlining the challenges presented by misinformation.

The main working datasets include records that have only one image linked to them, which makes sure that the dataset is consistent. The textual data was processed to include "clean" text, which means that it doesn't have emoji characters, stop words, URLs, Twitter handles, timestamps, certain punctuation, or words that are shorter than two characters. This made the textual data better. The text also went through normalization steps, which included changing all the words to lowercase, turning multiple spaces into one space, and lemmatizing the text to make sure it was all the same. This dataset only includes tweets that are in English or a language that is very similar to English. It uses the Langid Python package and the International Organization for Standardization (ISO) language codes. Also, the dataset doesn't include tweets that were tagged as retweets, which keeps the original content intact. Records that produced null values in prior feature extraction phases have been omitted from the dataset. For the image data, preprocessing meant making all the pictures the same size (224x224 pixels) and normalizing them so that the

[1] https://github.com/MKLab-ITI/image-verification-corpus

dataset was consistent.

The training dataset is notably composed of:

- 9,140 tweet records associated with a total of 352 distinct images, providing a solid foundation for analytical endeavors.
- This dataset encapsulates a total of 15 unique events, contributing to the diversity of contexts present in the study.
- Of these, 5,127 tweets are labeled as fake, illustrating the prevalence of misinformation within this sample.
- Conversely, 4,013 tweets are classified as real, offering a contrasting element critical for model training and evaluation.
- Five of the events captured in this dataset contain both real and fake tweets, facilitating a more nuanced understanding of media representation.
- In contrast, ten events are represented solely by fake tweets, drawing attention to specific instances of misinformation.

The test dataset comprises:

- 796 tweet records that correspond with 92 different images, thereby allowing effective assessment of model performance.
- This test dataset represents a variety of 23 unique events, further augmenting the diversity and complexity of the evaluation.
- Within the test set, 467 tweets are labeled as fake, contributing to the intricacies of classification tasks.
- Additionally, 329 tweets are labeled as real, facilitating a balanced comparison for classification algorithms.
- Seven events within this dataset contain both real and fake tweets, adding additional layers to the evaluation criteria.
- Furthermore, one event is characterized by solely real tweets, while fifteen events consist entirely of fake tweets, emphasizing the widespread nature of misinformation in various contexts.

We have also added features from earlier studies to the basic datasets. We added a 300-dimensional Word2Vec (Mikolov *et al.*, 2013) embedding vector to each tweet record as a text feature. This allowed us to represent the tweets by averaging the word embedding vectors of the words they contained. We made this feature using a Word2Vec model that was

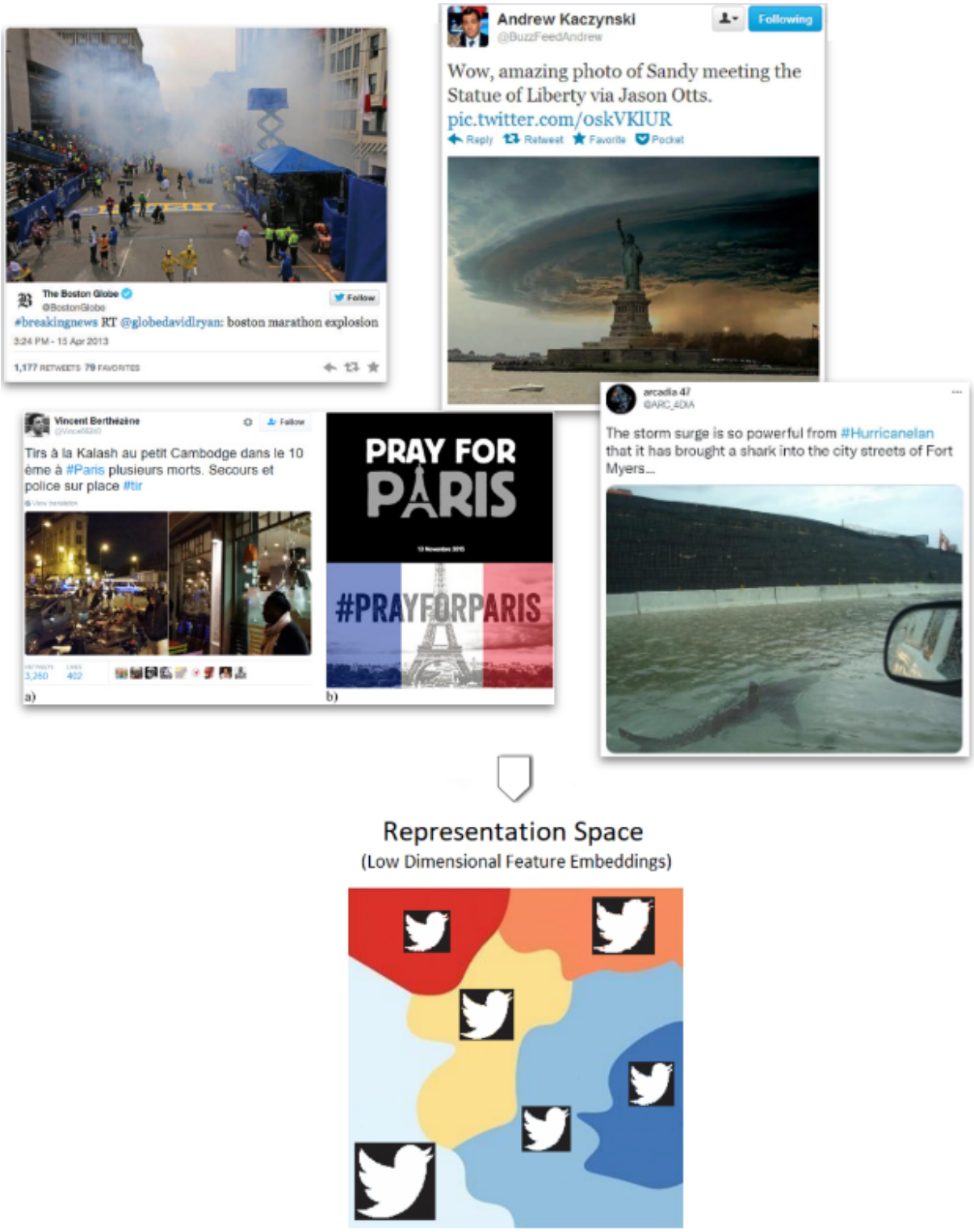
trained on the large Google News corpus[2]. For the image data, we incorporated a high-level image feature: a 4,096-dimensional fully connected layer derived from a pre-trained VGG-16 model for each image associated with an individual tweet record. These specific features had previously demonstrated the most promising classification results in earlier assessments. It is significant that we evaluated features produced by Bidirectional Encoder Representations from Transformers, referred to as BERT (Devlin *et al.*, 2019). Nonetheless, the classification accuracy derived from this specific dataset was not as reliable as that obtained through the utilization of Word2Vec-based embedding vectors. For example, combining BERT-based embedding vectors with VGG-based image features gave an F1-score of 70.68%. On the other hand, the F1-score goes up a lot to 77.45% when Word2Vec embeddings are combined with VGG features, as shown in **??**(b). It is important to keep in mind that BERT comes in many versions, and looking at all of the different versions of the F1-score would be too much for our current research. Still, this path is an interesting and promising one for future research projects.

4.2.2 High Level Feature Extraction

In order to conduct a comprehensive evaluation of the effects of incorporating new features and modalities on the classification accuracy and overall performance of our algorithms, we developed two innovative features that will be incorporated into our analytical framework. The initial feature that has been developed is referred to as "Image2text." This feature is produced by a sophisticated Image Captioning model, which is implemented in the PyTorch model[3]. This model employs pre-trained ResNet101 features to generate descriptive captions for each image in our datasets. Subsequently, we employ the same Word2Vec model that was previously employed to generate a 300-dimensional Word2Vec embedding. The word embedding vectors associated with each individual word in the generated captions are averaged to produce this embedding. This method introduces a text-based feature that functions as a unique modality for our evaluations, thereby enriching the dataset with valuable semantic information that is derived from the visual content. The second feature we developed for our analysis is a 200-dimensional vector that contains the top 200 topics assigned to each tweet. This vector is generated using an effective topic modeling technique. To initiate the topic modeling process, we employ the term frequency-inverse document frequency (TF-IDF) vectorizer of the scikit-learn Python package to generate

[2] https://code.google.com/archive/p/word2vec/
[3] https://github.com/ruotianluo/ImageCaptioning.pytorch

Figure 11 – Real-world instances of multimodal misinformation during crises, demonstrating the difficulty of detection amid various deceptive tactics. The figure shows real social media posts from big events like the Boston Marathon bombing (2013) and Hurricane Sandy (2012), where real pictures were used with false stories or were taken from events that had nothing to do with them. The post in the top right uses a picture of a different storm to falsely say that Hurricane Sandy affected the Statue of Liberty. The posts about the Paris attacks in 2015 show how real solidarity campaigns (#PrayForParis) can spread false information through edited images and wrong attributions. The bottom panel shows the representation space that our multimodal framework learned. It shows how different posts group together in low-dimensional feature embeddings based on their statistical properties.



Source: Prepared by the author.

a TF-IDF matrix that encompasses the textual content of the tweets. The vocabulary employed for this purpose has been meticulously selected to encompass only words that are present in less than 95% and more than 1% of tweets across our datasets. Next, we employ non-negative matrix factorization (NMF), a potent tool for identifying latent topics, to assign 200 distinct topics to each tweet. The deliberate selection of 200 topics guarantees that the dimensionality of this feature is consistent with the other matrices we will employ in our analysis, thereby facilitating a more comprehensive and coherent analysis of the data. It is anticipated that the discriminative power of our models will be significantly enhanced by these new features, which will lead to more precise classification results.

4.2.3 Multimodal Classification Framework

The task of multimodal misinformation detection is formulated as a blind source separation (BSS) problem, where each data modality, text and image, represents a mixed observation of shared latent semantic structures. This formulation naturally aligns with the Independent Vector Analysis (IVA) framework, which models statistical dependencies across modalities while maintaining independence across individual samples.

Consider a dataset comprising training and testing samples, where each instance consists of textual and visual content. We denote the observation matrices for each modality k as $\mathbf{X}_{\text{train}}^{[k]} \in \mathbb{R}^{d_k \times V_{\text{train}}}$ and $\mathbf{X}_{\text{test}}^{[k]} \in \mathbb{R}^{d_k \times V_{\text{test}}}$, where d_k represents the dimensionality of the initial high-level features, and V_{train} and V_{test} indicate the number of training and testing samples, respectively.

The processing pipeline consists of four integrated stages:

Stage 1: Dimensionality Reduction and Data Preparation. In order to extract N principal components, each modality is subjected to mean centering and Principal Component Analysis (PCA). Representations with reduced dimensions are produced by this transformation. $\hat{\mathbf{X}}_{\text{train}}^{[k]} \in \mathbb{R}^{N \times V_{\text{train}}}$ for each modality $k \in \{1,2,\ldots,K\}$, which are subsequently arranged into a three-dimensional tensor $\hat{\mathbf{X}}_{\text{train}} \in \mathbb{R}^{N \times V_{\text{train}} \times K}$.

Stage 2: Source Separation via IVA. The tensor representation serves as input to the IVA algorithm, which estimates K demixing matrices $\mathbf{W}^{[k]} \in \mathbb{R}^{N \times N}$. These matrices transform the observed features into source component vectors (SCVs) according to:

$$\mathbf{Y}_{\text{train}}^{[k]} = \mathbf{W}^{[k]} \left(\hat{\mathbf{X}}_{\text{train}}^{[k]} \right)^{\top}, \quad k \in \{1,2,\ldots,K\}. \tag{4.1}$$

The underlying assumption is that observations arise from mixing latent sources $\mathbf{s} \in \mathbb{R}^D$ through unknown mixing matrices $\mathbf{A}^{[k]}$:

$$\mathbf{x}^{[k]} = \mathbf{A}^{[k]}\mathbf{s}. \tag{4.2}$$

The recovered SCVs $\mathbf{y}_n = [y_n^{[1]}, y_n^{[2]}, \ldots, y_n^{[K]}]^\top$ capture cross-modal dependencies while maintaining statistical independence across samples. This property is particularly valuable for detecting semantic inconsistencies characteristic of misinformation.

Stage 3: Adaptive Density Estimation. In order to account for the intricate, non-Gaussian distributions that are prevalent in real-world data, we implement the Maximum Entropy with Measuring Functions and Kernels (M-EMK) estimator within the IVA framework. The source distribution is adaptively estimated by this method, which employs:

$$\hat{p}(\mathbf{y}_n) = \exp\left(-1 + \sum_{i=0}^{M} \lambda_i r_i(\mathbf{y}_n)\right), \tag{4.3}$$

where $r_i(\mathbf{y}_n)$ are measuring functions and λ_i are Lagrange multipliers that are optimized iteratively. This data-driven estimation allows the model to capture the skewed, multimodal, or heavy-tailed distributions that are inherent in misinformation data.

Stage 4: Feature Fusion and Classification. The extracted SCVs undergo fusion through concatenation, averaging, or max-pooling operations to form the final feature representation $\mathbf{Y}_{\text{train}}$. For the test set, the same preprocessing pipeline is applied using the training statistics: the training mean is subtracted, the PCA transformation from training is applied, and the learned demixing matrices transform the test data:

$$\mathbf{Y}_{\text{test}}^{[k]} = \mathbf{W}^{[k]} \left(\hat{\mathbf{X}}_{\text{test}}^{[k]}\right)^\top. \tag{4.4}$$

The Support Vector Machine (SVM) classifier, which has been chosen for its reliable performance on datasets with limited samples and its theoretical foundations, is fed by the fused representations $\mathbf{Y}_{\text{train}}$ and $\mathbf{Y}_{\text{test}}$. In order to guarantee statistical convergence, model selection and hyperparameter optimization employ grid search and five-fold cross-validation, which are conducted five times with data shuffling.

Independent Vector Analysis (IVA) and supervised classification are employed to capitalize on the complementary strengths of unsupervised feature extraction in this integrated framework.

4.3 Experimental Setup, Evaluation Protocol and Performance Results

The theoretical framework and multimodal classification architecture outlined in the previous sections provide the essential foundation for the practical identification of misinformation. Nonetheless, the proposed IVA-M-EMK approach requires a systematic assessment through carefully designed experiments that measure performance across various dimensions. The dimensions encompass classification accuracy, assessed via suitable metrics for imbalanced data; computational efficiency, measured through execution time analysis; scalability characteristics, which fluctuate according to feature dimensionality and training sample size; and the comparative performance of the proposed method against established baseline techniques. This section presents a thorough experimental validation demonstrating that the principled multimodal integration facilitated by IVA-M-EMK offers substantial practical benefits compared to both simplistic fusion techniques and other structured multimodal methods. The method attains enhanced classification efficacy while preserving computational feasibility appropriate for operational implementation.

4.3.1 Experimental Setup and Evaluation Protocol

The experimental validation employs a stringent evaluation protocol designed to mitigate potential biases or overfitting, ensuring reliable and reproducible performance assessments. The rigorous distinction between training and testing partitions is upheld at every stage of the pipeline to guarantee that performance metrics accurately represent true generalization ability rather than mere memorization of training instances. All experiments employ the datasets derived from MediaEval2016, as detailed in Section 4.2.1. The performance envelope of the proposed approach is thoroughly characterized, and operational regimes where particular design choices are most beneficial are identified through systematic variations in experimental conditions, including feature dimensionality, training sample size, fusion strategies, and comparison baselines, as assessed by the evaluation framework.

4.3.1.1 Performance Metrics and Evaluation Criteria

The macro-averaged F1-score serves as the principal metric for assessing classification performance. It was selected for its capacity to balance precision and recall, as well as its effectiveness in addressing the class imbalance in the dataset, where fake content constitutes

the minority class. The macro-averaging strategy, which entails calculating the F1-score for each class independently and averaging without weighting by class frequency, ensures that the detection performance of the minority fake class is afforded equal consideration to that of the majority real class. This mitigates the scenario where simplistic accuracy metrics may seem deceptively elevated by merely forecasting the predominant class for all instances. This metric selection aligns with operational deployment requirements, where false negatives (failure to identify authentic misinformation) and false positives (incorrect identification of legitimate content) entail distinct but comparable costs: False negatives facilitate the unregulated spread of detrimental misinformation, while false positives can lead to unjust censorship and erode user trust.

Alongside the assessment of classification accuracy, computational efficiency is quantified by measuring the total CPU execution time in seconds, encompassing both the training and testing phases. This temporal metric is essential for practical application: a detection system that attains a slightly higher F1-score but requires computational resources that are significantly greater may be inappropriate for real-time operational deployment, where processing latency directly impacts system responsiveness and infrastructure expenses. The assessment explicitly observes the correlation between execution time and critical problem parameters, including feature dimensionality and training sample size. This facilitates the identification of computational bottlenecks and the evaluation of whether the proposed method retains manageable complexity as the problem size escalates.

4.3.1.2 Baseline Comparison Methods

Experiments assess performance against various baseline methods that embody different paradigms for multimodal fusion to demonstrate the benefits of the IVA-M-EMK framework through thorough comparison:

Unimodal Baselines: The distinct information derived from separate modalities is measured by independent classifiers trained solely on textual or visual features, thereby setting lower limits on attainable performance. These baselines ascertain whether multimodal integration provides authentic value beyond the selection of the most robust individual modality, or if a straightforward single-channel analysis suffices.

Naive Concatenation Fusion: The simplest multimodal fusion approach entails the direct concatenation of high-level textual and visual features into cohesive vectors, which

are subsequently input into conventional classifiers. This baseline does not employ structured modeling of cross-modal dependencies; instead, it depends on the supervised classifier to implicitly acquire relevant interaction patterns from the training data. The value contributed by IVA's explicit dependency modeling is measured by a performance comparison with this methodology.

Integration of Principal Component Analysis (PCA): In an alternative dimensionality reduction approach, the features of each modality undergo PCA independently, and the resultant lower-dimensional representations are subsequently amalgamated through simple averaging. This baseline provides a systematic approach to dimensionality reduction akin to the preprocessing phase of IVA. Nevertheless, it excludes IVA's explicit modeling of cross-modal dependencies and the maximum likelihood separation framework.

Text in bold Canonical Correlation Analysis (CCA): This classical multivariate method identifies linear projections of two feature sets that optimize correlation, thus offering a systematic approach for integrating heterogeneous modalities by aligning them in a shared latent space. Unlike IVA, CCA does not enforce explicit probabilistic models on latent components; instead, it optimizes the correlation structure. IVA's probabilistic modeling framework is differentiated from simpler correlation-based fusion when contrasted with CCA.

Alternative IVA Variants: The experiments are juxtaposed with established IVA implementations utilizing fixed parametric source models to assess the distinct contribution of flexible semi-parametric density estimation through M-EMK. IVA-Gaussian (IVA-G) presumes multivariate Gaussian distributions and relies solely on second-order statistics; IVA-Laplacian (IVA-L) presumes Laplacian distributions and prioritizes higher-order statistics appropriate for sparse sources; IVA with adaptive Generalized Gaussian Distribution (IVA-A-GGD) provides intermediate flexibility via data-driven estimation of shape parameters within the constrained GGD family. This analysis directly evaluates whether M-EMK's distributional flexibility yields substantial performance improvements compared to simpler parametric alternatives.

4.3.1.3 Experimental Design and Parameter Selection

The experimental design methodically alters essential parameters controlling the multimodal fusion pipeline to thoroughly evaluate performance across different operational conditions. The dimensionality of features, defined by the number of principal components retained post-PCA, ranges from minimal values that capture only the predominant modes of variation

to greater values that maintain more intricate feature details, facilitating the evaluation of the accuracy-complexity tradeoff associated with dimensionality selection. The training sample size is similarly diverse to assess sample efficiency: whether methods attain satisfactory performance with restricted labeled data (essential due to annotation expenses) and how performance improves as more training examples are introduced.

The IVA-M-EMK approach entails fusion strategy experiments that evaluate three distinct methods for amalgamating estimated source component vectors into final feature representations for the supervised classifier: concatenation, which retains all extracted source information but results in increased dimensionality; maximum pooling, which identifies the most significant activation within each source across modalities; and averaging, which offers a balanced solution by preserving moderate dimensionality while synthesizing cross-modal information. A systematic comparison determines the strategy that best balances information retention with complexity arising from dimensionality.

All experimental configurations utilize Support Vector Machine (SVM) classifiers with linear kernels for the final supervised classification phase, chosen for their robust theoretical underpinnings, interpretability via learned weight vectors, and proven efficacy on moderate-scale datasets with meticulously crafted features, attributes that are well-suited to the IVA-extracted representations. Hyperparameter optimization is conducted via five-fold cross-validation in conjunction with grid search across the SVM regularization parameter space, with the entire cross-validation process reiterated five times utilizing distinct random data shuffling to guarantee robust parameter selection that is not excessively reliant on specific train-validation splits. This stringent optimization protocol protects against both underfitting (inadequate model capacity) and overfitting (excessive model complexity resulting in poor generalization).

4.3.1.4 Computational Infrastructure and Implementation Details

All experiments were performed on standardized computational infrastructure to facilitate equitable timing comparisons: computations were carried out on systems with Intel Core i7 processors and 16GB RAM, utilizing MATLAB implementations of the IVA algorithms and Python's scikit-learn library for SVM classification and evaluation metric calculations. No GPU acceleration or distributed computing was utilized, guaranteeing that the reported execution times represent performance attainable on standard hardware appropriate for deployment in resource-limited operational settings. Implementation details adhere to established best practices:

features are normalized to zero mean and unit variance prior to PCA application to prevent bias in dimensionality reduction from arbitrary feature scaling, convergence criteria for iterative IVA algorithms are established at relative gradient norms below specified thresholds to balance accuracy and computational expense, and identical random seeds are utilized across experimental replications to guarantee reproducibility.

This research fundamentally questions whether authentic multimodal integration offers additional benefits beyond merely choosing the most dominant individual modality or employing simplistic feature concatenation. Table 1 delineates the classification performance and computational expenses for three baseline configurations: text-only classification, image-only classification, and the direct concatenation of high-level textual and visual features devoid of structured multimodal fusion.

Table 1 – Comparison of classification performance between unimodal and naive multimodal fusion methodologies. F1-scores evaluate detection precision, whereas CPU execution times, expressed in seconds, gauge computational efficiency.

Method	**F1-Score**	**CPU Time (seconds)**
Text Only	40.04%	2.7×10^3
Image Only	65.78%	1.03×10^4
Naive Concatenation	77.59%	1.7×10^4

Source: Prepared by the author.

The experimental results reveal a number of critical insights regarding the importance of multimodal integration and modality-specific discriminative ability. The text-only classification achieves a 40.04% F1-score, which is significantly lower than the random classification baseline. This indicates that linguistic features alone do not possess the requisite discriminative power for the effective detection of misinformation in this dataset. This subpar performance exemplifies the intricate nature of contemporary textual misinformation, which increasingly employs syntactically accurate and linguistically proficient prose that exhibits few detectable anomalies through conventional natural language processing techniques. Unimodal textual analysis is rendered inadequate due to the deliberate imitation of authentic discourse in style by misleading text.

In this dataset, image-based indicators, such as forensic manipulation artifacts, contextual discrepancies, or statistical signatures acquired by deep visual feature extractors, possess more potent discriminative signals, as evidenced by the significantly enhanced individual performance of visual features with a 65.78% F1-score. However, image-only performance

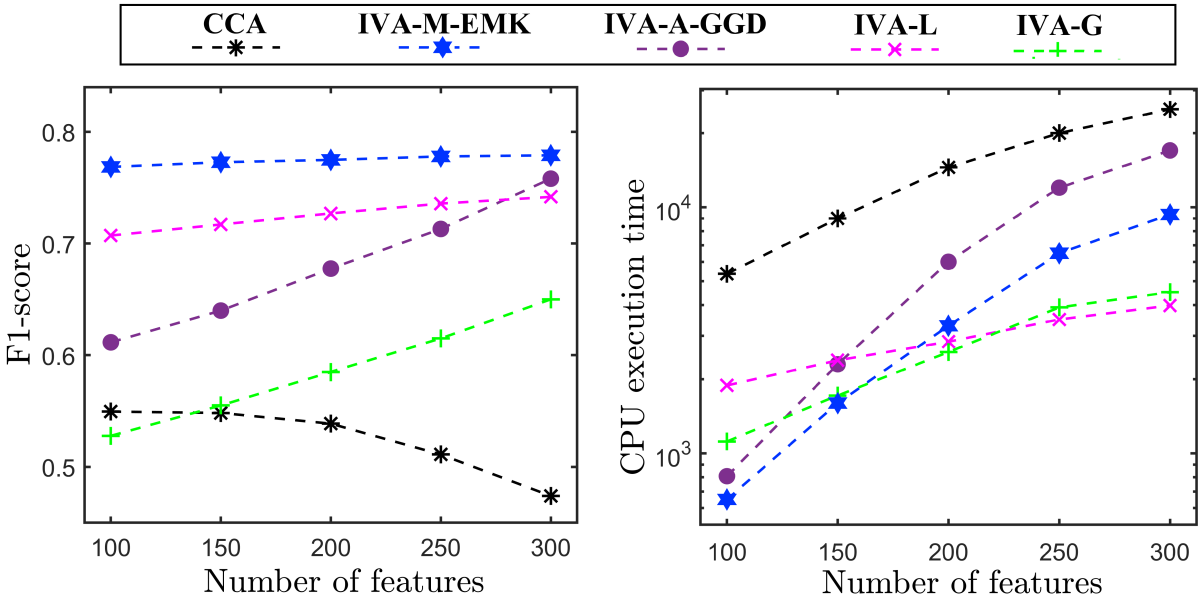
remains insufficient for operational deployment, as it is unable to identify over one-third of misinformation instances and generates false positive rates that would generate an intolerable number of erroneous flags in extensive content moderation contexts.

The naive concatenation baseline achieves a 77.59% F1-score, which confirms the fundamental assertion that textual and visual data provide complementary discriminative signals and demonstrates the substantial benefit of integrating modalities. The integration of textual features with visual representations significantly improves accuracy, demonstrating that the whole is greater than the sum of its parts, despite the fact that text-only performance slightly exceeds chance. This straightforward fusion method has substantial practical drawbacks: the combined feature vector dimensionality increases to 4,396 dimensions, resulting in high-dimensional representations that exacerbate the curse of dimensionality, increase the risk of overfitting, and significantly increase computational expenses to 1.7×10^4 seconds. Furthermore, naive concatenation does not have a mechanism for distinguishing between authentic discriminative signals and noise in cross-modal interactions, does not offer interpretability when it comes to the interaction of modalities in supporting classification decisions, and does not provide a systematic framework for addressing absent modalities or evaluating relative modality reliability.

In order to demonstrate practical applicability, any proposed technique must surpass a 77.59% F1-score. Additionally, computational expenses should be reduced below the 1.7×10^4 seconds baseline to facilitate real-time operational implementation. These baseline results serve as a framework for evaluating structured multimodal fusion methodologies. The subsequent subsections demonstrate that IVA-M-EMK satisfies both criteria by achieving improved classification accuracy through systematic cross-modal dependency.

The initial comprehensive evaluation looks at the relationship between computational efficiency and classification performance as the number of retained features increases. This is accomplished by increasing the number of principal components retained following PCA dimensionality reduction from 100 to 300. This analysis demonstrates each method's ability to effectively use higher-dimensional feature representations. It determines whether supplementary features provide true discriminative information that improves classification accuracy, or if they simply introduce noise and the risk of overfitting, compromising generalization performance. The evaluation of computational efficiency quantifies the relationship between execution time and feature dimensionality, identifying methods that maintain manageable complexity as problem size grows.

Figure 12 – Performance evaluation of multimodal fusion techniques based on the dimensionality of their features. The left panel shows the F1-score classification accuracy, emphasizing IVA-M-EMK's consistent and superior performance across all dimensions. In contrast, fixed-prior IVA variants (IVA-A-GGD, IVA-L) produce intermediate results, while CCA deteriorates significantly at higher dimensions. The CPU execution time is shown in the right panel, emphasizing the trade-offs in computational costs. IVA-G and IVA-L are simpler parametric models that achieve faster execution times at the expense of accuracy. In contrast, IVA-M-EMK maintains moderate computational costs while improving performance. All experiments use the entire training dataset, which is dimensionally adjusted by retaining PCA components.



Source: Prepared by the author.

Figure 12 shows the detailed results of the feature dimensionality experiment. The left panel shows F1-score trajectories, while the right panel shows the CPU execution times as the feature count increases from 100 to 300. IVA-M-EMK consistently achieves superior classification accuracy across the full dimensionality spectrum while maintaining competitive computational efficiency, with the performance landscape revealing distinct stratification among the tested methods.

4.3.1.5 Classification Accuracy Analysis

IVA-M-EMK shows great classification performance, with F1-scores ranging from 77% at lower feature dimensionality to a steady 78% as dimensionality rises. The method is very stable because it doesn't get worse because of the curse of dimensionality or get better because of overfitting to high-dimensional noise. Performance stays the same across the range of dimensions that were looked at. The M-EMK method quickly adapts to the real underlying

source distributions, no matter how many features there are. This avoids the systematic biases that fixed parametric models have when feature spaces grow and distributional complexity rises. This consistent behavior shows that the flexible semi-parametric density estimation framework works.

The fixed-prior IVA variants show a unique performance stratification that shows how flexible their distributions are. IVA-A-GGD gets an F1-score of about 61% at low dimensionality and 75% at high dimensionality, which is an average score. This shows a gradual improvement because extra features make it easier to fit the adaptive Generalized Gaussian family to the data. The unimodal, symmetric nature of the GGD remains a fundamental constraint, despite the fact that the shape parameter's adaptation provides substantial flexibility, which is significantly superior to entirely fixed parametric forms. IVA-L shows moderate performance of 72–73% across a range of dimensionalities. This is because it uses higher-order statistics that work well with sparse sources. However, it faces distributional mismatch when the real source structures differ from Laplacian assumptions.

IVA-G is the least effective of the IVA variants. It stabilizes at about a 65% F1-score in all dimensions and only slightly beats the image-only baseline. IVA-G only uses second-order statistics to model sources, which means it misses higher-order statistical information that sets different source types apart and captures the complex, non-Gaussian distributions that are common in real-world misinformation data. This important limitation clearly shows how wrong the Gaussian assumption is at its core. The lack of any performance improvement as dimensionality increases suggests that the distributional assumptions of IVA-G, rather than inadequate features, are the constraining factor in classification accuracy.

When the number of dimensions goes up, the CCA performance drops a lot. For example, it goes from about 55% at 100 features to less than 48% at 300 features, which is lower than random classification baselines. This is important. This significant failure exemplifies the intrinsic limitation of CCA: as dimensionality escalates, CCA becomes progressively vulnerable to spurious correlations and overfitting, owing to its singular emphasis on correlation maximization without a probabilistic framework. The method finds projections that maximize correlation in training data, even if those correlations are just noise or real patterns that help distinguish between different groups. This leads to a lot of generalization failure. This outcome corroborates the theoretical assertions of probabilistic modeling frameworks, such as IVA, which enforce structured assumptions that limit the scope of learned representations.

4.3.1.6 Computational Efficiency Analysis

The right panel of Figure 12 illustrates the computational efficiency characteristics, demonstrating the relationship between CPU execution time and feature dimensionality across various methods. IVA-M-EMK incurs moderate computational expenses, approximately 600 seconds at low dimensionality and about 10,000 seconds at 300 features, demonstrating near-linear scalability indicative of the algorithm's efficient optimization methods and QMC-based integration techniques. Although IVA-M-EMK does not attain the absolute minimum execution times, its computational expenses are fully manageable for operational implementation, especially when considered alongside its enhanced classification accuracy.

Among parametric IVA variants, IVA-G and IVA-L exhibit the lowest computational expenses, with IVA-G being especially efficient owing to its Gaussian assumption, which streamlines gradient calculations and facilitates second-order optimization techniques. This computational advantage offers negligible practical benefit due to the significantly lower classification performance of these methods; sacrificing 10-15 percentage points in F1-score for reduced computation time constitutes an unfavorable tradeoff in practical misinformation detection, where accuracy directly influences system efficacy.

IVA-A-GGD incurs the greatest computational expenses among the assessed methods, surpassing IVA-M-EMK at elevated dimensionalities, despite yielding lower classification accuracy. The computational load arises from the iterative shape parameter estimation processes of IVA-A-GGD, which necessitate multiple density evaluations and numerical optimizations during each algorithm iteration. The outcome indicates that adaptive parameterization within constrained distributional families enhances fixed parametric assumptions but entails significant computational costs without attaining the performance improvements achieved by IVA-M-EMK's entirely flexible semi-parametric method.

Canonical Correlation Analysis (CCA) demonstrates intricate computational scaling properties: although it is conceptually more straightforward than Independent Variable Analysis (IVA) methods, CCA's execution time increases significantly with dimensionality and can surpass that of several IVA variants at elevated feature counts. This paradoxical outcome illustrates the convergence challenges stemming from CCA's model misalignment; the algorithm necessitates numerous iterations to achieve convergence when optimizing correlation objectives that inadequately correspond with the underlying data structure, thereby negating any computational benefits per iteration gained from circumventing explicit density modeling.

The accuracy-efficiency analysis demonstrates IVA-M-EMK's practical benefits: the method attains superior classification performance across all assessed dimensionalities while preserving competitive computational costs, circumventing the accuracy compromises necessitated by simpler parametric models and the computational demands imposed by other adaptive methods. Moreover, IVA-M-EMK's consistent performance across various dimensionalities offers significant operational flexibility, allowing practitioners to choose feature dimensionality according to computational budget limitations without compromising classification accuracy, thus facilitating deployment in diverse resource availability contexts.

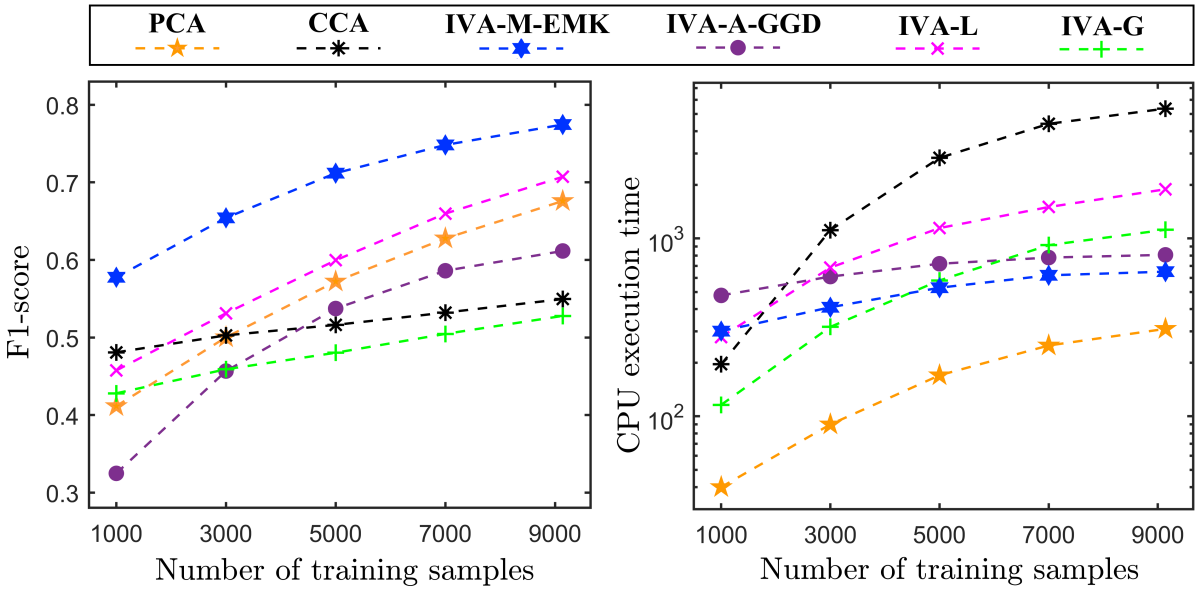
While the preceding analysis established IVA-M-EMK's superior performance given abundant training data across varying feature dimensionalities, practical deployment scenarios frequently confront limited labeled data availability. Manual annotation of multimodal misinformation is expensive, time-consuming, and requires expert judgment, creating strong incentives for methods achieving acceptable performance with minimal training data. This section examines sample complexity characteristics: how classification accuracy and computational costs scale as training dataset size varies from 1,000 to approximately 9,000 samples, revealing each method's data efficiency and assessing whether IVA-M-EMK's advantages persist in low-data regimes.

Figure 13 presents comprehensive sample complexity results, with the left panel displaying F1-score evolution as training set size increases and the right panel showing corresponding computational costs. Based on the previous analysis demonstrating IVA-M-EMK's stable performance across feature dimensionalities, all experiments in this section employ $N = 100$ principal components, a moderate dimensionality providing strong performance while maintaining computational efficiency. This experimental design isolates the effect of training sample availability from confounding influences of feature dimensionality variations.

4.3.1.7 Classification Performance Across Sample Regimes

IVA-M-EMK demonstrates exceptional sample efficiency, achieving strong classification performance even with limited training data and exhibiting consistent improvement as sample size increases. At the minimum evaluated training size of approximately 1,000 samples, IVA-M-EMK already achieves 58% F1-score, substantially exceeding most competing methods. Performance improves monotonically to approximately 78% at maximum training set size, demonstrating effective leverage of additional labeled data to refine density estimates and improve source separation quality. The method's strong low-sample performance validates the

Figure 13 – Comparison of performance among multimodal fusion techniques based on training sample size. The left panel displays F1-score trajectories, illustrating IVA-M-EMK's superior performance across all training set sizes, especially in low-data scenarios where flexible density estimation offers significant advantages. All IVA variants exhibit consistent enhancement with increased data, whereas CCA and PCA fusion display inadequate sample efficiency. Right panel: CPU execution time scaling characteristics, demonstrating IVA-M-EMK's advantageous computational profile with consistent, moderate costs across varying sample sizes. All experiments utilize $N = 100$ features, informed by prior analyses indicating consistent performance of IVA-M-EMK across various dimensionalities.



Source: Prepared by the author.

M-EMK framework's capacity to adapt to underlying distributions even from limited observations, avoiding the overfitting or convergence difficulties that plague more complex parametric alternatives.

IVA-L exhibits comparable sample efficiency characteristics, achieving approximately 53% F1-score at minimum training size and improving to roughly 71% at maximum sample availability. This favorable sample complexity reflects the Laplacian model's relatively simple parametric form, requiring estimation of fewer parameters compared to more flexible alternatives, enabling reliable density fitting even from limited data. However, IVA-L's ultimate performance ceiling remains substantially below IVA-M-EMK's, demonstrating that while simple parametric assumptions aid sample efficiency, they impose fundamental accuracy limitations that flexible semi-parametric methods overcome.

IVA-A-GGD and IVA-G demonstrate poorer sample complexity characteristics, requiring substantially larger training sets to achieve acceptable performance. IVA-A-GGD starts near 32% F1-score with minimal training data, gradually improving to approximately

62% at maximum sample size. This slow improvement reflects the challenge of simultaneously estimating shape parameters and performing source separation from limited observations, the adaptive mechanism provides little benefit when insufficient data exists to reliably estimate distributional characteristics. IVA-G exhibits even more severe sample efficiency limitations, achieving only 43% F1-score at minimum training size and reaching merely 53% even with maximum available data. The Gaussian assumption's fundamental distributional mismatch proves insurmountable regardless of training set size, validating that IVA-G's limitations stem from model inadequacy rather than mere parameter estimation challenges.

PCA fusion demonstrates moderate sample efficiency, achieving 41% F1-score initially and improving to 68% with maximum training data. While providing superior performance to IVA-G, PCA fusion's lack of explicit cross-modal dependency modeling limits ultimate classification accuracy compared to IVA methods that explicitly capture statistical relationships across modalities. CCA exhibits particularly poor sample complexity, starting at 48% and reaching only 55% despite maximum training data availability. The method's sensitivity to training set size likely reflects its tendency toward overfitting correlations that fail to generalize, a tendency exacerbated in low-data regimes where spurious correlations proliferate.

4.3.1.8 Computational Cost Scaling Analysis

The right panel of Figure 13 reveals how computational costs scale with training set size, providing critical insights for operational deployment planning. IVA-M-EMK exhibits remarkably stable computational costs, maintaining moderate execution times between approximately 300-700 seconds across the entire range of evaluated sample sizes. This favorable scaling characteristic reflects the algorithm's efficient optimization procedures and the QMC integration technique's superior convergence properties, additional training samples improve density estimation quality without proportionally increasing per-iteration computational costs. The stable cost profile provides valuable operational flexibility: practitioners can utilize maximum available training data to optimize classification accuracy without concern for prohibitive computational burdens.

CCA demonstrates the most favorable computational scaling among all evaluated methods, with execution times remaining below 1,000 seconds even at maximum training set size and exhibiting the flattest growth trajectory. However, This computational advantage offers negligible practical benefit due to the significantly lower classification performance of

these methods; sacrificing 10-15 percentage points in F1-score for reduced computation time constitutes an unfavorable tradeoff in practical misinformation detection, where accuracy directly influences system efficacy.

PCA fusion similarly maintains low computational costs below 400 seconds across all sample sizes, benefiting from the simplicity of independent PCA application followed by averaging. While computationally attractive, the approach sacrifices the performance advantages realized through explicit cross-modal dependency modeling, achieving inferior F1-scores compared to IVA methods despite lower execution times.

Among parametric IVA variants, IVA-G and IVA-L exhibit moderate computational costs with relatively flat scaling as training size increases, ranging from 200-1,000 seconds depending on method and sample size. IVA-A-GGD demonstrates the most severe computational burden, with costs ranging from approximately 500 seconds at minimum training size to nearly 1,000 seconds at maximum sample availability. The higher computational cost reflects IVA-A-GGD's iterative shape parameter estimation procedures that require repeated density evaluations at each optimization iteration. Despite this computational investment, IVA-A-GGD fails to achieve competitive classification accuracy relative to IVA-M-EMK, demonstrating that adaptive parameterization within restricted distributional families incurs computational costs without delivering corresponding performance benefits.

4.3.1.9 Synthesis: Optimal Method Selection for Practical Deployment

The definitive practical guidance for selecting methods in operational misinformation detection systems is provided by the extensive experimental validation of variations in feature dimensionality and training sample size. IVA-M-EMK is the most effective method in nearly all of the conditions that have been evaluated, as it ensures competitive computational efficiency and achieves the highest classification accuracy. The method demonstrates a number of unique strengths, such as exceptional performance stability across a wide range of feature dimensionalities, high sample efficiency that enables effective implementation with minimal labeled training data, smooth performance scaling with the addition of training samples, and manageable, predictable computational costs that facilitate capacity planning for operational deployment.

These findings corroborate the theoretical premises underpinning the IVA-M-EMK framework: Flexible semi-parametric density estimation through maximum entropy princi-

ples offers substantial practical benefits compared to both fixed parametric methods and basic correlation-based fusion techniques. The performance enhancements are significant rather than trivial; IVA-M-EMK consistently surpasses rival methods by 5-10 percentage points in F1-score across various experimental conditions, indicating the distinction between marginal utility and truly dependable misinformation detection capability. Moreover, these improvements in accuracy are attained without excessive computational expenses, confirming IVA-M-EMK's appropriateness for practical application in resource-limited environments.

4.4 Interpretability and Explainability of Multimodal Detection

The implementation of automated misinformation detection systems in critical areas such as content moderation, public health communication, and electoral integrity verification requires capabilities that surpass simple predictive accuracy. Stakeholders, such as platform operators, content creators, fact-checkers, and impacted users, necessitate a clear comprehension of the classification rationale: the specific textual assertions, visual components, or cross-modal discrepancies that influence individual detection determinations, the interaction of various information modalities in corroborating or disputing veracity evaluations, and the conditions under which the system demonstrates reliable or uncertain performance. The necessity for interpretability arises from several converging demands: fostering user trust in automated decisions affecting information access and freedom of expression, enabling human oversight and appeals processes for rectifying erroneously flagged content, facilitating error analysis to inform system enhancements, and supporting adversarial robustness analysis to anticipate potential evasion tactics.

This section provides a thorough interpretability analysis of the IVA-M-EMK detection framework by systematically applying Local Interpretable Model-agnostic Explanations (LIME), a prevalent method for generating comprehensible explanations of black-box classifier decisions. LIME functions by deriving local linear approximations to intricate decision boundaries, determining which input features most significantly affect predictions for particular instances. By analyzing various test cases, including accurately classified misinformation, effective generalization to new semantic variations, and systematic failure modes, we illustrate that the acquired multimodal representations encapsulate semantically significant cross-modal relationships rather than mere statistical anomalies. We confirm that the integration of visual information offers authentic discriminative value beyond textual analysis alone and pinpoint spe-

cific distributional traits, particularly images with extreme brightness distributions, that present systematic challenges necessitating future methodological enhancements.

4.4.1 Methodology: Local Interpretable Model-Agnostic Explanations

The primary explanatory framework for the interpretability analysis is LIME, which was chosen because of its model-agnostic design, allowing it to be applied to any classifier without requiring internal access to learned parameters or gradient calculations. LIME generates explanations in a methodical manner: for each test instance that requires elucidation, the approach generates a dataset of perturbed variants by randomly masking or modifying input characteristics. It then obtains classifier predictions for these modified instances, fits a locally weighted sparse linear model that approximates the classifier's behavior near the original instance, and identifies the features with the highest absolute weight in this local approximation, indicating their relative importance to the classification decision.

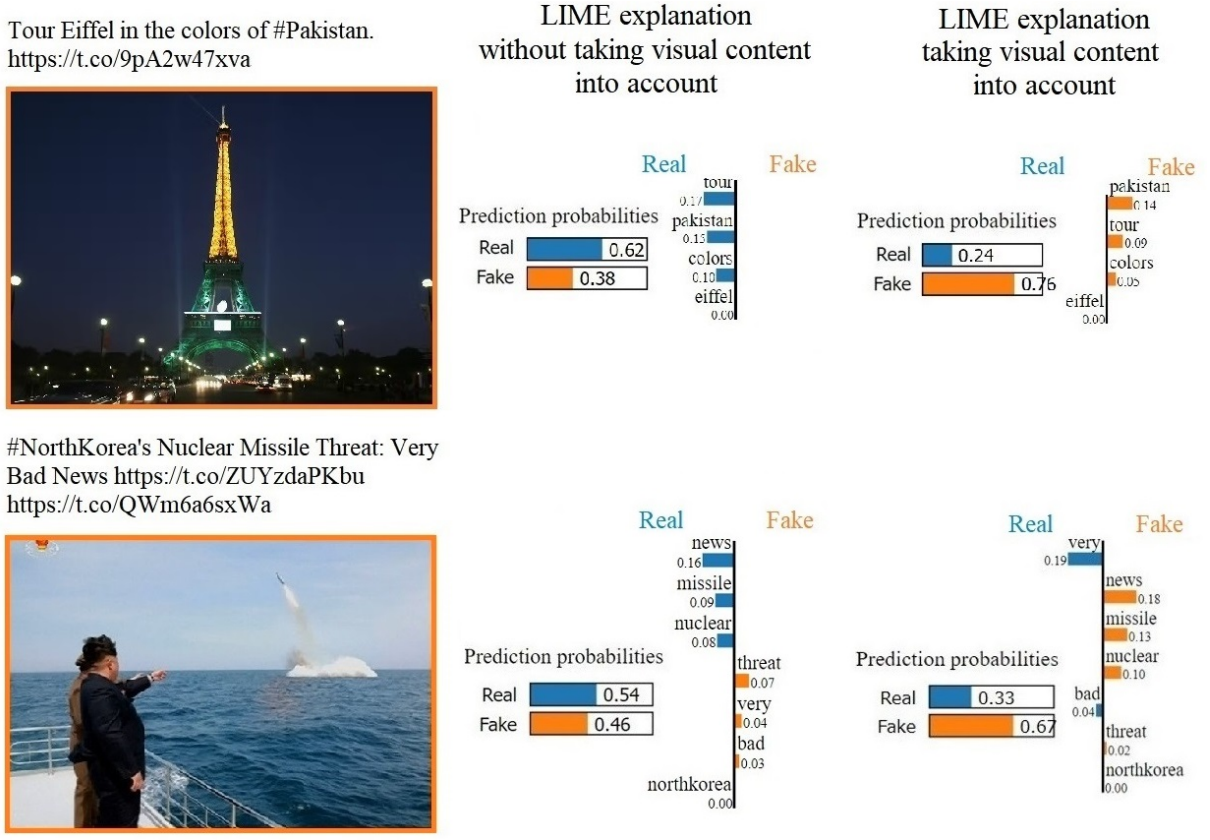
In the multimodal misinformation detection task, LIME's feature attribution determines the textual tokens and visual attributes that have the greatest impact on veracity predictions. The framework makes comparative analysis easier by isolating the contributions of specific modalities: by comparing explanations for classifiers trained with and without specific information channels, we quantify the extent to which visual content influences textual interpretation and determine whether cross-modal integration yields coherent semantic understanding or simply capitalizes on spurious correlations.

4.4.2 Scenario I: Cross-Modal Dependency Analysis Through Demixing Matrix Influence

The initial interpretability scenario investigates the impact of the demixing matrix $\mathbf{W}^{[1]}$, acquired through joint optimization of textual and visual modalities, on classification decisions when solely textual features are utilized as input during inference. This experimental design directly investigates whether the IVA framework effectively captures authentic cross-modal dependencies instead of merely concatenating independently processed modality features. . Specifically, we compare two classifier configurations: a baseline text-only system where $\mathbf{Y}^{[1]}_{\text{train}} = (\hat{\mathbf{X}}^{[1]}_{\text{train}})^{\top}$ employs raw PCA-reduced textual features without multimodal processing, versus a multimodally-informed system where $\mathbf{Y}^{[1]}_{\text{train}} = \mathbf{W}^{[1]}(\hat{\mathbf{X}}^{[1]}_{\text{train}})^{\top}$ applies the demixing matrix learned from joint text-image analysis despite receiving only textual input during classification.

Figure 14 illustrates two cases that exemplify the significant influence of multimodal

Figure 14 – Cross-modal influence analysis revealing how multimodally-learned transformations enhance text-only classification through implicit visual grounding. Left column: Original multimodal content comprising textual claims and accompanying imagery. Middle column: LIME explanations for baseline text-only classifier without multimodal demixing, showing feature attributions based purely on linguistic patterns. Right column: LIME explanations for multimodally-informed classifier employing $\mathbf{W}^{[1]}$ learned from joint text-image analysis, demonstrating altered feature importance reflecting implicit visual context. Upper row: Fake tweet claiming Eiffel Tower illumination in Pakistani flag colors following Lahore attacks, image actually depicts 2007 Rugby World Cup celebration. Lower row: Fake tweet alleging North Korean nuclear missile threat with manipulated missile imagery. Both examples demonstrate how multimodal learning fundamentally reshapes textual interpretation even when visual content is not directly provided during inference.



Source: Prepared by the author.

learning on the interpretation of textual features. The upper row analyzes a fraudulent tweet asserting solidarity with the victims of the Lahore bombing by illuminating the Eiffel Tower in the colors of the Pakistani flag. The image presented seems credible, depicting the Eiffel Tower illuminated in green and white; however, it represents contextual misinformation, as the photograph actually illustrates the 2007 Rugby World Cup celebration rather than a reaction to the 2016 Lahore attacks. This case illustrates complex multimodal deception, wherein both textual and visual elements seem independently credible, with the falsehood arising from their discordant temporal and contextual association.

The LIME explanations demonstrate significant disparities in feature attribution patterns between text-only and multimodal classifiers. The baseline text-only system (middle column) allocates the highest weights to geographically and thematically pertinent terms: "pakistan" (0.15 towards fake), "tour" (0.21 towards real), "eiffel" (0.10 towards fake), and "colors" (0.10 towards fake). This attribution pattern demonstrates plausible linguistic heuristics; references to Pakistan and colors near mentions of the Eiffel Tower may indicate displays of solidarity. However, it is inadequate for reliable classification, resulting in a weak prediction (0.62 probability real, 0.38 probability fake) that erroneously favors the acceptance of the false claim.

The multimodally-informed classifier (right column) exhibits markedly altered feature importance, despite receiving identical textual input. Terms directly linked to the visual assertion demonstrate markedly increased weights: "Pakistan" rises to 0.14 concerning falsehood, whereas "tour" and "colors" maintain substantial relevance. The overall prediction strongly supports accurate classification, with a 0.24 probability of being genuine and a 0.76 probability of being fraudulent. This transformation cannot be attributed to the classifier's direct access to visual features during inference, as the experimental design exclusively offers textual input. The updated interpretation reveals that $\mathbf{W}^{[1]}$, obtained through the joint optimization of text-image pairs, embodies implicit expectations concerning visual-textual coherence. Terms like "Pakistan" and "colors" acquire negative weights not merely due to their linguistic compositions, but because the demixing matrix has identified that statements about particular colored illumination often correlate with specific visual characteristics in genuine content, characteristics whose expected statistical signatures are implicitly indicated even without visual input.

The bottom row of Figure 14 depicts an additional scenario involving a simulated North Korean nuclear missile threat, alongside modified imagery of a missile ascending from

water. The text-only baseline classifier persists in underperforming, producing almost random predictions (0.54 probability of real, 0.46 probability of fake), despite terms like "missile" (0.09 bias toward fake) and "nuclear" (0.08 bias toward fake) displaying negligible signs of deception. The linguistic content is insufficient, as alarmist rhetoric concerning military threats is widespread in both reputable news articles and sensationalized misinformation, providing minimal discernment.

The multimodally-informed classifier exhibits markedly improved performance (0.33 probability real, 0.67 probability fake), effectively identifying the content as misleading through fundamentally altered feature interpretation. Terms like "missile" (0.13 falsehood association), "nuclear" (0.10 falsehood association), "threat" (0.02 falsehood association), and "news" (0.18 falsehood association) demonstrate increased correlations with the false category. This transition exemplifies the acquired correlations of the demixing matrix: during multimodal training, the system recognized that sensationalist military threat rhetoric combined with dramatic yet inferior imagery (as opposed to professional news photography) signifies a characteristic feature of misinformation. Although direct visual access is lacking during inference, the multimodal learning process has enhanced textual interpretation to recognize these patterns, effectively "envisioning" the visual characteristics that would typically accompany such textual claims and identifying their statistical variances.

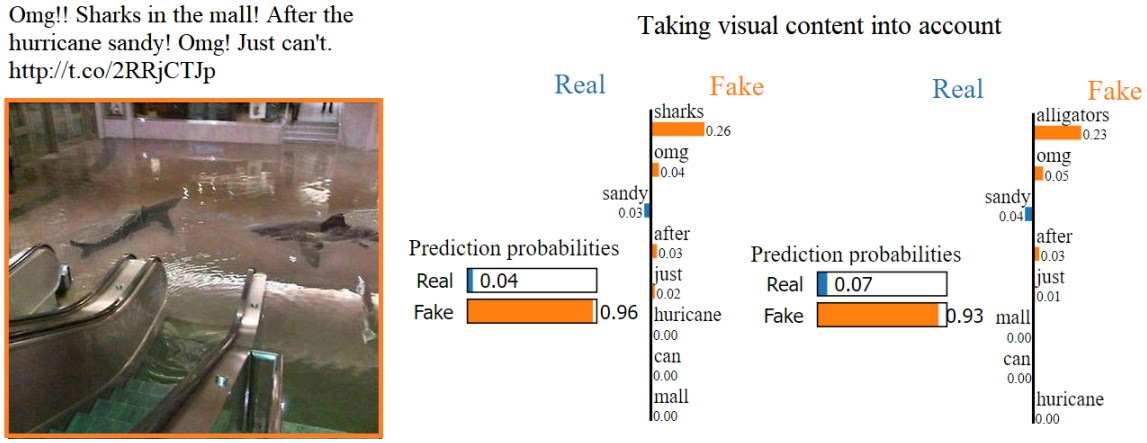
These examples demonstrate that IVA's structured dependency modeling through source component vectors effectively encapsulates genuine semantic cross-modal relationships instead of mere statistical correlations. The demixing matrices perform dimensionality reduction and feature decorrelation while also embodying expectations about the consistency of accurate content across modalities. The implicit cross-modal grounding remains intact when individual modalities are analyzed separately, fundamentally transforming the interpretation of textual features according to established visual-textual co-occurrence patterns. The framework successfully executes semantic completion through multimodal training, which cultivates intricate internal representations of the interactions between various information channels, thereby enabling more informed decision-making based on partial observations.

4.4.3 Scenario II: Semantic Robustness and Generalization to Novel Lexical Variations

The second interpretability scenario examines the framework's capacity for semantic generalization: whether the learned representations capture abstract conceptual relationships

enabling robust classification when confronted with lexical variations not observed during training. This capability proves critical for operational deployment where adversaries deliberately employ synonym substitution, euphemistic language, or coded terminology to evade detection systems that rely on brittle keyword matching. A detection framework exhibiting genuine semantic understanding rather than superficial pattern matching should maintain classification accuracy when semantically similar yet lexically distinct terms replace training vocabulary.

Figure 15 – Showing that semantic robustness works through a lexical substitution test. Left: A real fake tweet claiming that sharks were in a shopping mall after Hurricane Sandy, correctly labeled with LIME giving "sharks" a lot of weight (0.26 toward falsehood). The modified version replaces "sharks" with the unseen word "alligators." The classifier keeps a very close fake classification (0.93 versus 0.96 probability), and "alligators" show similar feature importance (0.23 towards fake), which shows that the pattern matching is based on meaning rather than words. The preserved classification accuracy despite vocabulary modifications indicates that the acquired representations embody abstract conceptual categories (aquatic predators in unlikely urban contexts) rather than merely retaining specific training terminology.



Source: Prepared by the author.

Figure 15 illustrates a controlled experiment that directly evaluates semantic robustness. The instance pertains to a widely disseminated false image claiming to depict sharks swimming in a submerged shopping mall following Hurricane Sandy, a misinformation that garnered significant viral propagation during the 2012 catastrophe. The initial tweet explicitly mentions "sharks," a term included in the training data due to the image's extensive dissemination. The left panel displays the classifier's evaluation of this original version: LIME assigns the greatest weight to "sharks" (0.26 for fake), with other pertinent terms such as "omg" (0.04), "sandy" (0.03), "after" (0.03), "hurricane" (0.00), and "mall" (0.00) receiving lesser weights. The classifier demonstrates high confidence in the classification of false content (0.04 probability of being real, 0.96 probability of being fake), accurately recognizing the deceptive material.

The essential test entails substituting "sharks" with "alligators," a semantically related but lexically different term that, crucially, was not encountered during training in connection with this specific image or analogous inundated indoor situations. If the classifier depends on fragile keyword matching, this substitution is likely to significantly impair performance: the specifically trained pattern "sharks + mall + flooding" would cease to match, potentially resulting in classification failure. A system demonstrating authentic semantic comprehension should acknowledge that alligators and sharks possess significant conceptual parallels; both are aquatic predators, both would be improbable in indoor shopping settings, both have been targets of analogous disaster-related misinformation campaigns, and both retain accurate classification.

The experimental results presented in the right panel of Figure 15 confirm authentic semantic robustness. Notwithstanding the novel terminology, the classifier exhibits nearly identical performance (0.07 probability of real, 0.93 probability of fake) with only slight deterioration from the original 0.96 probability of fake. LIME demonstrates that "alligators" receives similar feature attribution (0.23 towards fake) as the original "sharks" (0.26), suggesting the system acknowledges the semantic equivalence despite lexical variation. Supplementary terms exhibit comparable attributions: "omg" (0.05), "sandy" (0.04), "after" (0.03), and "mall" (0.00) obtain nearly equivalent weights, indicating that the overarching interpretative strategy remains consistent amid lexical perturbation.

This semantic robustness indicates essential characteristics of the acquired multimodal representations. The feature extraction pipeline utilizes pre-trained word embeddings that encapsulate distributional semantic relationships, whereby words occurring in analogous linguistic contexts obtain comparable vector representations. Terms such as "sharks" and "alligators" consequently reside in close proximity within embedding space, both linked to aquatic habitats, predatory conduct, and discussions pertaining to animals. The IVA framework, upon mastering the separation of sources and the identification of discriminative patterns, functions using continuous semantic embeddings instead of discrete keyword identities. The acquired demixing matrices and source component distributions therefore encapsulate abstract conceptual patterns, perilous aquatic creatures in improbable urban settings during emergencies, rather than retaining specific lexical instances.

Moreover, the multimodal characteristics of the representation enhance its robustness. Despite the significant semantic alteration caused by textual substitution, the visual content remains unaltered, continuing to portray the dubious inundated mall interior. The cross-modal

consistency verification intrinsic to IVA's dependency modeling guarantees that classification is based on coherent multimodal evidence rather than on isolated textual or visual attributes. An adversary seeking to avoid detection via synonym substitution must concurrently modify both textual and visual elements in semantically coherent manners, presenting a significantly more complex evasion challenge than merely altering text alone.

The semantic generalization ability differentiates principled multimodal frameworks such as IVA-M-EMK from more basic keyword-based or exclusively supervised methods that may overfit to the training vocabulary. The acquired representations encapsulate fundamental semantic frameworks, facilitating effective generalization to new linguistic expressions of analogous concepts, ensuring resilience against prevalent adversarial evasion strategies while preserving interpretability through semantically consistent feature attributions.

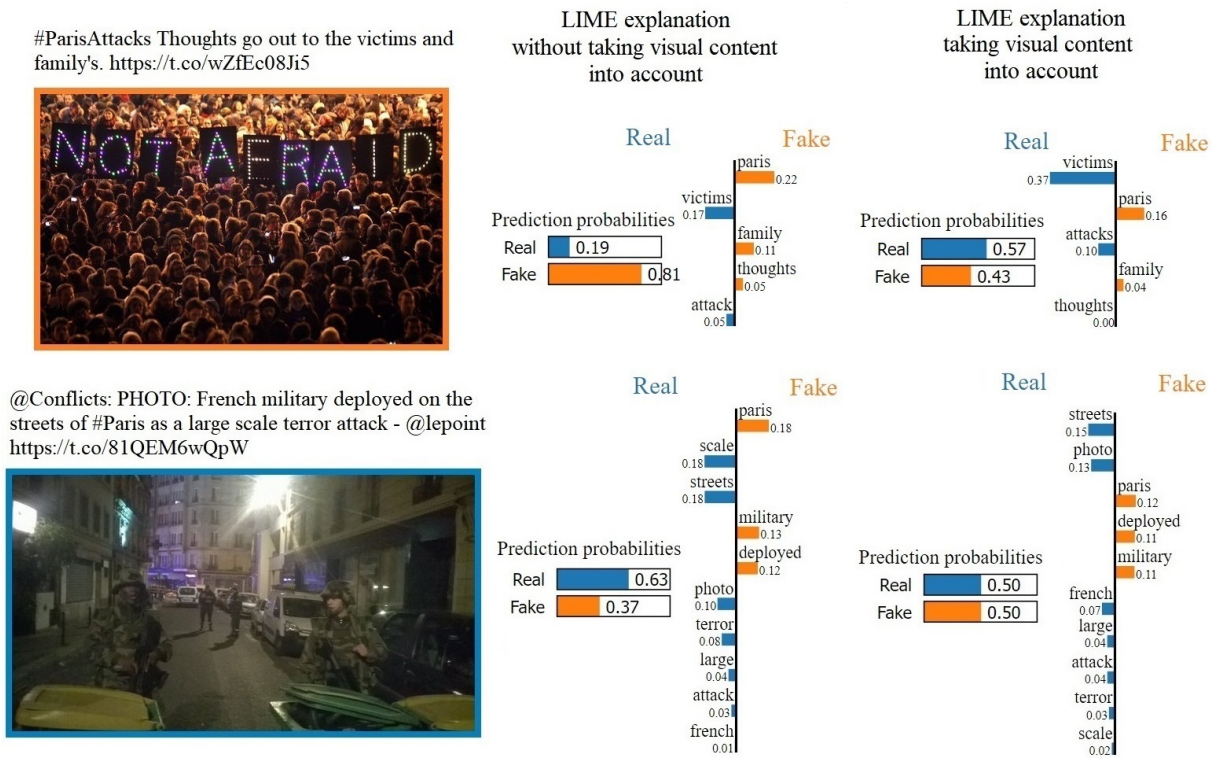
4.4.4 Scenario III: Systematic Failure Mode Analysis and Limitations

The prior analyses illustrate IVA-M-EMK's interpretable representation of authentic cross-modal semantic relationships; however, a thorough evaluation necessitates a candid assessment of systematic failure modes: particular distributional traits or content patterns where the existing framework shows diminished performance. Recognizing these failure modes fulfills several objectives: setting realistic performance expectations for operational implementation, informing future methodological improvements that rectify identified deficiencies, and illustrating the utility of interpretability tools for error analysis that uncovers failure mechanisms instead of simply indicating overall accuracy decline.

Figure 16 presents two representative failure cases where multimodal integration paradoxically degrades rather than improves classification performance relative to text-only analysis. These examples reveal a consistent pattern: images exhibiting extreme brightness distributions, predominantly dark pixels with concentrated bright regions, systematically introduce misleading visual features that confound the classifier despite carrying minimal genuine semantic information.

The upper row analyzes a fabricated tweet purporting to depict a solidarity vigil for the victims of the Paris attacks, featuring the text "Thoughts go out to the victims and family's" alongside an image of a crowd holding lights that form the phrase "NOT AFRAID." The content represents misinformation due to contextual manipulation: although the image illustrates a genuine memorial gathering, the specific photograph was captured at a separate

Figure 16 – Systematic failure mode analysis revealing visual feature distribution weaknesses. Left column: Original multimodal content. Middle column: LIME explanations for text-only classifier. Right column: LIME explanations for multimodally-informed classifier. Upper row: Fake tweet showing solidarity vigil for Paris attack victims, classifier incorrectly predicts real (0.57 probability) when incorporating visual features, versus correct fake prediction (0.81 probability) from text alone. The image contains extensive dark regions with concentrated bright elements ("NOT AFRAID" in lights) that introduce misleading visual features. Lower row: Real tweet depicting French military deployment after Paris attacks, multimodal classifier produces uncertain predictions (0.50 probability for both classes) versus correct real classification (0.63 probability) from text alone. Both cases demonstrate how images with extreme brightness distributions (predominantly dark pixels with concentrated highlights) systematically confound visual feature extractors, degrading rather than improving classification performance.



Source: Prepared by the author.

event and repurposed to accompany this tweet, thereby generating misleading associations. The text-only classifier performs adequately in this instance, yielding a 0.81 probability of being fake, with LIME assigning weights to emotionally charged terms such as "Paris" (0.22 towards fake), "victims" (0.17), and "family" (0.11 towards fake). The classifier identifies that generic emotional language, when combined with event references, often manifests in misinformation aimed at exploiting trending topics.

Nevertheless, the integration of visual features via $\mathbf{W}^{[1]}$ results in classification failure: the multimodal system assigns a probability of 0.57 to real (0.43 to fake), erroneously endorsing the deceptive content. LIME elucidates the source of ambiguity: visual integration fundamentally modifies the interpretation of textual features, with "victims" (0.37 toward real) now robustly endorsing the real class, in contrast to the text-only attribution. This reversal indicates the visual feature extractor's reaction to the image attributes; the photograph depicts a genuine memorial gathering, characterized by an appropriate emotional tone and visual components (crowd, candlelight, solidarity messaging). The profound visual attributes, developed for general image classification tasks, effectively identify the scene category but lack the contextual foundation to determine if this specific image is temporally and causally linked to the event mentioned in the text.

Critically, the image's extreme brightness distribution, predominantly dark background with concentrated bright regions (illuminated letters), creates an additional confound. Visual feature extractors, particularly deep convolutional networks, exhibit sensitivity to overall brightness and contrast distributions as these provide informative cues for scene recognition in standard computer vision tasks. However, in the misinformation detection context, low-light photography (common in nighttime events, indoor gatherings, or deliberate aesthetic choices) carries no inherent veracity signal. The system's learned associations between dark imagery and particular content categories introduce spurious correlations that degrade classification when these brightness characteristics appear in novel contexts.

The lower row presents a complementary failure case: a real tweet describing French military deployment after Paris attacks, accompanied by a photograph showing armed personnel on a street. The text-only classifier correctly identifies the content as real (0.63 probability) based on factual, descriptive language: "military" (0.13), "deployed" (0.12), "photo" (0.10), and "streets" (0.18 toward real) all contribute to appropriate classification. The textual content exhibits characteristics of legitimate news reporting, specific, verifiable claims with geographic

and institutional detail.

The multimodal classifier, however, produces maximally uncertain predictions (0.50 probability for both classes), effectively failing to classify the content. LIME shows that visual integration creates conflicting signals: terms like "streets" (0.15 toward fake) now incorrectly suggest misinformation, while "photo" (0.13 toward fake) and "deployed" (0.11 toward fake) similarly acquire inappropriate negative weights. This confusion again stems from extreme image brightness distribution, the photograph exhibits predominantly dark tones with selective bright regions (street lights, building illumination), similar to the previous failure case. The visual features extracted from this low-light nighttime photography apparently activate patterns the system associates with particular types of misinformation, overwhelming the correct signals from textual analysis.

These systematic failures reveal a clear limitation in the current visual feature extraction pipeline: deep convolutional networks trained on standard computer vision benchmarks (ImageNet, Places) develop representations optimized for object and scene recognition across well-lit, high-quality photography. These representations prove suboptimal for misinformation detection where: image quality varies dramatically from professional photojournalism to low-resolution mobile phone captures; lighting conditions span extreme ranges including nighttime events, indoor gatherings, and emergency situations; and semantic content relevance matters more than photographic aesthetics. The learned visual features consequently exhibit spurious sensitivity to brightness distributions, introducing systematic biases that degrade rather than improve classification for particular image types.

The thorough interpretability analysis reveals several essential conclusions about the IVA-M-EMK framework's ability to accurately represent authentic semantic cross-modal relationships in bimodal text-image contexts. The system effectively acquires significant cross-modal dependencies, demonstrated by the enduring impact of multimodally-learned demixing matrices on textual interpretation in the absence of visual input during inference, and by strong semantic generalization to novel lexical variations that indicate abstract conceptual comprehension rather than fragile keyword matching. The LIME-based explanatory framework demonstrates that learned representations encapsulate semantically coherent structures, identifying implausible combinations of textual assertions and visual contexts, generalizing conceptual categories beyond the training vocabulary, and displaying interpretable failure modes that aid in systematic error diagnosis and correction.

These failure cases underscore the essential significance of interpretability for operational implementation. Automated misinformation detection systems will inevitably demonstrate imperfect performance, as no framework achieves flawless classification across all content types. The difference between acceptable and unacceptable systems is not the eradication of all errors, but rather the assurance that failures manifest in predictable, diagnosable patterns that can be systematically addressed, and importantly, in offering transparent explanations that allow for human oversight to recognize and rectify erroneous decisions. The LIME-based analysis indicates that IVA-M-EMK meets the specified criteria: the framework's failures exhibit discernible distributional traits instead of obscure model anomalies, the acquired representations retain interpretability even in instances of failure, and the explanatory framework facilitates systematic error analysis that informs ongoing system enhancement.

The thorough interpretability analysis reveals several essential conclusions about the IVA-M-EMK framework's ability to accurately represent authentic semantic cross-modal relationships in bimodal text-image contexts. The system effectively acquires significant cross-modal dependencies, demonstrated by the enduring impact of multimodally-learned demixing matrices on textual interpretation in the absence of visual input during inference, and by strong semantic generalization to new lexical variations that indicate abstract conceptual comprehension rather than fragile keyword matching. The LIME-based explanatory framework demonstrates that learned representations encapsulate semantically coherent structures, identifying implausible combinations of textual claims and visual contexts, generalizing conceptual categories beyond the training vocabulary, and displaying interpretable failure modes that aid in systematic error diagnosis and remediation.

These capabilities affirm the essential principle that structured cross-modal dependency modeling via Independent Vector Analysis offers a robust basis for multimodal misinformation detection, attaining not only elevated overall classification accuracy but also transparent, semantically grounded decision-making that is accessible to human comprehension and subject to ongoing enhancement. Nevertheless, a pivotal inquiry persists: the analyses conducted to date exclusively focus on bimodal scenarios that incorporate both textual and visual information. Although these two modalities represent the most common combination in social media misinformation, extensive detection systems functioning at scale face significantly more complex multimodal environments.

Yet naively extending the current framework to incorporate these additional modali-

ties confronts a fundamental challenge: the challenge of dimensionality becomes dramatically more severe as modality count increases. The joint feature space dimensionality grows substantially with each added modality, not merely through additive feature accumulation but through exponential expansion of potential cross-modal interaction patterns requiring statistical modeling. A system integrating five modalities must capture not only individual channel characteristics but all pairwise, three-way, four-way, and five-way interaction patterns, creating representational complexity that rapidly overwhelms available training data and computational resources.

The subsequent analysis is guided by a critical methodological question that is motivated by this observation. Rather than merely determining whether multimodal integration enhances detection performance, it is necessary to systematically investigate which modalities and specific cross-modal interactions carry genuine discriminative signals versus introducing spurious correlations or computational burden without corresponding accuracy gains. Identifying the sparse subset of truly informative cross-modal relationships within the exponentially large space of potential dependencies is essential for effective high-dimensional multimodal fusion, which necessitates principled mechanisms for automatic modality selection and interaction pruning. This challenge is specifically addressed in the subsequent section, which expands the IVA framework by incorporating structured sparsity constraints that facilitate scalable multimodal fusion beyond bimodal scenarios. The IVA-SPICE variant utilizes graphical lasso regularization to enforce sparse precision matrix estimation, facilitating the automatic identification of modalities that demonstrate significant statistical dependencies while effectively nullifying weak or spurious correlations. This methodical framework for addressing multimodal dimensionality offers theoretical assurances of statistical efficiency and practical computational feasibility, facilitating a thorough exploration of the primary inquiry: Are two modalities adequate for effective misinformation detection, or do supplementary information channels offer significant enhancements that warrant their inclusion despite added complexity?

4.5 Are Two Modalities Sufficient? Addressing High-Dimensional Multimodal Complexity Through Structured Sparsity

The validation and interpretability analyses in previous sections demonstrate the IVA-M-EMK framework's ability to accurately capture authentic semantic cross-modal relationships in bimodal text-image contexts, attaining a 77.45% F1-score through rigorous joint statistical modeling. This robust performance prompts a critical architectural inquiry with significant

implications for operational implementation: are the two main modalities, textual content and visual imagery, adequate for capturing the essential discriminative signals necessary for misinformation detection, or does the inclusion of additional information sources yield meaningful accuracy enhancements that warrant the added computational complexity? This question cannot be resolved through intuition alone, as simplistic reasoning implying that "increased information enhances detection" encounters situations where indiscriminate inclusion of modalities frequently diminishes rather than enhances performance. The response is fundamentally reliant on the statistical modeling framework utilized, significantly impacting system design, computational resource distribution, and overall detection reliability.

Modern social media platforms offer extensive multimodal information that surpasses mere text-image combinations. Contemporary posts often incorporate temporal metadata that elucidates propagation dynamics, posting timestamps, resharing patterns, and virality trajectories, which differentiate organic information diffusion from orchestrated inauthentic amplification. Network features encapsulate indicators of source credibility and structural propagation characteristics: originating account authority, follower network topology, and cross-platform coordination patterns. User engagement metrics quantify audience response indicators: sentiment distributions of comments, sharing velocity, and demographic targeting trends. Automated image-to-text caption generation systems create synthetic textual descriptions of visual content, potentially exposing discrepancies between the actual depicted content and the accompanying human-generated text. Topic modeling uncovers underlying thematic structures within textual corpora, revealing coordinated narrative campaigns across multiple posts. Every supplementary information channel potentially offers complementary discriminative signals that, if effectively integrated, could improve detection accuracy beyond the capabilities of bimodal analysis.

A significant tension arises between theoretical potential and practical implementation. As the quantity of integrated modalities escalates from two to three, four, or beyond, the joint probability space expands exponentially rather than linearly. Each new modality introduces not only its unique feature dimensions but also, importantly, potential statistical interactions with all existing modalities. The shift from two to four modalities amplifies potential pairwise cross-modal interactions from one to six, representing a six-fold increase. Higher-order interactions increase at an accelerated rate: three-way interactions among modality triplets and four-way interactions encompassing all channels concurrently. This exponential complexity arises from various mechanisms that systematically impair statistical learning performance. The

curse of dimensionality results in training data becoming increasingly sparse in relation to the enlarged parameter space, as fixed sample sizes yield diminishing information per estimated parameter with increasing dimensionality. Spurious correlations abound in high-dimensional spaces where random chance creates seemingly statistical relationships that do not represent true discriminative patterns. The computational expenses increase significantly as density estimation, gradient calculation, and optimization processes scale superlinearly with joint dimensionality.

4.5.1 IVA-SPICE: A Principled Solution Through Structured Sparsity

Recognizing the sparsity inherent in multimodal misinformation patterns, we introduce IVA with Sparse Inverse Covariance Estimation (IVA-SPICE) (Rexhepi, 2022) as an evolution of the IVA framework specifically designed to handle high-dimensional multimodal spaces. Rather than attempting to model all possible dependencies, IVA-SPICE employs structured sparsity constraints to automatically identify and preserve only the most informative cross-modal relationships.

4.5.1.1 Theoretical Foundation

IVA-SPICE reformulates the density estimation problem through sparse precision matrix learning. Under the assumption that each SCV follows a multivariate Gaussian distribution, the score function becomes:

$$\phi_n^{[k]}(\mathbf{y}_n) = \mathbf{y}_n^\top \boldsymbol{\Sigma}_n^{-1} \mathbf{e}_k, \tag{4.5}$$

where $\mathbf{e}_k$ is the unit vector. The sparsity of each precision matrix $\boldsymbol{\Sigma}_n^{-1}$ becomes an informative characteristic, as it captures the conditional dependencies among the underlying features within an SCV by representing conditional independence as zero values in the inverse covariance.

To estimate these sparse precision matrices, IVA-SPICE employs graphical lasso, a regularized maximum likelihood estimation approach that uses an ℓ_1 penalty to encourage sparsity:

$$\mathscr{L}_{\text{SPICE}} = \log\det\boldsymbol{\Theta}_n - \text{Tr}(\mathbf{S}_n\boldsymbol{\Theta}_n) + \lambda\|\boldsymbol{\Theta}_n\|_1, \tag{4.6}$$

where $\Theta_n = \Sigma_n^{-1}$ is the precision matrix, $\mathbf{S}_n$ is the covariance, and λ controls the sparsity level. This formulation explicitly encourages conditional independence between weakly related source components, acting as an automatic feature selection mechanism that identifies which modal interactions contribute meaningfully to detection while suppressing noise and redundancy.

The complete IVA-SPICE procedure is summarized in Algorithm 1. The algorithm iterates over each source component, estimating its sparse precision matrix via graphical lasso, computing the orthogonal projection vectors, and updating the demixing matrices through gradient descent on the mutual information cost function until convergence.

Algorithm 1: IVA-SPICE algorithm (Rexhepi, 2022)

Input: $\mathbf{X} \in \mathbb{R}^{N \times V \times K}$
for *each* $k = 1, \ldots, K$ **do**
  Initialize $\mathbf{W}^{[k]} \in \mathbb{R}^{N \times N}$;
end
for $n = 1 : N$ **do**
  Estimate Σ_n^{-1} using graphical lasso to fully characterize $\phi_n^{[k]}(\mathbf{y}_n)$;
  Compute $\mathbf{h}_n^{[k]}$ for $k = 1, \ldots, K$;
  for $k = 1 : K$ **do**
    Calculate the derivative $\dfrac{\partial J_{\text{IVA}}}{\partial \mathbf{w}_n^{[k]}}$;
    $\left(\mathbf{w}_n^{[k]}\right)^{\text{new}} \leftarrow \left(\mathbf{w}_n^{[k]}\right)^{\text{old}} - \gamma \dfrac{\partial J_{\text{IVA}}}{\partial \mathbf{w}_n^{[k]}}$;
  end
end
$J_{\text{IVA}} = \sum_{n=1}^{N} H(\mathbf{y}_n) - \sum_{k=1}^{K} \log \left| \det \left(\mathbf{W}^{[k]} \right) \right|$;
Repeat steps 3 to 12 until convergence in $\mathscr{W}$;
return $\mathscr{W}$

4.5.2 Investigation of Modality Sufficiency

Image-to-text features use vision-language models that have already been trained to automatically write text descriptions of visual content. These descriptions include parts of scenes, objects, and actions that may be different from or better than human-generated captions. Topic modeling uses Latent Dirichlet Allocation on text in the corpus to find hidden thematic structures that reveal coordinated narrative campaigns or unusual topic combinations that point

to false information. These improved models make it easier to test all possible combinations of modes: six pairwise configurations test every possible two-modality fusion, four three-modality configurations test how adding an extra channel affects performance, and the full four-modality integration test whether comprehensive fusion achieves the highest level of accuracy.

Table 2 – Classification performance comparison across modality configurations revealing phase transition in optimal fusion strategy.

Modalities (K)	**Method**	**F1-Score**
$K = 2$: Text and Image	**IVA-M-EMK**	**77.45%**
	IVA-SPICE	67.92%
$K = 2$: Text and Topic Modeling	**IVA-M-EMK**	**57.25%**
	IVA-SPICE	53.45%
$K = 2$: Text and Image2Text	IVA-M-EMK	51.40%
	IVA-SPICE	**56.97%**
$K = 2$: Image and Topic Modeling	**IVA-M-EMK**	**72.59%**
	IVA-SPICE	63.25%
$K = 2$: Image and Image2Text	**IVA-M-EMK**	**73.45%**
	IVA-SPICE	66.37%
$K = 2$: Topic Modeling and Image2Text	**IVA-M-EMK**	**53.20%**
	IVA-SPICE	48.71%
$K = 3$: Text, Image, and Image2Text	IVA-M-EMK	61.73%
	IVA-SPICE	**69.43%**
$K = 3$: Text, Image, and Topic Modeling	IVA-M-EMK	58.43%
	IVA-SPICE	**67.99%**
$K = 3$: Text, Image2Text, and Topic Modeling	IVA-M-EMK	55.71%
	IVA-SPICE	**62.95%**
$K = 3$: Image, Image2Text, and Topic Modeling	IVA-M-EMK	62.07%
	IVA-SPICE	**68.58%**
$K = 4$: Text, Image, Image2Text, and Topic Modeling	IVA-M-EMK	61.22%
	IVA-SPICE	**73.01%**

Source: Prepared by the author.

Table 2 displays detailed results indicating a significant phase transition in optimal fusion methodology with an increase in modality count. The bimodal configurations exhibit IVA-M-EMK's distinct superiority in five out of six pairwise combinations, with the text-image pairing attaining the highest recorded performance of 77.45% F1-score. This dominance confirms the efficacy of the flexible semi-parametric density estimation method when joint dimensionality is manageable, as in the case of two modalities, where adequate training data allows for the reliable estimation of complex non-Gaussian distributions via the M-EMK framework, avoiding significant curse-of-dimensionality challenges. The singular bimodal configuration in which IVA-SPICE surpasses IVA-M-EMK, specifically in text and image-to-text fusion, is enlightening:

this particular combination demonstrates atypical traits wherein the two modalities offer predominantly redundant information (both conveying textual content, with image-to-text features being synthetically generated from images that could themselves be associated with original text). The intrinsic redundancy facilitates sparsity-based techniques to discern and leverage conditional independence frameworks, indicating IVA-SPICE's benefits in higher-dimensional contexts.

The three-modality configurations demonstrate significant performance reversals. IVA-M-EMK experiences significant deterioration compared to its optimal bimodal performance, decreasing from 77.45% to a range of 55.71-62.07% across the four assessed three-modality combinations, with performance declines surpassing 15 percentage points in certain instances. This degradation illustrates a fundamental limitation of the flexible density estimation framework: modeling intricate joint distributions across three modalities using limited training data leads to significant overfitting, as the method captures spurious correlations unique to the training set that do not generalize. In contrast, IVA-SPICE exhibits strong performance between 62.95% and 69.43%, surpassing several bimodal results of IVA-M-EMK despite the heightened dimensionality. The structured sparsity constraints offer implicit regularization, automatically discerning which cross-modal interactions possess authentic discriminative signals as opposed to noise, thereby executing principled feature selection at the modality combination level.

The results from the four modalities elucidate this phase transition with remarkable clarity. IVA-M-EMK's performance plummets to 61.22%, significantly lower than its poorest three-modality outcomes and markedly inferior to its 77.45% bimodal baseline, indicating a 16.23 percentage point decline despite utilizing double the information channels. The extensive density estimation method is entirely surpassed by the exponentially enlarged joint space, resulting in an overparameterized model that reflects training set peculiarities instead of authentic statistical patterns. IVA-SPICE attains a 73.01% F1-score, nearing the optimal bimodal performance while incorporating four modalities, and securing an 11.79 percentage point lead over IVA-M-EMK in the four-modality setup. This notable outcome illustrates that principled sparsity-based modeling facilitates efficient scaling to high-dimensional multimodal spaces, achieving a substantial portion of the bimodal baseline performance while potentially uncovering complementary discriminative signals from additional modalities that bimodal analysis inherently overlooks.

The systematic performance trends indicate essential statistical principles that regulate multimodal fusion. IVA-M-EMK's trajectory, characterized by robust bimodal performance followed by gradual deterioration, exemplifies typical overfitting dynamics in high-dimensional

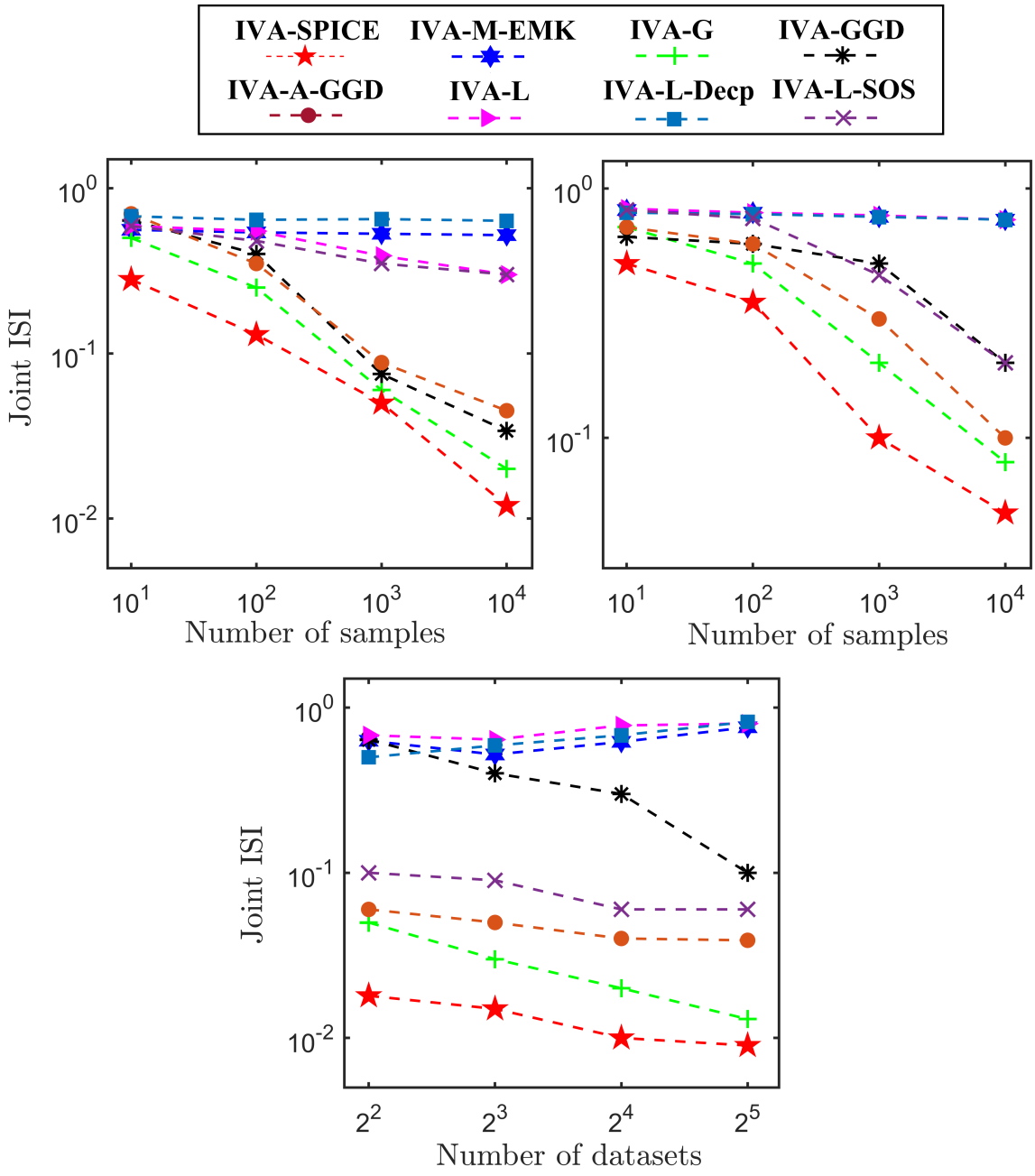
statistical learning. As the number of parameters increases more rapidly than the effective sample size with rising dimensionality, models tend to capture training noise instead of generalizable patterns. IVA-SPICE's divergent path and competitive bimodal performance, characterized by consistent enhancement with increasing modalities, confirm

Figure 17 illustrates detailed performance characteristics across three essential experimental dimensions, highlighting fundamental disparities in the handling of varying problem complexities by different IVA variants. The upper left panel analyzes sample complexity in the complex four-modality configuration, where Table 2 illustrated IVA-SPICE's significant advantages. At the minimum training set size, IVA-SPICE attains an F1-score of approximately 55%, surpassing the performance of IVA-M-EMK and illustrating effective parameter estimation despite limited data. With an increase in training samples, IVA-SPICE demonstrates a steady monotonic enhancement, achieving an F1-score of approximately 73% at maximum sample availability, closely nearing the 77.45% bimodal baseline despite a four-fold increase in dimensionality. The advantageous sample complexity illustrates the implicit regularization effect of sparsity constraints: by removing extraneous parameters linked to weak dependencies, the effective model complexity stays manageable in relation to the available training data, facilitating reliable parameter estimation across the sample size spectrum.

The parametric IVA variants exhibit significantly poorer sample complexity traits in this high-dimensional context. IVA-M-EMK demonstrates predominantly flat performance between 55-60% across the complete sample size spectrum, unable to utilize supplementary training data for enhanced generalization, a typical indication of model overparameterization where an expanded sample size yields negligible advantages as the model already accommodates training noise instead of generalizable patterns. IVA-A-GGD exhibits marginal enhancement with supplementary samples but levels off significantly beneath IVA-SPICE, whereas IVA-L and IVA-G reveal even poorer sample efficiency, thereby affirming that rigid parametric assumptions are insufficient for intricate high-dimensional distributions. The analysis of performance stratification confirms that the assumptions of the structural model, particularly sparsity, offer significantly greater advantages than simply increasing sample size when addressing the challenges posed by the curse of dimensionality.

The upper right panel displays divergent outcomes for the bimodal text-image configuration, wherein IVA-M-EMK attained its optimal performance. IVA-M-EMK exhibits exceptional sample efficiency, achieving a 77% F1-score with a moderate amount of training

Figure 17 – Comparative performance evaluation of IVA variants under diverse experimental conditions. Upper left panel: F1-score variation with training sample size for four-modality fusion ($K = 4$, $N = 100$ features), illustrating IVA-SPICE's enhanced sample efficiency and optimal performance threshold. Upper right panel: F1-score variation with training samples for two-modality fusion ($K = 2$), illustrating IVA-M-EMK's superiority in low-dimensional contexts. Lower panel: F1-score progression as the number of integrated modalities increases from $K = 2^1$ to $K = 2^2$, demonstrating IVA-SPICE's resilience to dimensionality expansion in contrast to the significant deterioration of unregularized methods. All experiments utilize uniform feature extraction pipelines and classifier configurations, thereby isolating the impact of the multimodal fusion methodology.



Source: Prepared by the author.

data and maintaining stable performance subsequently. IVA-SPICE demonstrates competitive performance, albeit marginally inferior to IVA-M-EMK across the majority of the sample size spectrum, corroborating the conclusion in Table 2 that flexible density estimation is superior when dimensionality is kept within reasonable limits. The sparsity constraints, although advantageous in high-dimensional contexts, offer negligible benefits and may even detrimentally affect performance in low-dimensional situations where ample data is available to accurately estimate complete covariance structures. This observation delineates a crucial practical guideline: method selection must consider anticipated dimensionality, with IVA-M-EMK favored for bimodal scenarios and IVA-SPICE indispensable for scenarios involving three or more modalities.

The lower panel offers a compelling illustration of the dimensionality scaling challenge and the solution provided by IVA-SPICE. The x-axis denotes the quantity of integrated modalities on a logarithmic scale ranging from $2^1 = 2$ to $2^2 = 4$ (including intermediate results for three modalities), whereas the y-axis illustrates the attained F1-score. IVA-M-EMK's trajectory demonstrates significant degradation: commencing at approximately 77% with two modalities, decreasing to around 60% with three modalities, and plummeting to 55% with four modalities, reflecting a substantial 22 percentage point decline, equating to a relative performance loss of over 28%. This disastrous scaling failure illustrates the exponentially increasing joint dimensionality that surpasses the limitations of static training data. IVA-A-GGD and other parametric variants exhibit analogous degradation patterns with differing intensities, affirming that the dimensionality issue surpasses particular distributional assumptions and instead signifies inherent statistical constraints of unregularized high-dimensional estimation.

IVA-SPICE's divergent trajectory exhibits notable resilience: performance remains consistently stable across the dimensionality spectrum, decreasing only marginally from approximately 68% at two modalities to 73% at four modalities, indicating a slight enhancement rather than deterioration. This paradoxical outcome, enhanced performance with greater dimensionality, demonstrates IVA-SPICE's ability to selectively utilize genuinely informative cross-modal interactions while mitigating spurious correlations. The supplementary modalities offer complementary discriminative signals that, when effectively integrated via sparse dependency modeling, improve classification accuracy instead of diminishing it. The juncture at which IVA-SPICE surpasses all rival methodologies is located at the three-modality threshold, specifically where Table 2 delineates the phase transition. The consistent pattern observed across various experimental manipulations confirms that the effects are authentic statistical phenomena,

not artifacts of the dataset or experimental noise.

4.5.3 Interpretability Through Learned Sparsity Patterns

In addition to quantitative performance benefits, IVA-SPICE's sparse precision matrices offer interpretable insights into the cross-modal dependency structures that underpin misinformation detection. The analysis of learned sparsity patterns within the four-modality configuration uncovers systematic trends that correspond with the intuitive comprehension of which modality interactions possess authentic discriminative signals. The interaction between text and image consistently demonstrates the highest dependency strength, as indicated by the non-zero precision matrix entries with the greatest magnitude, confirming the pivotal role of these modalities established in prior analyses. This significant reliance illustrates essential traits of multimodal misinformation, wherein textual assertions and visual evidence must uphold semantic coherence in accurate content, with deception frequently emerging through discrepancies between text and images.

The image-to-text features, which are synthetically generated textual descriptions of visual content, demonstrate intermediate dependency patterns. Robust correlations arise between image-to-text conversions and both the inherent features of the original image (as anticipated due to their derivational relationship) and the features of the original text (indicating semantic congruence when both accurately depict the same content). Nonetheless, the dependency between image-to-text and topic modeling often reveals nearly zero precision matrix entries, signifying conditional independence when considering other modalities. This pattern is interpretable: after considering text and image information, automated image descriptions offer negligible additional discriminative insight into corpus-level thematic patterns, primarily reiterating information already obtained through direct text and image analysis.

Topic modeling features exhibit the most minimal connectivity patterns, with numerous near-zero dependencies indicating that these features predominantly capture orthogonal discriminative signals. The comparatively weak associations probably indicate that topic modeling captures overarching thematic structures at the corpus level, rather than specific text-image consistency, rendering these features complementary yet less directly interactive with other modalities. This interpretation indicates potential merit in alternative fusion architectures that approach topic modeling as a distinct discriminative stream integrated via late fusion, rather than through joint density estimation, a direction deserving further exploration.

The sparsity patterns systematically adjust to the availability of training data, demonstrating a compelling data-dependent regularization effect. In data-scarce environments, acquired precision matrices attain maximal sparsity, imposing conditional independence among almost all non-essential modality pairs while retaining only the most significant dependencies. As the training sample size expands, sparsity diminishes moderately, permitting the retention of additional weak yet authentic dependencies, as adequate data becomes accessible to reliably estimate their parameters without the risk of overfitting. This adaptive behavior, characterized by the automatic adjustment of model complexity in relation to the available training data, offers a form of implicit cross-validation, wherein the ℓ_1 regularization strength is effectively calibrated during the optimization process to balance the bias-variance tradeoff suitable for the current sample size. This self-tuning feature is especially advantageous for operational deployment, where the availability of training data may fluctuate across domains or timeframes, guaranteeing consistent performance without necessitating manual hyperparameter adjustments for each deployment context.

4.5.4 Practical Implications for Operational Deployment

The comprehensive experimental validation establishes clear practical guidelines for designing multimodal misinformation detection systems deployed in operational environments. The choice between IVA-M-EMK and IVA-SPICE depends critically on the specific deployment context along multiple dimensions: number of available modalities, expected training data volume, computational resource constraints, and interpretability requirements. For systems integrating exclusively two modalities, the most common configuration in current social media platforms where text-image pairs dominate, IVA-M-EMK provides the optimal approach when computational resources permit flexible density estimation. The method achieves maximum performance through comprehensive modeling of potentially complex non-Gaussian joint distributions without imposing restrictive sparsity assumptions that might discard subtle but genuine dependencies. The text-image combination achieving 77.45% F1-score establishes a strong operational baseline for practical deployment.

However, as systems expand to incorporate three or more modalities, an increasingly common scenario as platforms provide richer metadata, automated caption generation becomes ubiquitous, and network analysis capabilities mature, IVA-SPICE transitions from optional enhancement to essential requirement. The method's superior performance across all

three-modality and four-modality configurations, combined with favorable sample complexity characteristics, establishes it as the clear choice for high-dimensional fusion scenarios.The 11.79 percentage point advantage over IVA-M-EMK in four-modality settings denotes the distinction between marginal utility and authentic reliability in detection capability, which directly impacts operational metrics such as false positive rates, false negative rates, and user confidence in automated moderation decisions. Additionally, the optimization process of IVA-SPICE is simplified by the computational advantages of sparse precision matrices, which enable efficient storage and computation. This results in wall-clock training and inference durations that are comparable or even superior to those of IVA-M-EMK in high-dimensional contexts.

The selection of optimal methods is similarly influenced by considerations regarding the availability of training data. In low-dimensional contexts, both methods can achieve robust performance, with IVA-M-EMK exhibiting marginal advantages for bimodal fusion, in situations with ample labeled data, comprising tens of thousands of annotated instances that encompass various misinformation strategies, subjects, and modality combinations. "'Operational reality frequently encounters limited annotation availability: the manual labeling of multimodal misinformation requires expert judgment, is labor-intensive and expensive, and rapidly becomes obsolete as adversarial strategies evolve. IVA-SPICE's exceptional sample efficiency, which achieved a 73% F1-score across four modalities despite the limited training data, provides substantial benefits in resource-constrained environments. In comparison to unregularized methods, the implicit regularization of sparsity constraints enables the estimation of parameters with greater reliability from smaller sample sizes, thereby reducing annotation costs while maintaining detection reliability.

Operational value that surpasses straightforward performance measurement is provided by the interpretability that learned sparsity patterns provide. The modality interactions that the system employs for classification are indicated by sparse precision matrices, which enable developers to determine whether the learned dependencies are consistent with domain knowledge and expert insights regarding the manifestation of misinformation across channels. Unanticipated dependency patterns, such as unexpectedly robust associations between modalities that were previously considered largely independent by domain experts, may indicate dataset biases, spurious correlations, or adversarial manipulation tactics that necessitate additional investigation. This transparency facilitates the continuous improvement of the system by means of informed feature engineering, focused data collection to address identified deficiencies, and adversarial

robustness analysis to anticipate potential evasion strategies. The interpretable structure is a significant departure from end-to-end deep learning methods that produce opaque high-dimensional representations devoid of clear semantic significance. This structure provides the requisite explainability for implementation in socially impactful content moderation applications.

4.5.5 Synthesis: Adaptive Multimodal Fusion Through Principled Sparsity

The comprehensive analysis of modality sufficiency demonstrates that the fundamental inquiry of whether two modalities are sufficient for the effective detection of misinformation does not yield a simple universal response, as it is contingent upon the operational context and fusion methodology. The experimental evidence indicates that the quantity of training data, the modeling methodology, and dimensionality are the primary factors that influence optimal architectural decisions. In resource-rich environments with ample training data, two modalities are both adequate and even advantageous when utilizing flexible density estimation techniques such as IVA-M-EMK. This approach achieves a robust 77.45% F1-score by thoroughly modeling intricate bimodal distributions and reducing the risk of overfitting. However, the implementation of appropriate regularization through structured sparsity is necessary to achieve a 73% improvement in performance, despite a fourfold increase in dimensionality, when extending beyond two modalities. IVA-SPICE enables effective scaling by autonomously identifying sparse cross-modal dependency structures, distinguishing genuine discriminative relationships from spurious correlations, and providing implicit regularization to reduce overfitting, a critical component of high-dimensional statistical learning.

The phase transition observed at the three-modality threshold has substantial implications for the practical design of systems. Designers must recognize that simplistic fusion methods can be detrimental in high-dimensional contexts, rather than perceiving multimodal fusion as a simple aggregation of information sources with the belief that additional modalities always enhance detection. The indiscriminate inclusion of modalities, without appropriate methodological adjustments, severely undermines efficacy, resulting in a 16 percentage point decline in IVA-M-EMK's performance when increasing from two to four modalities. Consequently, comprehensive fusion is inferior to selecting the single most effective modality. This discovery promotes the use of an adaptive architectural framework: systems should begin with fundamental bimodal fusion, which should be implemented using flexible methodologies. Subsequently, sparsity-aware techniques should be employed to integrate supplementary modalities only when

specific deployment criteria, such as varied misinformation strategies that require multimodal consistency verification, operating in low-data environments where implicit regularization is essential, or requiring interpretable dependency structures for transparency, justify the increased complexity.

Structured sparsity for implicit regularization, adaptive density estimation for distributional flexibility, and explicit dependency modeling for interpretability are all principled statistical methods that enable scaling beyond basic bimodal scenarios without compromising reliability. This is demonstrated by the efficacy of IVA-SPICE in managing multimodal dimensionality while maintaining competitive performance. The framework recognizes and capitalizes on the inherent sparsity of cross-modal misinformation patterns. Deception is manifested through specific targeted inconsistencies rather than widespread effects across all modality combinations, resulting in dependency structures that are well-suited for sparse representation. By aligning modeling assumptions with the genuine statistical characteristics of the detection problem, IVA-SPICE achieves advantageous scalability properties that are essential for operational deployment in progressively multimodal information environments. This investigation demonstrates that the aggregation of information sources is not sufficient for the effective high-dimensional fusion of data. Additionally, principled statistical frameworks are necessary to differentiate between signal and noise through structured regularization. These insights provide practical advice for the development of advanced misinformation detection systems that can leverage the diverse multimodal environment of modern social media while avoiding the dimensionality challenges that afflict simplistic fusion methods.

5 CONCLUSIONS AND FUTURE WORK

5.1 General Conclusions

This research aimed to ascertain the viability of a multimodal model for misinformation detection through Independent Vector Analysis. Following an extensive series of tests and validations, we have concluded that the proposed method is both operational and effective for its intended purpose.

It is essential to acknowledge that the model's success relies on specific premises. The statistical model of source component vectors (SCV), demonstrated to be reliable in all our experiments, serves as the foundation of the approach. An additional phase in its operation involves the extraction of features from diverse data modalities, a critical and indispensable step in the process. This thesis establishes that an efficient multimodal system for the proposed task can be constructed, contingent upon the fulfillment of specific conditions.

Throughout the experiments, a notable secondary conclusion emerged: the quantity of data modalities employed directly influences the classifier's performance. The experiments demonstrated that augmenting the number of modalities from two to three, four, or five enhances detection performance, albeit in certain cases the improvement is merely marginal. This was evidenced by the administration of the tests. This discovery has substantial implications for practical applications, necessitating meticulous evaluation of the trade-off between computational complexity and performance enhancement.

5.2 Summary of Contributions

5.2.1 Theoretical Advancements

This thesis aims to provide several theoretical contributions to the domains of blind source separation and multimodal analysis. Initially, we devised the IVA-M-EMK algorithm. This algorithm integrates semiparametric density estimation via multivariate entropy maximization within the established IVA framework. This formulation employs both global and local measuring functions to construct flexible probability density functions without imposing strict parametric assumptions. The integration of quasi-Monte Carlo methods for multidimensional integration facilitates computations in higher-dimensional spaces, thereby overcoming a notable limitation of previous methods.

Secondly, we established a theoretical link between the quality of source separation and the precision of density estimation. We established, via analytical derivation and empirical validation, that the relative entropy between the true distribution and the estimated distribution directly influences separation performance. This insight provides a principled basis for comprehending why flexible density models outperform fixed parametric assumptions in complex multimodal scenarios.

5.2.2 Multimodal Application to Misinformation Detection

The application of IVA-based methods for detecting instances of misinformation constitutes a novel contribution to the domains of signal processing and computational journalism. We established that misinformation can be effectively modeled as a blind source separation problem. In this model, each modality (text, image, embeddings) signifies a cacophony of semantic signals that underlie the surface. This formulation enables the extraction of latent components that identify both cross-modal consistencies and inconsistencies, which are critical indicators of potential misinformation.

We have established a novel standard for multimodal fusion in misinformation detection through our experimental framework, which can concurrently process up to four distinct modalities. The results demonstrate that IVA-M-EMK performs optimally with pairs of strongly aligned modalities, while IVA-SPICE shows enhanced performance as the number of modalities rises. This underscores the importance of aligning the model's complexity with the data's attributes.

5.2.3 Comparative Evaluation and Explainability Insights

Our findings indicated that flexible density modeling methods outperform conventional fixed-prior IVA variants. This was achieved through comprehensive empirical assessments on both synthetic and real-world datasets. IVA-M-EMK consistently surpasses IVA-G, IVA-L, and IVA-GGD in scenarios involving intricate, non-Gaussian source distributions, as indicated by the conducted comparative analysis. The integration of sparsity constraints into IVA-SPICE provides a systematic approach for managing high-dimensional multimodal spaces while maintaining interpretability.

One of the most notable advantages of our approach compared to black-box deep learning methods is its explainability. The model elucidates the elements of the data that

influence classification decisions by deconstructing multimodal inputs into comprehensible source components. This enables the model to offer these opportunities. This requirement is essential for successful deployment in sensitive areas such as content moderation and fact-checking.

The practical implications of this work extend beyond academic contributions. This research has practical applications. The proposed framework has many practical advantages: The unsupervised feature extraction phase reduces reliance on labeled data, which is often scarce and expensive; the modular architecture allows seamless incorporation of new modalities; and the statistical basis quantifies uncertainty, which is essential for high-stakes decision-making.

5.3 Future Work

5.3.1 Integration of Discriminative Priors ino IVA Framework

The standard Independent Vector Analysis algorithm, due to its unsupervised nature, pursues statistical independence without regard for class structure. This represents a primary limitation of the algorithm. Although this method effectively reveals concealed sources, it does not ensure that the extracted components are ideally aligned with classification objectives. This observation motivates the integration of discriminative information directly into the IVA optimization framework.

The integration of class-aware constraints into the source separation process yields a solution to this limitation through discriminant-enhanced Independent Vector Analysis (IVA). The resulting framework concurrently aims to uphold statistical independence among sources while maximizing class separability. This is achieved by enhancing the conventional IVA objective with Fisher's Linear Discriminant criterion. The dual objective is formalized by employing a regularized cost function that balances unsupervised structure discovery with structured discriminative learning.

A trade-off parameter is introduced by the mathematical formulation, regulating the relative significance of independence and discrimination. This method can learn representations that decompose the multimodal signal into interpretable components and align these components with the decision boundaries pertinent to classification when class labels are accessible during training. This is particularly advantageous in identifying misinformation, where subtle differences between legitimate and misleading content may be obscured by statistically significant yet

irrelevant patterns.

The discriminative extension is particularly advantageous in scenarios marked by class imbalance, a prevalent issue in misinformation detection. In these situations, legitimate content vastly outnumbers instances that are misleading. When the method explicitly enhances class separability, it more effectively identifies and amplifies the discriminative signals that would otherwise be overshadowed by majority-class statistics in purely unsupervised methods. This is due to the method's explicit optimization for class separability.

Initial implementations of this concept, such as our recent work on Discriminant Independent Vector Analysis (DIVA) (Maia *et al.*, 2025), have clearly demonstrated consistent improvements in comparison to standard IVA variants across a wide range of classification benchmarks. This approach shows particular promise for the detection of multimodal misinformation, where the combination of unsupervised multimodal fusion and supervised discriminative learning has the potential to improve both the accuracy of detection and the interpretability of the model. Throughout the entirety of this thesis, this supervised extension has been developed to supplement the unsupervised methods that have been developed. It provides a comprehensive range of approaches that are contingent upon the availability of labeled training data.

5.3.2 Expansion of Multimodal Analysis

In addition to the text-image combinations that have been investigated up until this point, the framework that was developed in this thesis naturally extends to incorporate additional sensory modalities. The integration of temporal and auditory data via video and audio stream modalities is among the most compelling enhancements. These modalities carry rich signals for the detection of misinformation, but they also introduce substantial technical challenges.

The use of video analysis provides a number of distinct benefits when it comes to identifying sophisticated manipulation techniques. With the help of the temporal dimension, discontinuities and artifacts that were previously concealed in isolated frames are brought to light. These include unnatural motion patterns, inconsistent lighting across sequences, and temporal anomalies in the behavior of objects. The static features extracted from individual images are supplemented by these dynamic signatures of manipulation, which have the potential to reveal deep fakes and synthetic media that might be undetectable by a single-frame inspection.

Similar to how acoustic analysis opens up new detection channels, our framework does not yet have the capability to investigate these channels. Inconsistencies between voice

characteristics and visual lip movements, prosodic irregularities, unnatural formant distributions, and subtle markers of synthesis or manipulation are all things that can be found in speech patterns. Mismatched background sounds or artificially inserted audio tracks are frequently indicators of content manipulation, and environmental audio provides additional verification signals to help identify these instances.

It is not the conceptual integration of these modalities that presents the most significant challenge in terms of incorporating them; rather, the IVA framework was designed to easily accommodate additional data streams. It is more accurate to say that the difficulty lies in the computational complexity of processing continuous, high-dimensional temporal signals. When compared to static images, video streams produce orders of magnitude more data. In order to maintain temporal coherence, specialized architectures that preserve sequential dependencies across frames are required. When attempting to perform joint audio-visual analysis, the difficulty of the task exponentially increases because synchronization and cross-modal temporal alignment become extremely important.

5.3.3 Application of Large Language Models

It is possible that our IVA-based framework could be incorporated into the powerful new tools for detecting misinformation that have emerged as a result of the proliferation of Large Language Models (LLMs). It is possible to investigate this integration from two primary perspectives:

Multimodal Feature Extraction: Feature extraction that is more complex can be accomplished with the help of LLMs. A textual content analysis could be performed by an LLM, for instance, in order to determine the likelihood of the content being misleading. Vision-language models, on the other hand, could extract joint representations from text-image pairs. It is possible that native embeddings from these models could be used as input features for IVA, which could potentially capture more nuanced semantic relationships than traditional methods of feature extraction.

Explainable AI (XAI): One of the most important steps toward developing artificial intelligence that can be trusted is the ability to understand the reasons behind why a model identifies content as misleading. Binary classification is typically performed by traditional fake content detection models. However, these models do not provide explanations that are understandable by humans. This restricts the interpretability of the results and undermines

their credibility. In this context, large language models (LLMs) offer a promising path toward explainable misinformation detection by generating natural language rationales that align with human reasoning (Li *et al.*, 2025). This is accomplished within the context of the current situation. During the process of identifying a manipulated image, for instance, an LLM might justify its decision by highlighting particular visual inconsistencies or semantic contradictions. This approach surpasses conventional explanation methods such as LIME, as LLMs are capable of jointly analyzing visual and linguistic cues, producing coherent, human-intelligible explanations that bridge the gap between perception, reasoning, and transparency.

5.4 Final Remarks

This thesis has demonstrated that Independent Vector Analysis, when augmented with flexible density estimation and appropriate regularization, provides a powerful framework for multimodal misinformation detection. The journey from theoretical development to practical application has revealed both the potential and the challenges of statistical approaches to this critical societal problem.

As misinformation tactics continue to evolve, so too must our detection methods. The foundations laid in this work, namely, statistical modeling, multimodal fusion, and interpretable feature extraction, provide a solid foundation for future innovations. The ultimate goal remains clear: to develop detection systems that are not only accurate and efficient but also transparent and trustworthy, serving as reliable guardians of information integrity in our increasingly connected world.

The fight against misinformation is far from over, but the tools and insights developed in this research represent significant progress toward that goal. By continuing to bridge the gap between statistical theory and practical application, we can build systems that help preserve the integrity of our information ecosystem while respecting the complexity and nuance of human communication.

REFERENCES

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**, New York, 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**, New York, v. 31, n. 3, p. 18–33, 2014.

ANDERSON, M.; ADALI, T.; LI, X. Joint blind source separation with multivariate Gaussian model: algorithms and performance analysis. **IEEE Transactions on Signal Processing**, New York, v. 60, n. 4, p. 1672–1683, 2012.

ANDERSON, M.; ADALI, T.; LI, X.-L. Joint blind source separation with multivariate Gaussian model: algorithms and performance analysis. **IEEE Transactions on Signal Processing**, New York, v. 60, n. 4, p. 1672–1683, 2012.

ANDERSON, M.; FU, G.; PHLYPO, R.; ADALI, T. Independent vector analysis, the Kotz distribution, and performance bounds. *In*: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, 38., 2013, Vancouver. **Proceedings [...]**. New York: IEEE, 2013. p. 3243–3247.

ANDERSON, M.; FU, G.; PHLYPO, R.; ADALI, T. Independent vector analysis: identification conditions and performance bounds. **IEEE Transactions on Signal Processing**, New York, v. 62, n. 17, p. 4399–4410, 2014.

ANDERSON, M.; FU, G.; PHLYPO, R.; ADALI, T. Independent vector analysis: identification conditions and performance bounds. **IEEE Transactions on Signal Processing**, New York, v. 62, n. 17, p. 4399–4410, 2014.

ANDERSON, M.; FU, G.-S.; PHLYPO, R.; ADALI, T. Joint blind source separation with multivariate Gaussian model: algorithms and performance analysis. **IEEE Transactions on Signal Processing**, New York, v. 60, n. 4, p. 1672–1683, 2012.

ANDERSON, M.; LI, X.; RODRIGUEZ, P.; ADALI, T. An effective decoupling method for matrix optimization and its application to the ICA problem. *In*: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, 37., 2012, Kyoto. **Proceedings [...]**. New York: IEEE, 2012. p. 1885–1888.

BHINGE, S.; BOUKOUVALAS, Z.; LEVIN-SCHWARTZ, Y.; ADALI, T. IVA for abandoned object detection: exploiting dependence across color channels. *In*: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, 41., 2016, Shanghai. **Proceedings [...]**. New York: IEEE, 2016. p. 2494–2498.

BOUKOUVALAS, Z. **Development of ICA and IVA algorithms with application to medical image analysis**. Tese (Doutorado em Matemática Aplicada) — University of Maryland, Baltimore County, Baltimore, 2017. Disponível em: https://arxiv.org/abs/1801.08600. Acesso em: 24 nov. 2025.

BOUKOUVALAS, Z.; ELTON, D.; CHUNG, P.; FUGE, M. **Independent vector analysis for data fusion prior to molecular property prediction with machine learning**. 2018. Disponível em: https://arxiv.org/abs/1811.00628. Acesso em: 24 nov. 2025.

BOUKOUVALAS, Z.; FU, G.; ADALI, T. An efficient multivariate generalized Gaussian distribution estimator: application to IVA. *In*: ANNUAL CONFERENCE ON INFORMATION SCIENCES AND SYSTEMS, 49., 2015, Baltimore. **Proceedings [...]**. New York: IEEE, 2015. p. 1–4.

BOUKOUVALAS, Z.; SAID, S.; BOMBRUN, L.; BERTHOUMIEU, Y.; ADALI, T. A new Riemannian averaged fixed-point algorithm for MGGD parameter estimation. **IEEE Signal Processing Letters**, New York, v. 22, n. 12, p. 2314–2318, 2015.

COMON, P. Independent component analysis, a new concept? **Signal Processing**, Amsterdam, v. 36, n. 3, p. 287–314, 1994. Disponível em: https://www.sciencedirect.com/science/article/pii/0165168494900299. Acesso em: 24 nov. 2025.

COVER, T. M.; THOMAS, J. A. **Elements of information theory**. 2. ed. Hoboken: Wiley-Interscience, 2006.

DAMASCENO, L. d. P. **Independent vector analysis using semi-parametric density estimation via multivariate entropy maximization**. Dissertação (Mestrado em Engenharia de Teleinformática) — Universidade Federal do Ceará, Fortaleza, 2021.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. **BERT: pre-training of deep bidirectional transformers for language understanding**. 2019. Disponível em: https://arxiv.org/abs/1810.04805. Acesso em: 24 nov. 2025.

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

FU, G.; BOUKOUVALAS, Z.; ADALI, T. Density estimation by entropy maximization with kernels. *In*: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, 40., 2015, Brisbane. **Proceedings [...]**. New York: IEEE, 2015. p. 1896–1900.

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

JAYNES, E. T. Information theory and statistical mechanics. **Physical Review**, New York, v. 106, n. 4, p. 620–630, 1957.

JAYNES, E. T. Information theory and statistical mechanics. **Physical Review**, New York, v. 106, n. 4, p. 620–630, 1957.

KIM, T. Real-time independent vector analysis for convolutive blind source separation. **IEEE Transactions on Circuits and Systems I: Regular Papers**, New York, 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. **IEEE Transactions on Audio, Speech, and Language Processing**, New York, 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*: ROSCA, J.; ERDOGMUS, D.; PRíNCIPE, J. C.; HAYKIN, S. (Ed.). **Independent component analysis and blind signal separation**. Berlin: Springer, 2006, (Lecture Notes in Computer Science, v. 3889). p. 165–172.

LI, X.; ADALI, T. Independent component analysis by entropy bound minimization. **IEEE Transactions on Signal Processing**, New York, v. 58, n. 10, p. 5151–5164, 2010.

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

LI, Y.; LIU, X.; WANG, X.; LEE, B. S.; WANG, S.; ROCHA, A.; LIN, W. FakeBench: probing explainable fake image detection via large multimodal models. **IEEE Transactions on Information Forensics and Security**, New York, v. 20, p. 8730–8745, 2025.

MAIA, M.; DAMASCENO, L.; CORIZZO, R.; CAVALCANTE, C. C.; BOUKOUVALAS, Z. Class-constrained independent vector analysis for discriminative multimodal representation learning. *In*: IEEE INTERNATIONAL CONFERENCE ON DATA MINING, 25., 2025, Washington. **Proceedings [...]**. New York: IEEE, 2025.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. **Efficient estimation of word representations in vector space**. 2013. Disponível em: https://arxiv.org/abs/1301.3781. Acesso em: 24 nov. 2025.

NIEDERREITER, H. **Random number generation and quasi-Monte Carlo methods**. Philadelphia: Society for Industrial and Applied Mathematics, 1992.

O'LEARY, D. P. **Scientific computing with case studies**. Philadelphia: Society for Industrial and Applied Mathematics, 2009.

RAJABI, J.; OKECHUKWU, S.; MOUSAVI, A.; CORIZZO, R.; CAVALCANTE, C. C.; BOUKOUVALAS, Z. Event-based multi-modal fusion for online misinformation detection in high-impact events. *In*: IEEE INTERNATIONAL CONFERENCE ON BIG DATA, 2024, Washington. **Proceedings [...]**. New York: IEEE, 2024. p. 3301–3308.

REXHEPI, E. **Independent vector analysis with sparse inverse covariance estimation**. Dissertação (Mestrado em Matemática) — American University, Washington, 2022. Disponível em: https://aura.american.edu/articles/thesis/Independent_vector_analysis_with_sparse_inverse_covariance_estimation/23861118. Acesso em: 24 nov. 2025.

ROJO-ÁLVAREZ, J. L.; MARTíNEZ-RAMON, M.; MUñOZ-MARI, J.; CAMPS-VALLS, G. **Digital signal processing with kernel methods**. Hoboken: Wiley-IEEE Press, 2018.

SAHOO, S. R.; GUPTA, B. B. Multiple features based approach for automatic fake news detection on social networks using deep learning. **Applied Soft Computing**, Amsterdam, v. 100, p. 106983, 2021. Disponível em: https://www.sciencedirect.com/science/article/pii/S1568494620309224. Acesso em: 24 nov. 2025.

SCHWARZ, S.; THEóPHILO, A.; ROCHA, A. EMET: embeddings from multilingual-encoder transformer for fake news detection. *In*: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, 45., 2020, Barcelona. **Proceedings [...]**. New York: IEEE, 2020. p. 2777–2781.

VIA, J.; ANDERSON, M.; LI, X.-L.; ADALI, T. A maximum likelihood approach for independent vector analysis of Gaussian data sets. *In*: IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, 21., 2011, Beijing. **Proceedings [...]**. New York: IEEE, 2011. p. 1–6.