Optimal Transport in linear Independent Component Analysis

Master Thesis (M.Sc.)
in
Quantitative Data Science Methods
Psychometrics, Econometrics and Machine Learning
Faculty of Economics and Social Sciences
at the University of Tübingen
Methods Center

submitted by
Ashutosh Jha
from Tuebingen
(Matriculation number: 6639615)

Submitted in Tübingen
April 26, 2026

1. Supervisor:	Prof. Dr. Joachim Grammig[1]
2. Supervisor:	Prof. Dr. Michel Besserve[2,3]
3. Supervisor:	Dr. Simon Buchholz[2]

1. Reviewer:	Prof. Dr. Joachim Grammig[1]
2. Reviewer:	Prof. Dr. Augustin Kelava[4]

[1] Chair of Statistics, Econometrics and Empirical Economics, University of Tübingen

[2] Empirical Inference, Max Planck Institute for Intelligent Systems, Tübingen

[3] Institute for Machine Learning and AI, TU Braunschweig

[4] Methods Center, University of Tübingen

Declaration of Academic Integrity

Hereby, I declare that I have composed the presented paper independently on my own and without any other resources than the ones indicated. All thoughts taken directly or indirectly from external sources are properly denoted as such.

This thesis has neither been previously submitted as a whole or in any significant part as part of any other examination process, and the thesis has not been published as a whole or in any significant part, and that the copy submitted in electronic file form is identical in content to the bound copies submitted.

Name	Place, Date

Signature

Contents

1 Introduction **3**

1.1 Existing Methodologies 3

1.2 Applications 4

1.3 Main Contribution 4

1.4 Structure of the Thesis 5

2 Theoretical Foundations **6**

2.1 Linear Independent Component Analysis (ICA) 6

2.1.1 The Insufficiency of Principal Component Analysis 7

2.1.2 Identifiability and Ambiguities 7

2.1.3 Centering and Whitening 8

2.1.4 Statistical and Thermodynamic Intuitions 8

2.2 An Information Geometry Perspective on ICA 9

2.2.1 Product and Gaussian Manifolds 9

2.2.2 Kullback-Leibler Pythagorean Identities 9

2.2.3 Independence, Correlation, and Non-Gaussianity 11

2.3 Optimal Transport Theory 12

2.3.1 The Monge Problem: Push-Forwards and Pull-Backs 12

2.3.2 The Kantorovich Relaxation 13

2.3.3 The Wasserstein Distances (W_1 and W_2) 14

3 The Optimal Transport ICA (OT-ICA) Framework **16**

3.1 Optimal Transport Distance as a Contrast for ICA 16

3.2 Theoretical Motivation: Bounding Mixture Non-Gaussianity 17

3.2.1 The W_2 Upper Bound 17

3.2.2 Comonotonicity and the Proof of Strict Inequality 19

3.3 Empirical Landscape: W_1 versus W_2^2 21

3.3.1 Empirical Gradients and Derivative Volatility 21

4 Algorithmic Enhancements **24**

4.1 The Baseline OT-ICA Algorithm 24

4.2 Computational Bottlenecks and Solutions 24
4.2.1 Exact Analytical Gaussian Targets 24
4.2.2 Batched Vectorization for Restarts 25
4.3 Optimization on the Stiefel Manifold 26
4.3.1 The Riemannian Gradient 26
4.3.2 Retraction via Symmetric Decorrelation 27
4.4 Total Complexity and Statistical Efficiency 27
4.5 Navigating Discrete Optimization Landscapes 28
4.5.1 The Discontinuous Landscape of Discrete CDFs 28
4.5.2 Continuous Smoothing via Gaussian Dithering 29
4.5.3 Escaping Local Minima with Stochastic Mini-Batching . . . 30
4.6 OT-ICA Algorithm Overview 30
4.7 Limitations of Fixed-Point Algorithms for OT-ICA 31

5 Experimental Evaluation and Applications 34
5.1 Evaluation Metrics And Methodology 34
5.1.1 The Amari Performance Index 34
5.1.2 Computational Resource Regimes 34
5.2 OT-ICA Methodology Validation 36
5.3 Computational & Temporal Scaling 37
5.4 Algorithmic Limitations of FastICA 38
5.4.1 The Zero Negentropy Condition 39
5.4.2 Empirical Results: Zero Negentropy 39
5.4.3 The Vanishing Curvature Condition 39
5.4.4 Empirical Results: Vanishing Curvature 40
5.5 The Generalized Hybrid Mixture Stress Test 42
5.5.1 Experimental Setup 42
5.5.2 Empirical Results: Hybrid Mixture Stress Test 42
5.6 Discrete Only Mixtures: A Unique Challenge 43
5.6.1 Empirical Results: Discrete Only Mixtures 43
5.6.2 Optimization Properties of Discrete Data 43
5.6.3 Empirical Results: Harsh Discrete Environments 44
5.7 Synthesis: The Case for OT-ICA 45
5.8 Application: EEG Artifact Removal 46
5.8.1 The Challenge of Volume Conduction 46
5.8.2 Empirical Results on Clinical EEG Data 46
5.9 Application: Price Discovery and Information Shares 47
5.9.1 Non-Gaussianity in Econometric Market Microstructure . . . 47
5.9.2 Economic Identification Strategy 47

5.9.3 Empirical Results on Simulated Market Data 47

6 Conclusion **50**

6.1 Summary of Contributions 50

6.2 Limitations 51

6.3 Future Work 51

A Mathematical Proofs and Derivations **53**

A.1 Derivation of the FastICA Fixed-Point Failure Condition 53

References **56**

List of Figures

2.1 Information Geometry of ICA 10
2.2 Optimal Transport Push-Forward 12
2.3 The Three Scenarios of Kantorovich Optimal Transport 13

3.1 Gradient Landscapes of W_1 and W_2^2 22

4.1 Riemannian Optimization on the Stiefel Manifold 26
4.2 Gaussian Dithering of Discrete CDFs 29

5.1 Baseline Source Recovery via OT-ICA 36
5.2 Separation Accuracy on Laplacian Sources 37
5.3 Execution Time Scaling 38
5.4 Performance on Zero-Negentropy Failure Mode 40
5.5 The Vanishing Curvature Counterexample 41
5.6 Performance on Trimodal Failure Mode 41
5.7 General Hybrid Mixture Stress Test 42
5.8 Performance on Isolated Discrete Pools 43
5.9 Optimization Surface of Discrete Mixtures 44
5.10 Performance on Harsh Discrete Environments 45
5.11 EEG Artifact Removal via OT-ICA 46
5.12 Recovery of Structural Shocks via OT-ICA 48
5.13 Empirical Distribution of Estimated Information Shares 49

List of Tables

5.1 Heuristic thresholds for interpreting the Amari Performance Index. 35
5.2 Standardized Computational Regimes 35
5.3 Baseline Matrices and Amari Error 37
5.4 Non-Gaussianity Sensitivity Analysis 44
5.5 Simulation Results for Information Share Recovery 48

Chapter 1

Introduction

Blind source separation, specifically formulated as independent component analysis (ICA), is a fundamental problem in modern statistics (Jutten & Herault, 1985). The primary objective of ICA is to recover a set of hidden, statistically independent source signals from an observed set of mixtures. In this thesis, the problem is strictly bounded to the linear mixture setting, wherein the observations are assumed to be generated by an unknown, invertible linear combination of the latent sources.

To achieve exact separation, ICA imposes strict statistical independence among the recovered components. This approach is frequently contrasted with principal component analysis (PCA). While PCA diagonalizes the covariance matrix to ensure that the resulting components are uncorrelated, ICA enforces the mathematical condition of full statistical independence (Hyvärinen, Karhunen, & Oja, 2001, Chapter 1). This rigorous criterion allows ICA to systematically disentangle complex latent variables that remain mixed under simple decorrelation.

1.1 Existing Methodologies

In practice, the standard ICA pipeline begins by centering and whitening the observed mixture. Whitening is a necessary preprocessing step that removes linear correlations and scales the marginal variances to unity. Once the data is whitened, the search for independent components is geometrically restricted to finding an optimal orthogonal rotation matrix on the Stiefel manifold, a space of orthogonal matrices. (Absil, Mahony, & Sepulchre, 2008).

Traditional ICA algorithms navigate this rotational search by optimizing surrogate contrast functions to identify non-Gaussian projections (Hyvärinen et al., 2001, Chapter 8). These classical methods rely on surrogate information-theoretic measures, such as negentropy approximations or cumulant-based matching, to quantify the non-Gaussianity of the unmixed components (Amari, Cichocki, &

Yang, 1996).

While computationally efficient, these parametric approximations evaluate specific statistical moments rather than the full probability density. Consequently, we find them to fail to capture the complete geometric structure of the underlying empirical distributions, leading to optimization failures in heterogeneous mixtures.

1.2 Applications

The ability to blindly separate independent sources without parametric assumptions regarding their underlying distributions has widespread utility across diverse scientific domains. In neuroscience, ICA serves as the standard clinical technique for isolating and removing high-amplitude biological artifacts, such as ocular and muscular signals, from continuous electroencephalogram (EEG) and magnetoencephalogram (MEG) recordings (Hyvärinen et al., 2001, Chapter 21).

Beyond biological signal processing, the framework is foundational in digital communications and computer vision. In telecommunications, ICA is deployed to separate overlapping transmissions in code-division multiple access (CDMA) systems, while in image processing, it is utilized to extract localized, oriented, and bandpass features from natural images, effectively mimicking the processing functions of the mammalian primary visual cortex (Hyvärinen et al., 2001, Chapters 22-23).

Furthermore, the methodology is increasingly utilized in advanced econometrics and market microstructure. In financial price discovery, ICA facilitates the identification of latent structural shocks, allowing researchers to uniquely resolve the orthogonal ambiguities inherent in standard vector error correction models (Zema & Cordoni, 2025). Across all these distinct domains, the validity of the downstream scientific or economic analysis depends strictly on the reliable mathematical convergence of the underlying ICA optimization algorithm.

1.3 Main Contribution

The primary methodological contribution of this thesis is the formulation and investigation of an optimal transport framework for independent component analysis (OT-ICA). The core objective reduces to searching the Stiefel manifold for the rotation matrix that maximizes the non-Gaussianity of the projected data.

To achieve this, we propose utilizing the Wasserstein distance as the objective function for the rotational search. Our objective is to find the orthogonal candidates that maximize the Wasserstein distance to a standard normal distribution.

Because the Wasserstein metric evaluates the true global geometry of the empirical distribution using order statistics (Villani, 2003), it provides a non-parametric and highly sensitive measure of distance to Gaussianity. We find that this helps in successfully circumventing the structural blind spots of standard surrogate measures.

1.4 Structure of the Thesis

The remainder of this thesis is structured as follows. Chapter 2 establishes the theoretical foundations of linear ICA, information geometry, and optimal transport. Chapter 3 formalizes the OT-ICA framework, proving that the squared Wasserstein (W_2^2) distance reliably bounds mixture non-Gaussianity. Chapter 4 details the algorithmic enhancements required to optimize this metric efficiently on the Stiefel manifold. Chapter 5 presents our empirical evaluation, demonstrating the failure modes of standard proxy optimization and how OT-ICA addresses these limitations. Finally, Chapter 6 concludes the thesis with a summary of findings and directions for future research.

Chapter 2

Theoretical Foundations

This chapter establishes the mathematical and conceptual foundations necessary for linear independent component analysis. We first formalize the linear mixing model and its identifiability conditions, subsequently detailing the statistical and thermodynamic intuitions that motivate the framework. Following this, we introduce an information-geometric perspective linking independence to non-Gaussianity, and finally, formalize the principles of optimal transport that govern the proposed optimization metric.

2.1 Linear Independent Component Analysis (ICA)

The foundational assumption of the linear Independent Component Analysis model is that the observed signals are generated by a linear combination of statistically independent latent sources. We formalize this generative model and its required constraints as follows:

Theorem 2.1 The Linear ICA Model and Identifiability, (Comon, 1994). *Let the observed data be denoted by a random vector* $\mathbf{X} \in \mathbb{R}^d$*, and the latent sources by* $\mathbf{S} \in \mathbb{R}^d$*. The linear ICA model assumes:*

a) ***Linear Mixing:*** $\mathbf{X} = \mathbf{AS}$*, where* $\mathbf{A} \in \mathbb{R}^{d \times d}$ *is an unknown, invertible mixing matrix.*

b) The components $s_1, s_2, \ldots, s_d$ *of* $\mathbf{S}$ *are mutually statistically independent.*

c) At most one of the source components s_i *follows a Gaussian distribution.*

Under these conditions, the mixing matrix $\mathbf{A}$ *can be uniquely identified from the observations* $\mathbf{X}$*, up to a permutation and scaling of its columns.*

Expanded into vector notation, the structural mixing represents the observed

variables as

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1d} \\ a_{21} & a_{22} & \cdots & a_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ a_{d1} & a_{d2} & \cdots & a_{dd} \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_d \end{bmatrix}. \tag{2.1}$$

The core objective of ICA is to estimate an unmixing matrix $\mathbf{B} \approx \mathbf{A}^{-1}$ such that the recovered components $\mathbf{Z} = \mathbf{BX}$ are as statistically independent as possible, thereby reconstructing the original sources $\mathbf{S}$.

2.1.1 The Insufficiency of Principal Component Analysis

Principal Component Analysis (PCA) is frequently utilized as a baseline technique to decorrelate observed data. This corresponds to finding the orthogonal matrix $\mathbf{V}^\top$ derived from the eigenvalue decomposition of the data's covariance matrix. However, rendering a sample strictly decorrelated is mathematically insufficient for achieving blind source separation when the underlying data is non-Gaussian.

As noted by MacKay (2003), PCA is fundamentally an L_2-norm variance maximization technique; it is inherently sensitive to the scaling of the variables and heavily influenced by outliers. More critically, PCA relies exclusively on second-order statistics. Once the data is decorrelated and scaled (whitened) such that $\mathbb{E}[\mathbf{X}\mathbf{X}^\top] = \mathbf{I}_d$, the PCA criterion is satisfied for any arbitrary orthogonal rotation. While decorrelation implies strict independence for multivariate Gaussian distributions, higher-order dependencies remain intact for non-Gaussian data. Consequently, ICA must transcend second-order moments, traversing the space of orthogonal matrices to locate the rotation that eliminates these higher-order dependencies and achieves statistical independence.

2.1.2 Identifiability and Ambiguities

The non-Gaussianity constraint defined in Theorem 2.1 is a strict mathematical requirement. If two or more sources are Gaussian, any orthogonal transformation of those specific sources yields another set of strictly independent Gaussian variables, rendering the mixing matrix $\mathbf{A}$ impossible to determine uniquely.

Furthermore, even when the non-Gaussianity assumption is satisfied, the model contains two inherent ambiguities. The first is a scaling ambiguity: because both $\mathbf{A}$ and $\mathbf{S}$ are unknown, any scalar multiplier assigned to a source s_i can be perfectly cancelled by dividing the corresponding column of $\mathbf{A}$ by the same scalar. Therefore, the exact variances of the unobserved sources cannot be inferred. The second is a permutation ambiguity: the model treats the sum of sources symmetrically,

meaning the indexing order of the recovered components is arbitrary.

2.1.3 Centering and Whitening

To simplify the optimization landscape, the observed data $\mathbf{X}$ is subjected to a two-step preprocessing phase consisting of centering and whitening. Centering ensures that the data has a zero mean by subtracting the expectation, $\mathbb{E}[\mathbf{X}] = \mathbf{0}$. Given centered data, the covariance matrix simplifies to $\mathrm{Cov}(\mathbf{X}) = \mathbb{E}[\mathbf{X}\mathbf{X}^\top]$. Whitening transforms the observed vector into a new vector $\widetilde{\mathbf{X}}$ whose components are uncorrelated and possess unit variance. This is achieved using the eigenvalue decomposition of the covariance matrix

$$\mathrm{Cov}(\mathbf{X}) = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top \tag{2.2}$$

where $\mathbf{V}$ is the orthogonal matrix of eigenvectors and $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues λ_j. The whitening transformation matrix $\mathbf{W}$ is constructed by scaling the eigenvectors by the inverse square root of the eigenvalues

$$\mathbf{W} = \mathbf{\Lambda}^{-1/2}\mathbf{V}^\top. \tag{2.3}$$

Applying this transformation yields the whitened data $\widetilde{\mathbf{X}} = \mathbf{W}\mathbf{X}$, where the covariance is the identity matrix $\mathbf{I}_d$. After whitening, if we define the unmixing operation on the whitened data as $\widetilde{\mathbf{Z}} = \mathbf{U}\widetilde{\mathbf{X}}$, the covariance of the recovered components is

$$\mathrm{Cov}(\widetilde{\mathbf{Z}}) = \mathbf{U}\mathrm{Cov}(\widetilde{\mathbf{X}})\mathbf{U}^\top = \mathbf{U}\mathbf{I}_d\mathbf{U}^\top = \mathbf{U}\mathbf{U}^\top. \tag{2.4}$$

Because the independent components must have unit variance, we require $\mathbf{U}\mathbf{U}^\top = \mathbf{I}_d$. This reveals that the unmixing matrix $\mathbf{U}$ must be perfectly orthogonal, thus reducing the ICA optimization to a search over the Stiefel manifold, a space of orthogonal matrices.

2.1.4 Statistical and Thermodynamic Intuitions

The identifiability limitation in Theorem 2.1 leads to the translation of the challenge of finding independent components into a distinct optimization objective: maximizing non-Gaussianity. The theoretical justification for this approach is rooted in the Central Limit Theorem.

Theorem 2.2 (Central Limit Theorem Intuition). *Let $s_1, s_2, \ldots, s_d$ be independent random variables. Under general regularity conditions, the distribution of their normalized sum tends toward a Gaussian distribution as d increases (Billingsley, 1995).*

While Theorem 2.2 formally applies to limits of sums, its consequence in ICA is that any observed linear mixture of independent, non-Gaussian sources will inherently appear more Gaussian than the underlying sources themselves. (Hyvärinen et al., 2001, Chapter 1). Therefore, by extracting the unmixed components that are the least Gaussian, the algorithm effectively isolates the original independent signals.

The use of non-Gaussianity as a proxy for independence is further understood by an information theory and statistical mechanics perspective. For any continuous random variable with a fixed finite variance, it is a foundational mathematical result that the Gaussian distribution uniquely possesses the maximum possible Shannon entropy (Cover & Thomas, 2006, Chapter 8). In thermodynamics, isolated physical systems naturally evolve toward states of maximum entropy. Viewed through this interdisciplinary lens, the Central Limit Theorem serves as the statistical manifestation of this physical principle: the mathematical process of linearly mixing independent sources inherently increases the overall entropy of the system, inevitably driving the joint distribution toward the high-entropy Gaussian state.

2.2 An Information Geometry Perspective on ICA

2.2.1 Product and Gaussian Manifolds

Information geometry provides a framework for quantifying the relationship between independence and non-Gaussianity by analyzing probability distributions as points on geometric manifolds. Following the perspective introduced by Cardoso (2022), ICA can be understood as orthogonally projecting the empirical distribution onto two distinct subspaces. The first is the Product manifold $\mathcal{P}$, which contains all distributions where the marginal components are strictly independent. The second is the Gaussian manifold $\mathcal{G}$, which encompasses all multivariate Gaussian distributions with zero mean and a fixed covariance structure. We dive into this perspective with the knowledge that standard decorrelation merely ensures that the covariance is diagonal, but full ICA seeks to find a linear transform that minimizes the divergence between the empirical data and its projection onto the Product manifold (Cardoso, 2022).

2.2.2 Kullback-Leibler Pythagorean Identities

Distances between these distributions are measured using the Kullback-Leibler (KL) divergence (Cover & Thomas, 2006). For any random vector Y with a joint density P_Y, the mutual information $\mathscr{I}(Y)$ measures the distance to the closest

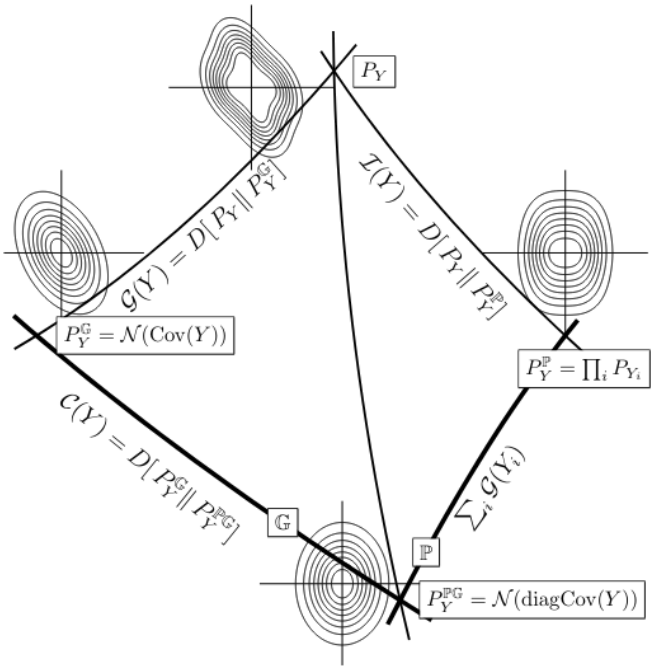

Figure 2.1: *The Pythagorean geometry of Independent Component Analysis. The empirical joint distribution P_Y is orthogonally projected onto the Product manifold $\mathcal{P}$ (yielding the independent marginals P_Y^P) and the Gaussian manifold $\mathcal{G}$ (yielding the best Gaussian fit $\mathcal{N}(\mathrm{Cov}\, Y)$). The intersection of these approximations yields the fully factorized Gaussian target $\mathcal{N}(\mathrm{diag}\,\mathrm{Cov}\, Y)$. The Kullback-Leibler divergences between these nodes geometrically represent distinct statistical properties: Mutual Information $\mathscr{I}(Y)$, Joint Non-Gaussianity $G(Y)$, and Correlation $C(Y)$. Figure adapted from Cardoso (2022).*

independent product distribution, denoted $P_Y^P = \prod_i P_{Y_i}$. This expands integrally as:

$$\mathscr{I}(Y) = D_{\mathrm{KL}}\left(P_Y \parallel P_Y^P\right) = \int P_Y(y) \log \frac{P_Y(y)}{\prod_i P_{Y_i}(y_i)}\, dy. \tag{2.5}$$

A Pythagorean identity exists on the Product manifold. For any candidate product distribution $Q = \prod_i q_i$, the divergence splits precisely into the mutual information and the marginal divergences. Expanding the logarithm reveals this additive structure:

$$\begin{aligned} D_{\mathrm{KL}}\left(P_Y \| Q\right) &= \int P_Y(y) \log \left(\frac{P_Y(y)}{P_Y^P(y)} \frac{P_Y^P(y)}{Q(y)} \right) dy \\ &= \int P_Y(y) \log \frac{P_Y(y)}{P_Y^P(y)}\, dy + \sum_i \int P_{Y_i}(y_i) \log \frac{P_{Y_i}(y_i)}{q_i(y_i)}\, dy \\ &= \mathscr{I}(Y) + \sum_i D_{\mathrm{KL}}\left(P_{Y_i} \| q_i\right). \end{aligned} \tag{2.6}$$

A similar Pythagorean identity holds for the Gaussian manifold $\mathcal{G}$. Let $\mathcal{N}(\Sigma)$ denote an arbitrary zero-mean multivariate Gaussian distribution parameterized

by a covariance matrix Σ. Furthermore, let $\mathcal{N}(\operatorname{Cov} Y)$ denote the specific Gaussian distribution that shares the exact covariance structure of the empirical data Y. Geometrically, $\mathcal{N}(\operatorname{Cov} Y)$ represents the orthogonal projection of P_Y onto the Gaussian manifold, yielding the best possible Gaussian approximation of the data.

The KL divergence from the empirical distribution P_Y to the arbitrary Gaussian target $\mathcal{N}(\Sigma)$ decomposes into the divergence to this best Gaussian fit and the divergence between the two Gaussian models:

$$D_{\mathrm{KL}}\big(P_Y \| \mathcal{N}(\Sigma)\big) = D_{\mathrm{KL}}\big(P_Y \| \mathcal{N}(\operatorname{Cov} Y)\big) + D_{\mathrm{KL}}\big(\mathcal{N}(\operatorname{Cov} Y) \| \mathcal{N}(\Sigma)\big). \tag{2.7}$$

2.2.3 Independence, Correlation, and Non-Gaussianity

These geometric identities can be combined to relate independence directly to non-Gaussianity. We define the non-Gaussianity of the joint variable Y as $G(Y) = D_{\mathrm{KL}}(P_Y \,\|\, \mathcal{N}(\operatorname{Cov} Y))$, and the scalar measure of correlation as $C(Y) = D_{\mathrm{KL}}(\mathcal{N}(\operatorname{Cov} Y) \,\|\, \mathcal{N}(\operatorname{diag} \operatorname{Cov} Y))$.

To establish the fundamental balance between these quantities, we evaluate the KL divergence from the empirical distribution P_Y to a fully factorized Gaussian target $Q = \mathcal{N}(\operatorname{diag} \operatorname{Cov} Y)$. We can compute this divergence via two distinct geometric paths shown in Figure 2.1.

Following the Pythagorean identity on the Product manifold (Equation 2.6):

$$D_{\mathrm{KL}}\big(P_Y \| \mathcal{N}(\operatorname{diag} \operatorname{Cov} Y)\big) = \mathscr{I}(Y) + \sum_i G(Y_i). \tag{2.8}$$

Following the Pythagorean identity on the Gaussian manifold (Equation 2.7):

$$D_{\mathrm{KL}}\big(P_Y \| \mathcal{N}(\operatorname{diag} \operatorname{Cov} Y)\big) = G(Y) + C(Y). \tag{2.9}$$

Equating these two paths yields Cardoso's unified theorem (Cardoso, 2022):

$$\mathscr{I}(Y) + \sum_i G(Y_i) = G(Y) + C(Y). \tag{2.10}$$

To demonstrate how maximizing marginal non-Gaussianities minimizes mutual information, we analyze the terms under the constraints of the ICA pipeline. Crucially, the joint non-Gaussianity $G(Y)$ is invariant under any invertible linear transformation (Cardoso, 2022). Furthermore, once the data has been whitened, the covariance matrix is forced to be the identity matrix. This implies that $\operatorname{Cov}(Y) = \operatorname{diag} \operatorname{Cov}(Y)$, which strictly forces the correlation term to zero ($C(Y) = 0$).

Applying these constraints, the relationship rearranges into a direct subtraction:

$$\mathscr{I}(Y) + \sum_i G(Y_i) = G(Y) + 0$$
$$\mathscr{I}(Y) = G(Y) - \sum_i G(Y_i). \tag{2.11}$$

Because $G(Y)$ remains constant under orthogonal rotation, minimizing the mutual information $\mathscr{I}(Y)$, finding independent components, is mathematically equivalent to maximizing the sum of the marginal non-Gaussianities $\sum_i G(Y_i)$. This shows that centering and whitening constrain the data to a subspace where true statistical independence can be achieved solely by maximizing non-Gaussianity.

2.3 Optimal Transport Theory

2.3.1 The Monge Problem: Push-Forwards and Pull-Backs

While information geometry utilizes Kullback-Leibler divergence, optimal transport provides an alternative metric space parameterized by the physical geometry of probability mass. First formulated by Gaspard Monge (1781), the problem was originally conceptualized as finding the most cost-effective way to move a distribution of mass.

Mathematically, this is framed as finding a transport map T that associates points from a continuous source measure μ to a target measure ν. We express this spatial modification of probability mass using the *push-forward* operator $T_{\#}$. A measure ν is the push-forward of μ by a map T, denoted $\nu = T_{\#}\mu$, if for any measurable set B in the target space:

$$\nu(B) = \mu(T^{-1}(B)). \tag{2.12}$$

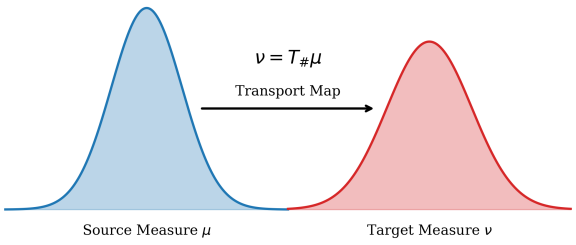

Figure 2.2: *The Monge transport map T pushes forward the probability mass from the source measure μ to construct the target measure ν, strictly conserving total mass during the transformation.*

As illustrated in Figure 2.2, the push-forward operator acts on measures, effectively pushing mass forward along the map T. It is conceptually distinct from the *pull-back* operator $T^{\#}$, which acts on functions. For a continuous, bounded test function $g : \mathcal{Y} \to \mathbb{R}$, the pull-back evaluates the function via the composition mapping ($T^{\#}g = g \circ T$). Because push-forward and pull-back are adjoint operators, the Monge problem seeks to minimize a transportation cost function $c(x, y)$ subject to the strict constraint that $T_{\#}\mu = \nu$:

$$\min_{T} \left\{ \int_{\mathcal{X}} c(x, T(x)) d\mu(x) \quad : \quad T_{\#}\mu = \nu \right\}. \tag{2.13}$$

2.3.2 The Kantorovich Relaxation

The fundamental limitation of the Monge problem is that the map T must be strictly deterministic; mass from a single source location cannot be split to multiple targets. Consequently, if the target distribution has more discrete points than the source distribution, a valid Monge map may not even exist.

To resolve this degeneracy, Kantorovich (1942) relaxed the deterministic requirement, proposing instead a probabilistic transport that permits mass splitting. The Kantorovich formulation replaces the deterministic map T with a joint probability distribution, or coupling, $\pi \in \Pi(\mu, \nu)$, which dictates exactly how fractions of mass are distributed from the source to the target.

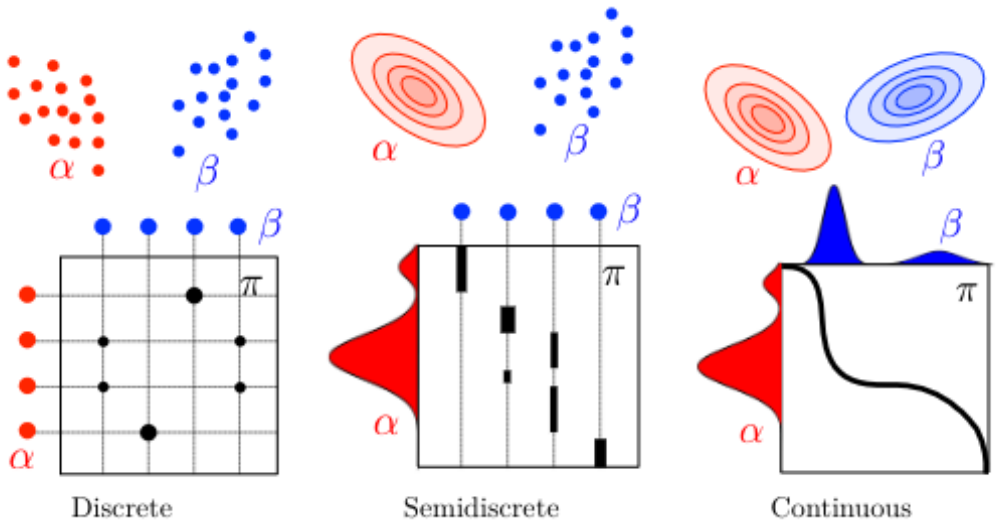

Figure 2.3: *A schematic view of the input measures and optimal couplings across the three primary scenarios of Kantorovich Optimal Transport: Discrete, Semi-discrete, and Continuous. The top row depicts the marginal distributions, while the bottom row visualizes the structural nature of the corresponding coupling spaces. Figure adapted from Peyré and Cuturi (2019).*

As illustrated in Figure 2.3, this relaxation is universally applicable across three

fundamental statistical scenarios. In the discrete setting depicted in the first column, the coupling manifests as a transportation matrix mapping distinct point masses. In the semi-discrete setting (second column), mass from a continuous density is optimally partitioned into specific discrete target regions. In the fully continuous setting (third column), the coupling forms a joint density over the continuous supports. By allowing fractional transport in all domains, Kantorovich transformed the non-convex combinatorial Monge problem into a relatively tractable linear optimization problem.

2.3.3 The Wasserstein Distances (W_1 and W_2)

The minimal cost required to transform one probability distribution into another is formalized as the Wasserstein distance. For continuous measures in arbitrary dimensions d, utilizing a Euclidean distance cost function $c(x, y) = \|x - y\|^p$, the general W_p distance is defined over the space of optimal couplings $\Pi(\mu, \nu)$ as:

$$W_p(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p \, d\pi(x, y) \right)^{1/p}. \tag{2.14}$$

While the integral definition above correctly formalizes optimal transport in high dimensions, direct analytical computation of $\pi(x, y)$ is highly restrictive for arbitrary multivariable spaces. The concept of sorting data to construct a cumulative distribution function (CDF) is strictly a one-dimensional phenomenon, as there is no natural, universal ordering of coordinates in $\mathbb{R}^d$ ($d > 1$).

However, in the specific one-dimensional setting, the geometry of optimal transport drastically simplifies (Villani, 2003, Chapter 2). Because the data can be unambiguously sorted, the W_1 distance analytically collapses to the integral of the absolute difference between the univariate cumulative distribution functions F and G:

$$W_1(\mu, \nu) = \int_{-\infty}^{\infty} |F(x) - G(x)| \, dx. \tag{2.15}$$

Similarly, the squared Wasserstein distance ($p = 2$), which heavily penalizes local geometric distortions, takes on a highly efficient analytical form in one dimension. It is most conveniently expressed using the quantile functions, which are the exact inverses of the cumulative distributions (F^{-1} and G^{-1}):

$$W_2^2(\mu, \nu) = \int_0^1 |F^{-1}(t) - G^{-1}(t)|^2 \, dt. \tag{2.16}$$

By circumventing the need to construct multi-dimensional couplings, the 1D W_2^2 metric provides an extremely computationally efficient mechanism for measur-

ing structural divergence. When evaluating empirical linear projections against standard normal noise, the 1D squared Wasserstein distance serves as the primary optimization function for the framework utilized in this thesis.

Chapter 3

The Optimal Transport ICA (OT-ICA) Framework

3.1 Optimal Transport Distance as a Contrast for ICA

As established in the preceding chapter, the process of centering and whitening restricts the search for independent components to the manifold of orthogonal matrices. Furthermore, Cardoso's unified theorem (Cardoso, 2022) demonstrates that within this whitened space, the mutual information of the projections is minimized precisely when the sum of their marginal non-Gaussianities is maximized. Traditional Independent Component Analysis algorithms approach this optimization by utilizing proxy contrast functions, such as kurtosis or negentropy approximations. While computationally inexpensive, these proxies evaluate specific statistical moments rather than the full probability density.

This thesis formulates ICA directly as an optimal transport problem, utilizing the squared Wasserstein distance, W_2^2, as a direct measure of non-Gaussianity. We formalize this framework as follows:

Definition 3.1 The OT-ICA Optimization Objective. *Let* $\mathbf{X}$ *be the whitened observed mixture, and let* $\mathbf{u} \in S_n$ *be a candidate projection vector on the unit sphere. Let the empirical distribution of the projected data be denoted by* $\mathbb{P}_{\mathbf{u}^\top \mathbf{X}}$*, and let the standard normal distribution be defined as* $\Gamma \equiv \mathcal{N}(0, 1)$*. The Optimal Transport ICA (OT-ICA) objective is to find the orthogonal rotation matrix whose column vectors* $\mathbf{u}$ *maximize the transport cost to* Γ*:*

$$\hat{\mathbf{u}} = \arg\max_{\mathbf{u} \in S_n} W_2^2\big(\mathbb{P}_{\mathbf{u}^\top \mathbf{X}}, \Gamma\big). \tag{3.1}$$

By framing the problem in this manner, we leverage the optimal transport metric's ability to evaluate geometric deviations from Gaussianity, providing a sensitive contrast function.

3.2 Theoretical Motivation: Bounding Mixture Non-Gaussianity

To mathematically justify the use of W_2^2 as an objective function for ICA, we demonstrate that maximizing this distance recovers the true independent sources. By the Central Limit Theorem, a mixture of non-Gaussian sources is more Gaussian than the constituent sources themselves. In optimal transport terms, this implies that the Wasserstein distance between a mixture and a Gaussian distribution is bounded by the distances of its independent components.

3.2.1 The W_2 Upper Bound

Theorem 3.1 W_2 **Upper Bound.** *Let* $\mathbf{Z} = (Z_1, \ldots, Z_d)^\top$ *be a vector of mutually independent latent sources, and let* $\mathbf{w}$ *be a weight vector on the unit sphere* ($\sum_{i=1}^d w_i^2 = 1$)*. The squared Wasserstein distance between the linear mixture* $\mathbf{w} \cdot \mathbf{Z}$ *and the standard normal distribution* Γ *is bounded by the weighted sum of the source distances to* Γ*:*

$$W_2^2(\mathbf{w} \cdot \mathbf{Z}, \Gamma) \leq \sum_{i=1}^{d} w_i^2 W_2^2(Z_i, \Gamma). \tag{3.2}$$

Proof. We construct a specific joint probability distribution, or coupling, using a common source of randomness. Let $\mathbf{N} = (N_1, \ldots, N_d)^\top$ be a vector of d independent standard normal variables, $\mathbf{N} \sim \mathcal{N}(0, \mathbf{I}_d)$, assumed to be statistically independent of the sources $\mathbf{Z}$. For each independent latent source Z_i, there exists an optimal transport map T_i that pushes forward the standard normal distribution to the source distribution, denoted as $(T_i)_{\#}\Gamma = Z_i$.

We construct a random variable X representing the mixture using the transport maps evaluated on the common noise vector:

$$X = \sum_{i=1}^{d} w_i T_i(N_i). \tag{3.3}$$

Because $T_i(N_i)$ follows the distribution of Z_i, X models the weighted mixture $\mathbf{w} \cdot \mathbf{Z}$. To compare this mixture to a Gaussian distribution, we construct a target variable

Y using the same underlying noise vector $\mathbf{N}$ and weights $\mathbf{w}$:

$$Y = \sum_{i=1}^{d} w_i N_i. \tag{3.4}$$

Given that the weights sum to unity and the components N_i are independent standard normals, the resulting variable Y follows the standard normal distribution, $Y \sim \Gamma$. The pair (X, Y) creates a valid coupling between the mixture distribution and the Gaussian target.

The squared Wasserstein distance is the infimum of the expected squared cost over all valid couplings. Therefore, the optimal distance is less than or equal to the cost of our constructed coupling (X, Y), yielding the inequality:

$$W_2^2(\mathbf{w} \cdot \mathbf{Z}, \Gamma) \leq \mathbb{E}\left[|X - Y|^2\right]. \tag{3.5}$$

Expanding the squared difference yields:

$$\begin{aligned} |X - Y|^2 &= \left| \sum_{i=1}^{d} w_i T_i(N_i) - \sum_{i=1}^{d} w_i N_i \right|^2 \\ &= \left(\sum_{i=1}^{d} w_i \big(T_i(N_i) - N_i\big) \right)^2. \end{aligned} \tag{3.6}$$

Squaring the summation produces diagonal terms ($i = j$) and cross terms ($i \neq j$):

$$|X - Y|^2 = \sum_{i=1}^{d} w_i^2 \big(T_i(N_i) - N_i\big)^2 + \sum_{i \neq j} w_i w_j \big(T_i(N_i) - N_i\big)\big(T_j(N_j) - N_j\big). \tag{3.7}$$

Taking the expectation simplifies the expression. Because the underlying noise variables N_i and N_j are independent, and the functions evaluate to a mean of zero, all cross terms vanish. We are left with the expectation of the diagonal terms:

$$\mathbb{E}\left[|X - Y|^2\right] = \sum_{i=1}^{d} w_i^2 \mathbb{E}\left[\big(T_i(N_i) - N_i\big)^2\right]. \tag{3.8}$$

Recognizing that $\mathbb{E}[(T_i(N_i) - N_i)^2]$ is the definition of the squared Wasserstein distance between the individual source Z_i and Γ, we arrive at the bound:

$$W_2^2(\mathbf{w} \cdot \mathbf{Z}, \Gamma) \leq \sum_{i=1}^{d} w_i^2 W_2^2(Z_i, \Gamma). \quad \blacksquare \tag{3.9}$$

This property demonstrates that mixing independent sources mathematically

reduces the Wasserstein distance to Gaussianity. Maximizing the left-hand side forces the weight vector $\mathbf{w}$ to collapse toward a canonical axis, isolating a single independent source.

3.2.2 Comonotonicity and the Proof of Strict Inequality

Optimizing ICA requires establishing that mixtures are strictly less non-Gaussian than the original sources. We determine the conditions under which the upper bound becomes a strict inequality.

Theorem 3.2 Strict W_2 Inequality for Non-Gaussian Sources. *Assuming at most one source Z_i is Gaussian, and the source distributions possess smooth and strictly positive densities, the upper bound derived in Theorem 3.1 is a strict inequality for any valid mixture where multiple weights w_i are non-zero:*

$$W_2^2(\mathbf{w} \cdot \mathbf{Z}, \Gamma) < \sum_{i=1}^{d} w_i^2 W_2^2(Z_i, \Gamma). \tag{3.10}$$

Proof. In a one-dimensional setting, the W_2 distance equals the expected squared difference of a specific coupling if and only if the random variables are comonotonic. By Brenier's Theorem (Brenier, 1991), this implies there exists a strictly increasing, deterministic function mapping one variable to the other. For the constructed mixture variable X and the Gaussian target Y to be comonotonic with respect to the shared noise vector $\mathbf{N} \in \mathbb{R}^d$, their gradients must be parallel at every point in the probability space. This requires the existence of a scalar function $\lambda(\mathbf{N})$ such that:

$$\nabla X(\mathbf{N}) = \lambda(\mathbf{N}) \nabla Y(\mathbf{N}). \tag{3.11}$$

Assuming the source distributions possess smooth and strictly positive densities, Caffarelli's regularity theory (Caffarelli, 1992) guarantees that the optimal transport maps T_i are continuously differentiable. We compute the gradients of our constructed variables with respect to $\mathbf{N}$. The gradient of the Gaussian target is the weight vector:

$$Y(\mathbf{N}) = \sum_{i=1}^{d} w_i N_i \implies \nabla Y = \mathbf{w}. \tag{3.12}$$

The gradient of the source mixture leverages the derivatives of the individual

transport maps:

$$X(\mathbf{N}) = \sum_{i=1}^{d} w_i T_i(N_i) \implies \nabla X = \begin{bmatrix} w_1 T_1'(N_1) \\ w_2 T_2'(N_2) \\ \vdots \\ w_d T_d'(N_d) \end{bmatrix}. \tag{3.13}$$

Substituting these gradients into the comonotonicity condition yields the system of equations:

$$\nabla X = \lambda(\mathbf{N})\mathbf{w}. \tag{3.14}$$

To determine if this condition holds, we differentiate both the Left-Hand Side (LHS) and the Right-Hand Side (RHS) with respect to $\mathbf{N}$ to compute their Hessian matrices.

On the LHS, because X separates into independent terms $w_i T_i(N_i)$, the cross-derivatives $\frac{\partial^2 X}{\partial N_i \partial N_j}$ are zero for all $i \neq j$. This yields a diagonal matrix containing the second derivatives of the transport maps. On the RHS, we differentiate the product $\lambda(\mathbf{N})\mathbf{w}$. Applying the product rule yields the outer product of $\mathbf{w}$ and $\nabla\lambda$, forming a Rank-1 matrix. Equating the matrices:

$$\underbrace{\begin{bmatrix} w_1 T_1''(N_1) & 0 & \cdots & 0 \\ 0 & w_2 T_2''(N_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_d T_d''(N_d) \end{bmatrix}}_{\text{Hessian of } X \text{ (Diagonal)}} = \underbrace{\begin{bmatrix} w_1 \frac{\partial \lambda}{\partial N_1} & w_1 \frac{\partial \lambda}{\partial N_2} & \cdots & w_1 \frac{\partial \lambda}{\partial N_d} \\ w_2 \frac{\partial \lambda}{\partial N_1} & w_2 \frac{\partial \lambda}{\partial N_2} & \cdots & w_2 \frac{\partial \lambda}{\partial N_d} \\ \vdots & \vdots & \ddots & \vdots \\ w_d \frac{\partial \lambda}{\partial N_1} & w_d \frac{\partial \lambda}{\partial N_2} & \cdots & w_d \frac{\partial \lambda}{\partial N_d} \end{bmatrix}}_{\text{Outer Product } \mathbf{w}(\nabla\lambda)^\top \text{ (Rank-1)}} \tag{3.15}$$

Examining any off-diagonal entry located at row i and column j ($i \neq j$), the diagonal LHS matrix dictates that this entry must be exactly zero. The RHS Rank-1 matrix evaluates this entry as $w_i \frac{\partial \lambda}{\partial N_j}$, establishing the requirement:

$$w_i \frac{\partial \lambda}{\partial N_j} = 0 \quad \text{for all } i \neq j. \tag{3.16}$$

Because we assume a mixture where multiple component weights are non-zero ($w_i \neq 0$), it follows that $\frac{\partial \lambda}{\partial N_j} = 0$. Since this holds across all dimensions, the gradient of the scaling function is zero ($\nabla\lambda = \mathbf{0}$). Consequently, $\lambda(\mathbf{N})$ is a global scalar constant, λ.

Equating the diagonal elements implies $T_i'(N_i) = \lambda$. Integrating this relationship indicates the optimal transport maps are linear functions of the form:

$$T_i(x) = \lambda x + c. \tag{3.17}$$

Applying a purely linear transformation to Gaussian noise N_i produces another Gaussian distribution. The identifiability assumption of ICA dictates that the original latent sources Z_i are non-Gaussian, which requires their optimal transport mappings from a Gaussian base to be non-linear. Therefore, the linear map requirement contradicts the non-Gaussianity of the sources. Because the condition for equality cannot be met, the relationship is a strict inequality.

$$W_2^2(\mathbf{w} \cdot \mathbf{Z}, \Gamma) < \sum_{i=1}^{d} w_i^2 W_2^2(Z_i, \Gamma). \quad \blacksquare \tag{3.18}$$

3.3 Empirical Landscape: W_1 versus W_2^2

Having established the theoretical foundations of the W_2^2 metric, we evaluate its practical suitability for manifold optimization. In the context of ICA, gradient ascent algorithms must climb the loss landscape to locate the maxima of the distance metric to isolate the independent sources.

While the W_1 distance (Earth Mover's Distance) is robust to outliers, its reliance on an L_1 absolute penalty introduces geometric challenges when applied to finite empirical samples. We demonstrate why the squared L_2 penalty of the W_2^2 metric is preferable for gradient-based convergence.

3.3.1 Empirical Gradients and Derivative Volatility

To observe this behavior, we generated a synthetic bivariate mixture of Student-t distributions ($\nu = 4$) and evaluated both the raw distances and their first-order numerical gradients across a continuous sweep of rotation angles $\theta \in [0, \pi]$. Because the unmixing matrix restricts the search to orthogonal vectors, the two true independent components are geometrically separated by exactly $\pi/2$ radians (90°).

The empirical results, visualized in Figure 3.1, illustrate the behaviors of the two metrics during continuous gradient ascent. In optimal transport, because the W_1 cost scales linearly with distance, the resulting optimization landscape lacks the convexity required to smooth finite sample noise.

The consequence of this lack of curvature is evident in the bottom gradient profile. Because W_1 relies on piecewise linear sorting mechanics without a squaring penalty, its empirical derivative fluctuates frequently as the projection angle rotates. Consequently, first-order gradient ascent solvers will encounter a jagged derivative surface, which hinders convergence to the true peak. This instability is exacerbated in higher dimensions, where the W_1 derivative surface becomes even more jagged and difficult to traverse. Furthermore, this volatility limits the utility of second-order Quasi-Newton approximations (such as L-BFGS), as a stable Hessian cannot

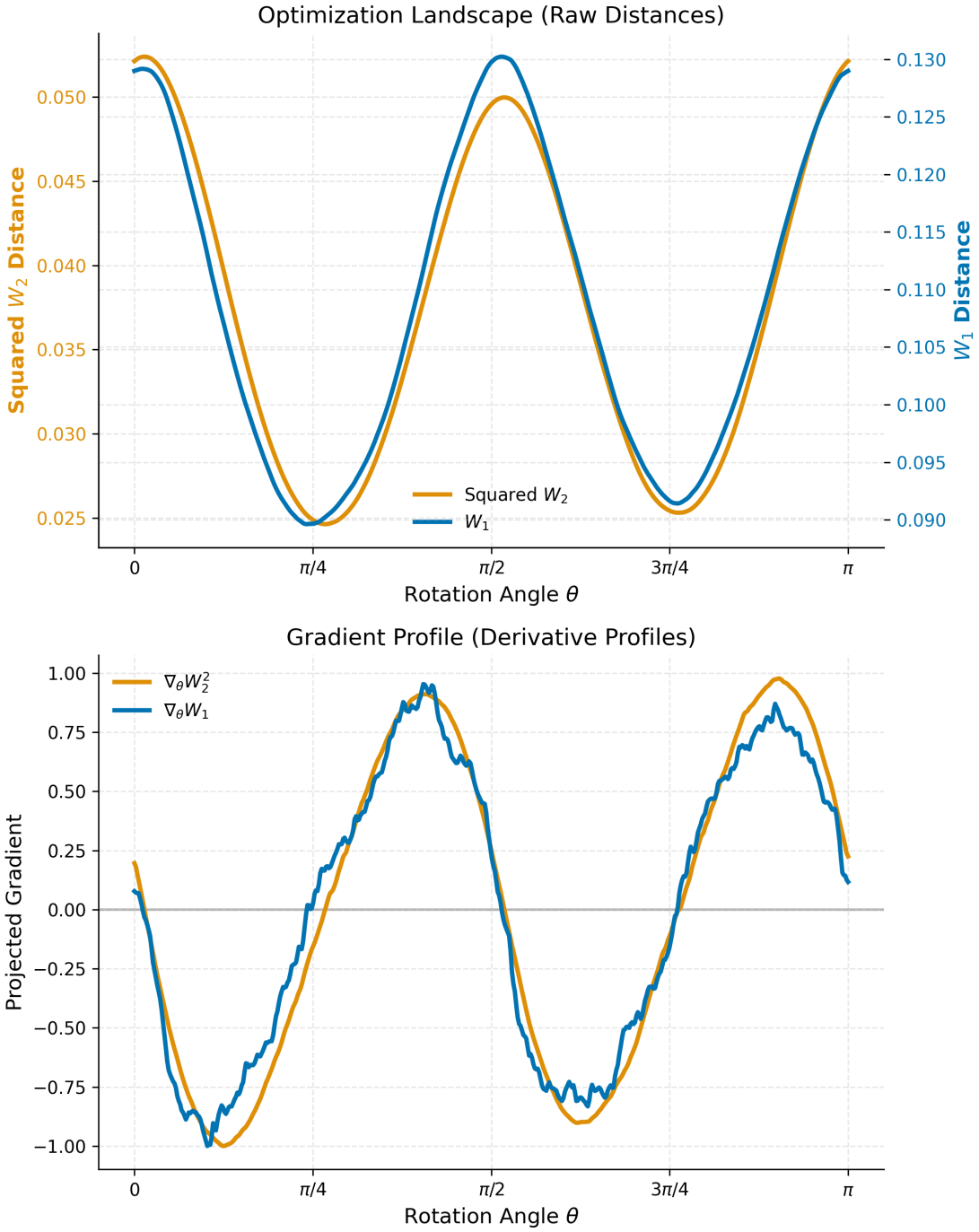

Figure 3.1: *The empirical optimization landscapes (top) and their normalized first derivatives (bottom). While both metrics peak near the true independent components ($\theta \approx 0, \pi/2, \pi$), their gradient behaviors differ. The W_2^2 gradient smoothly approaches zero, whereas the W_1 gradient exhibits jagged volatility on finite samples.*

be reliably estimated from such a jagged and difficult to traverse gradient.

In contrast, the squared penalty of the W_2^2 metric acts as a natural smoother by penalizing local geometric distortions. As seen in the gradient profiles, the W_2^2 derivative behaves continuously, decelerating predictably as it approaches the optimum. This retention of curvature provides a reliable signal for gradient-based solvers, allowing the OT-ICA framework to converge consistently.

Chapter 4

Algorithmic Enhancements

4.1 The Baseline OT-ICA Algorithm

Having established the theoretical optimality of the squared Wasserstein distance (W_2^2) for Independent Component Analysis, the challenge transitions to computationally minimizing this cost function. The baseline OT-ICA algorithm proceeds by first centering and whitening the observed mixture $\mathbf{X} \in \mathbb{R}^{d \times N}$, yielding the whitened data $\widetilde{\mathbf{X}}$. The algorithm then initializes an orthogonal unmixing matrix $\mathbf{W} \in \mathbb{R}^{d \times d}$. In the forward pass, the data is projected onto the candidate components, $\mathbf{Y} = \mathbf{W}\widetilde{\mathbf{X}}$. For each row (component) of $\mathbf{Y}$, the empirical cumulative distribution function is constructed by sorting the N observations. The W_2^2 distance is then computed by matching these sorted empirical quantiles to the corresponding quantiles of a standard normal distribution.

While theoretically sound, the standard implementation of this gradient descent loop faces computational and geometric limitations. The exact optimal transport matching requires sorting at every iteration, presenting a computational bottleneck. Furthermore, standard gradient updates do not inherently respect the strict orthogonality constraint of the mixing matrix, and discrete approximations of the target Gaussian introduce gradient noise. To make OT-ICA computationally competitive with existing methods, several specific algorithmic and geometric enhancements are required.

4.2 Computational Bottlenecks and Solutions

4.2.1 Exact Analytical Gaussian Targets

One source of instability in the baseline algorithm is the discrete approximation of the target Gaussian distribution. Traditionally, the target quantiles are approximated by evaluating the inverse cumulative distribution function at evenly spaced

intervals, such as $q_i = \Phi^{-1}((i - 0.5)/N)$. However, representing a continuous, infinite-tailed Gaussian distribution with a finite set of discrete points introduces approximation noise. When the algorithm computes the gradient of the W_2^2 distance, this noise translates into a non-smooth optimization landscape, hindering convergence as the components approach independence.

To eliminate this approximation noise, we replace the point-sampled targets with an exact analytical formulation. We compute the expected value of a standard normal variable within each discrete quantile bin.

Definition 4.1 Exact Analytical Gaussian Target. *Let the uniform probability bin edges for N samples be defined as $p_i = \frac{i}{N}$ for $i = 0, \ldots, N$, with corresponding Gaussian domain boundaries $z_i = \Phi^{-1}(p_i)$, where Φ is the standard normal cumulative distribution function. The exact analytical target value T_i for the i-th sorted sample is the conditional expectation within that interval:*

$$T_i = N \int_{z_{i-1}}^{z_i} x\phi(x)\, dx \tag{4.1}$$

where $\phi(x)$ is the standard normal probability density function. Evaluating this integral yields the closed-form target:

$$T_i = N\big(\phi(z_{i-1}) - \phi(z_i)\big). \tag{4.2}$$

By computing this analytical target once during initialization, the algorithm evaluates the W_2^2 distance against an exact discrete representation of a Gaussian, ensuring smooth gradients near the optimum.

4.2.2 Batched Vectorization for Restarts

Unlike Principal Component Analysis, which yields a unique global solution via singular value decomposition, the ICA optimization landscape is non-convex and characterized by local optima. Consequently, identifying the true independent components requires multiple random initializations (restarts) to explore the space.

Because evaluating the W_2^2 objective requires an $\mathcal{O}(N \log N)$ sorting operation for each candidate component, executing these random restarts sequentially via standard iterative loops is computationally expensive. To resolve this, we utilize a batched vectorization strategy. Rather than maintaining a single weight matrix, we aggregate B random restarts into a single high-dimensional tensor $\mathbf{W}_{\text{batch}} \in \mathbb{R}^{B \times d \times d}$.

This formulation allows the projection $\mathbf{Y} = \mathbf{W}_{\text{batch}}\widetilde{\mathbf{X}}$ to be computed simul-

taneously for all B trajectories. The elements of the resulting batch tensor can be sorted in parallel. The pre-computed exact analytical target vector $\mathbf{T}$ is subsequently broadcast-subtracted across the entire batch to compute the distances and gradients in a single step.

4.3 Optimization on the Stiefel Manifold

4.3.1 The Riemannian Gradient

Because the input data is whitened, the unmixing matrix $\mathbf{W}$ must remain orthogonal, satisfying $\mathbf{W}\mathbf{W}^\top = \mathbf{I}$. The geometric space of all such orthogonal matrices forms the Stiefel manifold, denoted $\mathcal{S}$. If we compute the standard Euclidean gradient of the squared Wasserstein distance, $\mathbf{G} = \nabla_{\mathbf{W}} W_2^2$, this vector points into the unconstrained ambient space $\mathbb{R}^{d\times d}$. Updating the matrix using this Euclidean gradient would violate the orthogonality constraint.

To navigate the geometry of the Stiefel manifold, we project the Euclidean gradient onto the tangent space of the manifold at the current point $\mathbf{W}$ (Figure 4.1).

Definition 4.2 Riemannian Gradient on the Stiefel Manifold. *Given a Euclidean gradient* $\mathbf{G}$*, the Riemannian gradient* $\nabla_{\mathcal{S}}\mathbf{W}$ *that projects* $\mathbf{G}$ *onto the*

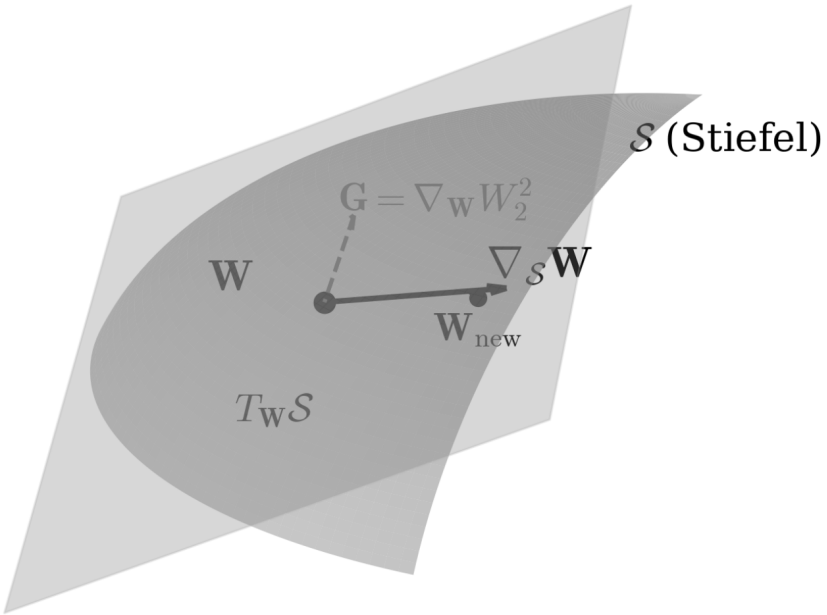

Figure 4.1: *Geometric sketch of optimization on the Stiefel manifold* $\mathcal{S}$*. The Euclidean gradient* $\mathbf{G}$ *is projected onto the tangent space* $T_{\mathbf{W}}\mathcal{S}$ *to yield the Riemannian gradient* $\nabla_{\mathcal{S}}\mathbf{W}$*. Following a step along the tangent plane, a retraction maps the estimate back to the orthogonal manifold surface at* $\mathbf{W}_{new}$*.*

tangent space of the Stiefel manifold at $\mathbf{W}$ *is given by (Absil et al., 2008, Chapter 3):*

$$\nabla_{\mathcal{S}}\mathbf{W} = \mathbf{G} - \frac{1}{2}\big(\mathbf{G}\mathbf{W}^\top + \mathbf{W}\mathbf{G}^\top\big)\mathbf{W}. \tag{4.3}$$

Geometrically, the term $(\mathbf{G}\mathbf{W}^\top + \mathbf{W}\mathbf{G}^\top)$ measures the symmetric violation of orthogonality induced by the Euclidean gradient. By subtracting this violation, we ensure that the update step aligns with the curvature of the manifold, preserving the decorrelation of the components.

4.3.2 Retraction via Symmetric Decorrelation

Although the Riemannian gradient ensures the update step is tangential to the Stiefel manifold, moving along this tangent plane for a finite step size η causes the new matrix $\mathbf{W}_{\text{step}} = \mathbf{W} - \eta\nabla_{\mathcal{S}}\mathbf{W}$ to depart from the curved manifold surface. To restore exact orthogonality, the matrix must be mapped back onto the manifold via a retraction (Absil et al., 2008, Chapter 4).

Definition 4.3 Symmetric Decorrelation Retraction. *Let* $\mathbf{W}_{step}$ *be the unmixing matrix after a tangent space update. The retraction mapping the matrix back to the orthogonal Stiefel manifold surface is:*

$$\mathbf{W}_{\text{new}} = (\mathbf{W}_{\text{step}}\mathbf{W}_{\text{step}}^\top)^{-1/2}\mathbf{W}_{\text{step}}. \tag{4.4}$$

This operation computes the overlap covariance of the stepped matrix and applies a uniform adjustment to all vectors simultaneously. This ensures that no single independent component is prioritized during the retraction process, maintaining the stability of the batched optimization.

4.4 Total Complexity and Statistical Efficiency

Having detailed the sorting-based objective and the Riemannian optimization steps, we formalize the total computational cost of the OT-ICA framework. For a dataset of dimension d with N samples, K random restarts, and T optimization iterations, the per-iteration complexity is dominated by:

$$\mathcal{O}(K \cdot (d \cdot N \log N + d^3)) \tag{4.5}$$

The $N \log N$ term stems from the sorting of projections, while the d^3 term represents the cost of the symmetric decorrelation (matrix inverse square root) used for retraction on the Stiefel manifold. For typical ICA applications where the number

of samples significantly exceeds the dimensionality ($N \gg d$), this d^3 cost is computationally negligible. It becomes a bottleneck only in extreme high-dimensional regimes, a scenario that is independently precluded by the statistical limits of the framework.

Beyond temporal complexity, we must account for the statistical sample complexity. While the 1D projection approach maintains a convergence rate of $\mathcal{O}(N^{-1/2})$, the global identification of the mixing matrix is implicitly subject to the curse of dimensionality. Theoretical results in optimal transport establish that the empirical Wasserstein distance in d dimensions converges at a rate of $\mathcal{O}(N^{-1/d})$ (Fournier & Guillin, 2015). In practice, as d increases, the number of samples N required to accurately resolve the maxima of non-Gaussianity on the manifold surface grows. This necessitates a corresponding increase in the number of parallel restarts K, creating a trade-off between statistical resolution and computational feasibility.

4.5 Navigating Discrete Optimization Landscapes

While the OT-ICA framework demonstrates convergence on continuous distributions, data often contains discrete signals, such as binary states or count-based events. Applying continuous optimal transport directly to discrete mixtures introduces geometric properties that complicate the optimization landscape.

4.5.1 The Discontinuous Landscape of Discrete CDFs

The exact analytical Gaussian target developed in Section 4.2 relies on matching the empirical cumulative distribution function (CDF) of the projected data. For a continuous variable, this empirical CDF approximates a strictly increasing curve. However, for highly discrete data, any one-dimensional projection creates a stepwise CDF.

This discreteness disrupts the continuous geometry of the W_2 metric due to the mechanics of sorting. Evaluating the Wasserstein distance requires sorting the projected points. As the optimization algorithm rotates the unmixing matrix $\mathbf{W}$, dense clusters of discrete data points shift relative to one another, causing the sorting order to change abruptly. Every change in the sorting permutation causes a discontinuous shift in the gradient direction. Consequently, the theoretically smooth W_2 landscape results in a non-differentiable surface filled with local optima. The optimizer frequently converges to spurious peaks (often corresponding to diagonal projections of the discrete hypercube), leading to unmixing failures.

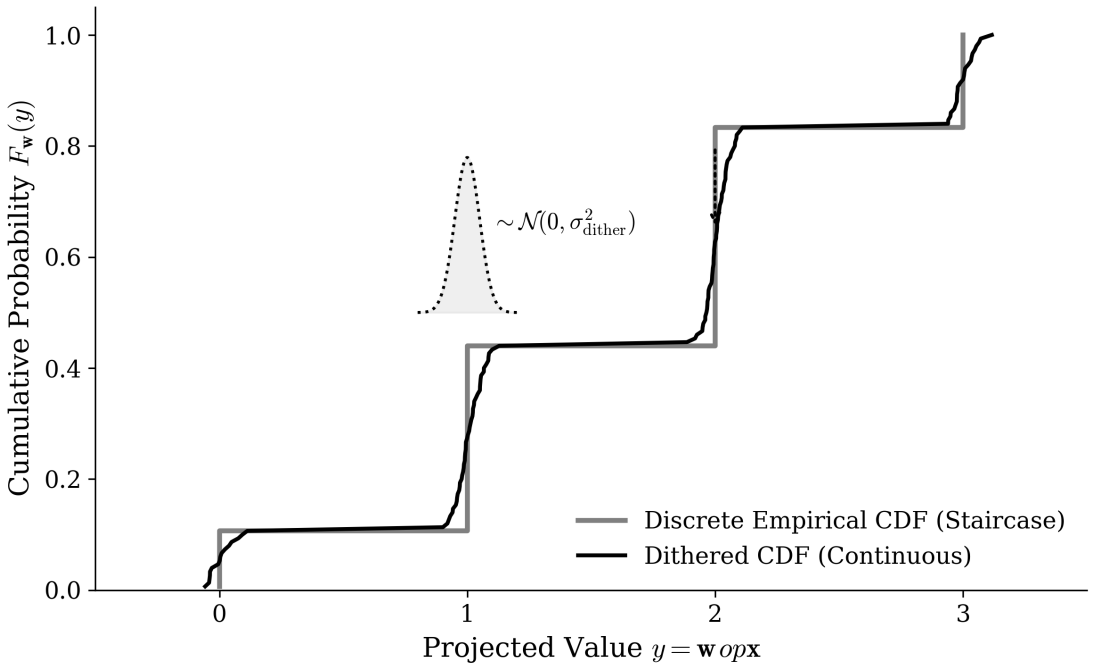

Figure 4.2: *The effect of algorithmic Gaussian Dithering. The non-differentiable staircase of a discrete empirical CDF (gray) is convolved with a narrow Gaussian kernel, yielding a strictly increasing, continuous curve (black) that restores stable Wasserstein gradients.*

4.5.2 Continuous Smoothing via Gaussian Dithering

To restore the smooth gradients of the Stiefel manifold without resorting to computationally intensive $\mathcal{O}(N^2)$ entropic regularization techniques (such as Sinkhorn distances), we implement a Gaussian dithering step.

Adapted from signal processing techniques used to decorrelate quantization error (Schuchman, 1964), we inject a small variance of continuous, zero-mean Gaussian noise ($\sigma_{\text{dither}} \approx 0.01$) into the projected components prior to the sorting operation in the forward pass. Mathematically, this operation convolves the discrete empirical distribution with a narrow Gaussian kernel (Figure 4.2).

This deforms the step-wise discrete CDF into a strictly increasing, continuous curve. By eliminating non-differentiable sorting ties, Gaussian dithering restores stable gradients while preserving the $\mathcal{O}(N \log N)$ computational efficiency of the 1D optimal transport formulation.

This dithering process does not constitute the injection of restrictive prior knowledge regarding the source distributions. Rather, it acts as an uninformed regularizer applied uniformly to the 1D projections $\mathbf{w}^\top \mathbf{X}$. For latent sources that are already continuous, the addition of small variance ($\sigma_{\text{dither}} \approx 0.01$) is statistically negligible and preserves the underlying Wasserstein geometry. For discrete sources, it prevents sorting failures.

From a theoretical perspective, this dithering process represents a form of kernel density regularization. Mathematically, adding Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2_{\text{dither}})$ to the projection Y is equivalent to convolving the empirical measure μ_N with a Gaussian kernel ϕ_σ:

$$\hat{\mu}_\sigma = \mu_N * \phi_\sigma \tag{4.6}$$

This operation transforms discrete measures into a smooth density. By Brenier's Theorem (Brenier, 1991) and Caffarelli's regularity (Caffarelli, 1992) results, this ensures that the resulting optimal transport map is unique and continuously differentiable. Furthermore, this approach is analogous to entropic regularization used in Sinkhorn distances, where a blur parameter is introduced to make the Wasserstein distance differentiable. By utilizing 1D dithering, we approximate the smoothness of Sinkhorn divergence while retaining the computational efficiency of sorted exact transport.

4.5.3 Escaping Local Minima with Stochastic Mini-Batching

Even with a smoothed CDF, the global optimization landscape of a discrete mixture remains non-convex. Evaluating the exact W_2 distance using the full dataset guarantees deterministic convergence to the nearest local maximum. In non-convex environments with discrete local minima, this deterministic behavior prevents global convergence.

To address this, we utilize stochastic mini-batching. At each optimization step, the algorithm evaluates the Wasserstein gradients using a randomly sampled subset of the data (e.g., $N_{\text{batch}} = 512$ or 1024). This stochastic subsampling introduces gradient noise, acting similarly to simulated annealing. A geometric local minimum that exists for the full dataset may have a different gradient profile for a random subset. This stochastic momentum allows the optimizer to escape spurious local minima. The combination of Gaussian dithering to smooth the discrete steps and stochastic mini-batching to navigate the non-convex topology enables OT-ICA to separate mixed environments.

4.6 OT-ICA Algorithm Overview

The theoretical components and algorithmic enhancements are synthesized into the final OT-ICA training procedure. The pseudo-code for extracting independent non-Gaussian components from a mixed dataset using stochastic Riemannian optimization is detailed below (Algorithm 1).

Algorithm 1 The Optimal Transport ICA (OT-ICA) Algorithm

Require: Observed mixture $\mathbf{X} \in \mathbb{R}^{d \times N}$, iterations K, learning rate η, batch size B, dither noise σ_{dither}

Ensure: Estimated unmixing matrix $\mathbf{W}_{\text{final}} \in \mathbb{R}^{d \times d}$

Phase 1: Preprocessing & Target Generation

Center the data: $\mathbf{X} \leftarrow \mathbf{X} - \mathbb{E}[\mathbf{X}]$

Whiten the data: $\widetilde{\mathbf{X}} \leftarrow (\mathbf{X}\mathbf{X}^\top)^{-1/2}\mathbf{X}$ ▷ Yields unit variance, uncorrelated data

Compute analytical target $\mathbf{T} \in \mathbb{R}^B$ via exact Gaussian CDF integration

Phase 2: Initialization

Initialize $\mathbf{W} \in \mathbb{R}^{d \times d}$ randomly from $\mathcal{N}(0, 1)$

Apply symmetric decorrelation:

$\mathbf{W} \leftarrow (\mathbf{W}\mathbf{W}^\top)^{-1/2}\mathbf{W}$ ▷ Initialize on Stiefel manifold

Phase 3: Stochastic Riemannian Optimization

for $k = 1$ to K **do**

  Sample stochastic mini-batch $\widetilde{\mathbf{X}}_{\text{batch}} \in \mathbb{R}^{d \times B}$ from $\widetilde{\mathbf{X}}$

  Project data onto candidates: $\mathbf{Y} \leftarrow \mathbf{W}\widetilde{\mathbf{X}}_{\text{batch}}$

  Apply Gaussian Dithering: $\mathbf{Y}_{\text{dither}} \leftarrow \mathbf{Y} + \mathcal{N}(0, \sigma^2_{\text{dither}})$

  Sort each row of $\mathbf{Y}_{\text{dither}}$ ascending to yield empirical quantiles $\mathbf{Y}_{\text{sorted}}$

  Compute Wasserstein objective (maximize non-Gaussianity):

  $\mathcal{L} = \frac{1}{B}\sum_{i=1}^{d} ||\mathbf{Y}^{(i)}_{\text{sorted}} - \mathbf{T}||_2^2$

  Backpropagate to compute Euclidean gradient: $\mathbf{G} = \nabla_{\mathbf{W}}\mathcal{L}$

  Project to tangent space (Riemannian Gradient):

  $\nabla_{\mathcal{S}}\mathbf{W} = \mathbf{G} - \frac{1}{2}\big(\mathbf{G}\mathbf{W}^\top + \mathbf{W}\mathbf{G}^\top\big)\mathbf{W}$

  Update matrix along tangent plane:

  $\mathbf{W}_{\text{step}} = \mathbf{W} + \eta\nabla_{\mathcal{S}}\mathbf{W}$ ▷ Gradient Ascent

  Retract to Stiefel manifold via symmetric decorrelation:

  $\mathbf{W} \leftarrow (\mathbf{W}_{\text{step}}\mathbf{W}_{\text{step}}^\top)^{-1/2}\mathbf{W}_{\text{step}}$

  Optional: Apply learning rate decay $\eta \leftarrow \eta \cdot 0.99$

end for

Phase 4: Finalization

$\mathbf{W}_{\text{final}} \leftarrow \mathbf{W}(\mathbf{X}\mathbf{X}^\top)^{-1/2}$ ▷ Map back to original data space

return $\mathbf{W}_{\text{final}}$

4.7 Limitations of Fixed-Point Algorithms for OT-ICA

A widely used algorithm for ICA, FastICA (Hyvärinen et al., 2001, Chapter 8), achieves efficient convergence by employing a Newton (second-order) fixed-point

step. FastICA utilizes this approach by maximizing proxy contrast functions that are static and easily differentiable. A natural question is whether a similar fixed-point iteration can be derived for the W_2^2 objective.

Theorem 4.1 Intractability of Exact Newton Updates. *For an empirical distribution mapped to a continuous target via sorting, computing an exact, closed-form Hessian $\mathbf{H} = \nabla^2_{\mathbf{w}} W_2^2$ is analytically intractable without introducing external density estimation mechanisms. The Map Derivative $\nabla_{\mathbf{w}} T_{\mathbf{w}}$ inherently relies upon the rate of change of cumulative probability mass, which strictly requires the continuous Probability Density Function (PDF) evaluated at the quantile boundaries.*

Proof. To implement a Newton step, the exact Hessian matrix of the squared Wasserstein distance, $\mathbf{H} = \nabla^2_{\mathbf{w}} W_2^2$, is required. We first derive the first derivative (the gradient) of the objective with respect to the weight vector $\mathbf{w}$. Utilizing the envelope theorem for optimal transport (Santambrogio, 2015, Section 7.2), the variation of the optimal map evaluates to zero within the cost function, allowing us to differentiate the squared cost directly:

$$\begin{aligned}
\nabla_{\mathbf{w}} \left(\frac{1}{2} W_2^2 \right) &= \nabla_{\mathbf{w}} \left(\frac{1}{2} \mathbb{E}\Big[\big(\mathbf{w}^\top \mathbf{X} - T_{\mathbf{w}}(\mathbf{w}^\top \mathbf{X}) \big)^2 \Big] \right) \\
&= \mathbb{E}\Big[\big(\mathbf{w}^\top \mathbf{X} - T_{\mathbf{w}}(\mathbf{w}^\top \mathbf{X}) \big) \nabla_{\mathbf{w}} (\mathbf{w}^\top \mathbf{X}) \Big] \\
&= \mathbb{E}\Big[\mathbf{X} \big(\mathbf{w}^\top \mathbf{X} - T_{\mathbf{w}}(\mathbf{w}^\top \mathbf{X}) \big) \Big]. \qquad (4.7)
\end{aligned}$$

To compute the Hessian, we differentiate the optimal transport map $T_{\mathbf{w}}(\mathbf{w}^\top \mathbf{X})$ with respect to $\mathbf{w}$. Applying the total derivative yields two terms:

$$\nabla_{\mathbf{w}} [T_{\mathbf{w}}(\mathbf{w}^\top \mathbf{X})] = \underbrace{T'_{\mathbf{w}}(\mathbf{w}^\top \mathbf{X}) \mathbf{X}}_{\text{Argument Derivative}} + \underbrace{(\nabla_{\mathbf{w}} T_{\mathbf{w}})(\mathbf{w}^\top \mathbf{X})}_{\text{Map Derivative}}. \qquad (4.8)$$

By Caffarelli's regularity theory (Caffarelli, 1992), the transport map $T_{\mathbf{w}}$ is continuously differentiable, guaranteeing the existence of the Argument Derivative. The limitation lies within the Map Derivative.

In the 1D optimal transport setting, the transport map to a standard Gaussian relies on the empirical cumulative distribution function, $F_{\mathbf{w}}$, of the projected data $y = \mathbf{w}^\top \mathbf{X}$. Expanding the Map Derivative yields:

$$(\nabla_{\mathbf{w}} T_{\mathbf{w}})(y) = \nabla_{\mathbf{w}} \big[\Phi^{-1}(F_{\mathbf{w}}(y)) \big] = \frac{1}{\phi(\Phi^{-1}(F_{\mathbf{w}}(y)))} \nabla_{\mathbf{w}} F_{\mathbf{w}}(y). \qquad (4.9)$$

The term $\nabla_{\mathbf{w}} F_{\mathbf{w}}(y)$ represents the rate of change of the cumulative probability mass as the projection plane $\mathbf{w}$ is rotated. The derivative of a cumulative distribution

function's boundary strictly requires the underlying probability density function (PDF) evaluated at that boundary, rendering the closed-form calculation intractable without explicit density estimation. ■

This theoretical requirement negates the computational advantage of the OT-ICA formulation. The efficiency of 1D optimal transport stems from its reliance on sorting (empirical CDFs), circumventing continuous density estimation entirely. Requiring the PDF to compute the Hessian necessitates density approximation methods like Kernel Density Estimation (KDE) or Normalizing Flows. These methods introduce statistical noise, hyperparameter tuning, and significant computational overhead.

Because computing a true Newton-based fixed-point step for exact OT-ICA is analytically intractable, the chosen strategy for the OT-ICA framework utilizes Quasi-Newton methods, specifically the Limited-memory BFGS (L-BFGS) algorithm (Liu & Nocedal, 1989) adapted for the Stiefel manifold (Absil et al., 2008, Chapter 6). L-BFGS approximates the inverse Hessian curvature by observing the history of first-order Wasserstein gradients. This approach allows the algorithm to achieve convergence while preserving the sorting-based methodology of the optimal transport formulation.

Chapter 5

Experimental Evaluation and Applications

5.1 Evaluation Metrics And Methodology

5.1.1 The Amari Performance Index

To empirically evaluate the success of an Independent Component Analysis (ICA) algorithm, a quantitative metric is required to measure the distance between the estimated unmixing matrix $\mathbf{W}_{\text{est}}$ and the true mixing matrix $\mathbf{A}$. Because ICA is subject to inherent scaling and permutation ambiguities, we utilize the Amari Performance Index (Amari et al., 1996).

Definition 5.1 Amari Performance Index. *Let $\mathbf{P} = \mathbf{W}_{est}\mathbf{A}$ be the global transfer matrix. For a $d \times d$ matrix $\mathbf{P}$ with elements p_{ij}, the Amari error measures the structural divergence from a generalized permutation matrix:*

$$E(\mathbf{P}) = \frac{1}{2d} \sum_{i=1}^{d} \left(\sum_{j=1}^{d} \frac{|p_{ij}|}{\max_k |p_{ik}|} - 1 \right) + \frac{1}{2d} \sum_{j=1}^{d} \left(\sum_{i=1}^{d} \frac{|p_{ij}|}{\max_k |p_{kj}|} - 1 \right). \quad (5.1)$$

This formulation normalizes each row and column by its maximum absolute value. If $\mathbf{P}$ is a generalized permutation matrix, indicating successful source separation, the index evaluates to zero. The experimental evaluations employ specific heuristic thresholds to interpret the Amari error (Table 5.1).

5.1.2 Computational Resource Regimes

Because OT-ICA relies on first-order gradient descent over the non-convex Stiefel manifold, its capacity to find the global optimum depends on the computational

Amari Error (E)	Interpretation / Quality of Separation
$E < 0.1$	Excellent to near-perfect source separation.
$0.1 \leq E < 0.3$	Good separation; the structural shape of the underlying sources is highly recoverable.
$0.3 \leq E < 0.5$	Moderate separation; some structural details may be lost.
$E \geq 0.5$	Severe degradation; multiple sources remain significantly mixed.
$E \geq 1.0$	The estimated matrix is effectively a random orthogonal projection with no meaningful relation to the true sources.

Table 5.1: *Heuristic thresholds for interpreting the Amari Performance Index.*

resources allocated to search the space, particularly as dimensionality increases. Conversely, FastICA utilizes a Newton-based fixed-point iteration that converges rapidly under standard conditions. FastICA requires extended computational limits in specific edge cases where its structural assumptions are violated, causing the solver to oscillate.

To accommodate OT-ICA's scaling requirements, and to evaluate whether FastICA's non-convergence in certain cases is due to mathematical properties rather than standard optimization timeouts, high-dimensional tests in this chapter are evaluated under two computational regimes. The hyperparameter configurations establish specific boundaries for these regimes (Table 5.2).

Regime & Objective	OT-ICA Configuration	FastICA Configuration
Low Compute Baseline *Represents standard, efficient execution.*	**Restarts:** $\min(4d, 150)$ **Deflation Phase:** 150 iterations **Symmetric Phase:** 200 iterations	**Max Iterations:** 1,000 *Standard limit for fixed-point convergence.*
High Compute Regime *Evaluates structural convergence independent of timeouts.*	**Restarts:** $\min(15d, 600)$ **Deflation Phase:** 300 iterations **Symmetric Phase:** 500 iterations	**Max Iterations:** 5,000 *Extended to ensure non-convergence is structural.*

Table 5.2: *Summary of the computational regimes used to evaluate algorithmic scaling. These iteration limits and restart thresholds separate standard optimization timeouts from mathematical convergence issues.*

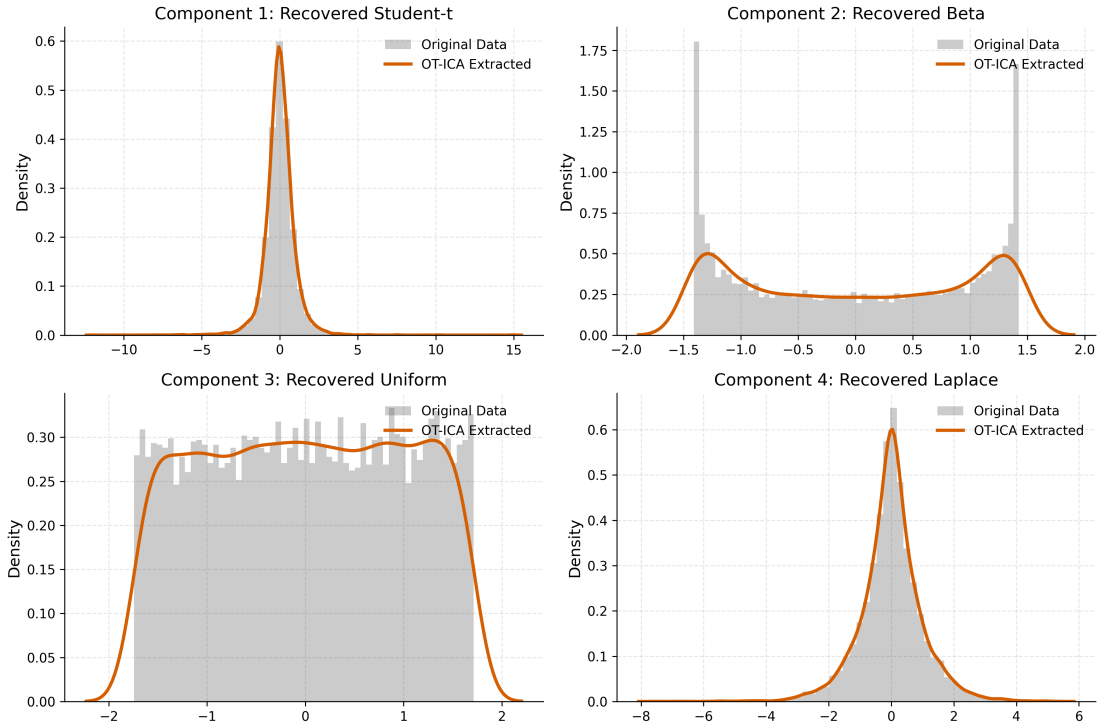

Figure 5.1: *Empirical probability distributions of the 4D baseline validation. The original generated sources are represented by gray histograms, overlaid with the sources extracted by the OT-ICA algorithm (orange curves). The framework recovers continuous non-Gaussian structures, navigating both super-Gaussian peaks and sub-Gaussian bimodality.*

5.2 OT-ICA Methodology Validation

We first verify that OT-ICA successfully recovers a mixture of non-Gaussian distributions in a standard, low-dimensional environment. We generated $N = 10,000$ samples from four continuous independent sources: Laplace (super-Gaussian), Uniform (sub-Gaussian), Student-t (continuous heavy-tailed), and a Beta distribution parameterized at $(0.5, 0.5)$ (bimodal arcsine distribution). These sources were linearly combined using a 4×4 mixing matrix $\mathbf{A}$.

The OT-ICA algorithm successfully extracted the four structural signals from the mixture (Figure 5.1). To quantify this separation numerically, we evaluate the true mixing matrix $\mathbf{A}$, the estimated unmixing matrix $\mathbf{W}_{\text{est}}$, and the global transfer matrix $\mathbf{P} = \mathbf{W}_{\text{est}}\mathbf{A}$. Because ICA is subject to permutation and sign ambiguities, comparing the estimated unmixing matrix to the true mixing matrix directly is challenging. The global transfer matrix $\mathbf{P}$ addresses this; if the algorithm inverted the mixing process, $\mathbf{P}$ approximates a generalized permutation matrix.

The global transfer matrix $\mathbf{P}$ forms a generalized permutation matrix (Table 5.3). This confirms that OT-ICA extracted the independent sources, yielding an Amari error below the 0.1 threshold. For comparison, the FastICA algorithm applied to the same 4D mixture achieved a comparable Amari error.

True Mixing Matrix (A)				Estimated Unmixing Matrix ($\mathbf{W}_{\text{est}}$)			
1.058	1.520	−0.249	1.012	0.154	−1.943	0.608	−0.456
0.042	−0.486	0.388	−0.523	−0.092	−0.284	−0.632	−1.977
−0.154	−1.572	−0.304	−1.234	0.146	0.719	1.097	1.691
−0.006	0.503	0.053	0.928	1.021	1.204	0.828	0.677

Global Transfer Matrix ($\mathbf{P} = \mathbf{W}_{\text{est}}\mathbf{A}$)

−0.010	−0.007	−1.000	−0.002
0.000	−0.003	0.000	−1.000
0.006	−1.000	−0.002	−0.012
1.000	0.006	−0.003	0.010

Amari Performance Index

OT-ICA Error: **0.0155** | FastICA Error: **0.0188**

Table 5.3: *The ground truth mixing matrix, estimated unmixing matrix, and resulting global transfer matrix for the OT-ICA baseline validation. The transfer matrix approximates a generalized permutation matrix, resulting in a comparable Amari error to FastICA.*

5.3 Computational & Temporal Scaling

To evaluate computational scaling in higher dimensions, we benchmarked OT-ICA against FastICA using a linear mixture of continuous super-Gaussian Laplacian sources. We evaluated both algorithms across increasing dimensions with a fixed sample size of $N = 10,000$. Because OT-ICA relies on gradient-based optimization over the Stiefel manifold, its capacity to find global optima depends on computational resources. We evaluated both algorithms under the Low Compute and High Compute regimes.

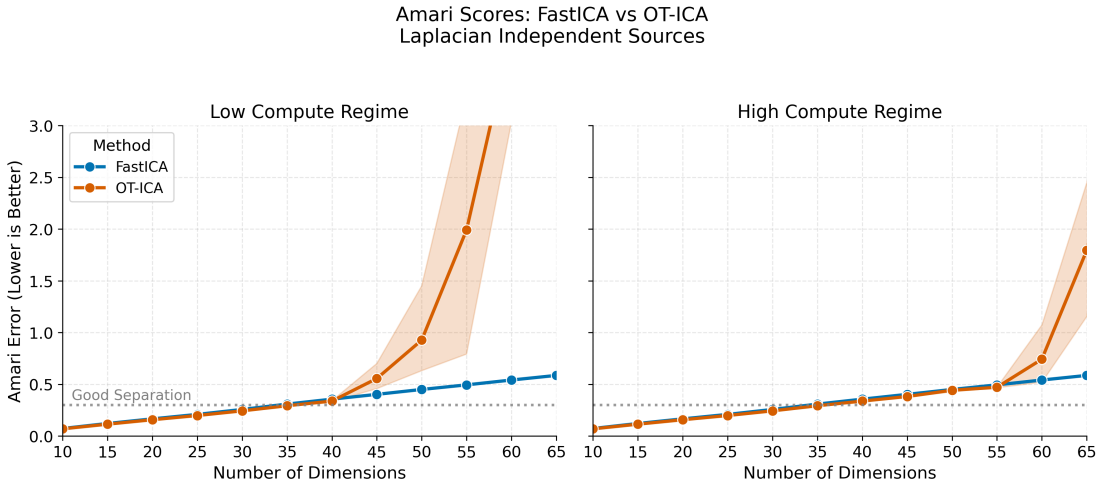

Figure 5.2: *Amari error comparison between FastICA and OT-ICA across dimensions for Laplacian sources. While OT-ICA degrades in the Low Compute regime due to the surface area of the Stiefel manifold, the High Compute regime achieves separation ($E < 0.3$), matching FastICA's capabilities.*

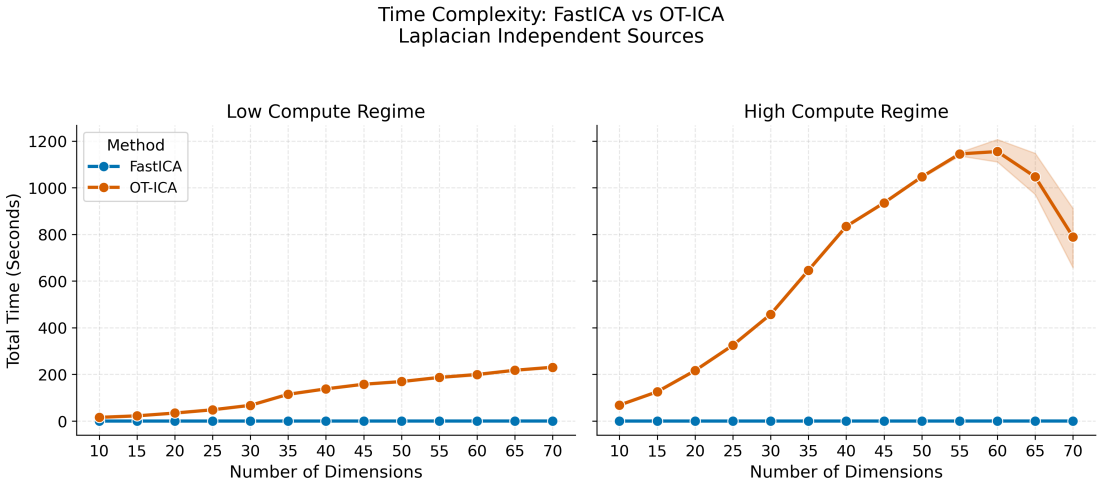

Figure 5.3: *Total execution time (in seconds) for FastICA and OT-ICA across dimensions. OT-ICA exhibits higher computational costs, particularly in the High Compute regime (right) where parallel restarts are required.*

FastICA performs well for lower dimensions, maintaining low error regardless of the compute regime (Figure 5.2). The Laplacian distribution aligns with FastICA's default *logcosh* contrast function. While OT-ICA scales similarly to FastICA, it requires the High Compute regime to navigate the higher-dimensional optimization landscape. As dimensionality approaches $d = 40$, the statistical resolution of the joint distribution for a sample size of $N = 10,000$ degrades. Both algorithms fail to maintain the $E < 0.3$ threshold, demonstrating a statistical limitation of the fixed sample size.

The computational cost for OT-ICA is significantly higher than FastICA (Figure 5.3). FastICA's Newton-based fixed-point iteration allows rapid convergence. In contrast, OT-ICA's first-order gradient descent and parallel restarts cause execution times to scale more steeply. This difference in algorithmic efficiency raises the question of whether a gradient-based framework like OT-ICA provides advantages over FastICA in specific settings.

While FastICA efficiently separates homogeneous mixtures of continuous distributions, empirical data is often heterogeneous. The following sections explore how specific structural properties of non-Gaussian sources affect FastICA's convergence.

5.4 Algorithmic Limitations of FastICA

We examine two specific mathematical conditions that affect FastICA's convergence on certain non-Gaussian independent components, resulting from its reliance on proxy contrast functions and Newton-based iteration.

5.4.1 The Zero Negentropy Condition

FastICA approximates negentropy $J(x)$ using a non-quadratic contrast function $G(\cdot)$, such as the *logcosh* function $G(x) = \frac{1}{a}\log\cosh(ax)$. The approximation is given by:

$$J(x) \propto \big(\mathbb{E}[G(x)] - \mathbb{E}[G(\nu)]\big)^2 \tag{5.2}$$

where ν is a standard Gaussian random variable. The algorithm seeks projections that maximize this squared difference.

If a latent independent source possesses a specific distribution such that its expected value under the contrast function equals that of a Gaussian ($\mathbb{E}[G(x)] = \mathbb{E}[G(\nu)]$), the approximated negentropy evaluates to zero. In this scenario, the objective function provides no gradient signal, and the algorithm does not extract the source.

5.4.2 Empirical Results: Zero Negentropy

To empirically demonstrate this condition, we constructed a symmetric Trimodal Gaussian Mixture parameterized by peak locations $\{-b, 0, b\}$ and variance σ^2. We solved for parameters such that the expected value of the *logcosh* function matched the theoretical baseline of a standard normal distribution. This distribution maintains unit variance, representing an independent non-Gaussian source that yields an approximated negentropy of zero.

We generated linear mixtures of this Trimodal Gaussian source across dimensions ($d \in [5, 25]$) with $N = 10,000$ samples. We separated the evaluation into the Low Compute and High Compute regimes.

The empirical results demonstrate that FastICA yields high Amari errors ($E > 0.5$) for this distribution (Figure 5.4). Providing FastICA with High Compute resources does not improve performance, confirming that the lack of convergence is related to the properties of the proxy objective function rather than iteration limits.

OT-ICA achieves source separation, bypassing parametric approximations by utilizing the empirical CDF via optimal transport. While OT-ICA shows performance degradation in the Low Compute regime for $d > 15$ due to the non-convexity of the Stiefel manifold, the High Compute regime restores accurate unmixing ($E < 0.3$) up to $d = 20$.

5.4.3 The Vanishing Curvature Condition

FastICA utilizes a Newton (second-order) fixed-point iteration step. The mathematical stability of this convergence relies on a condition governing the algorithmic

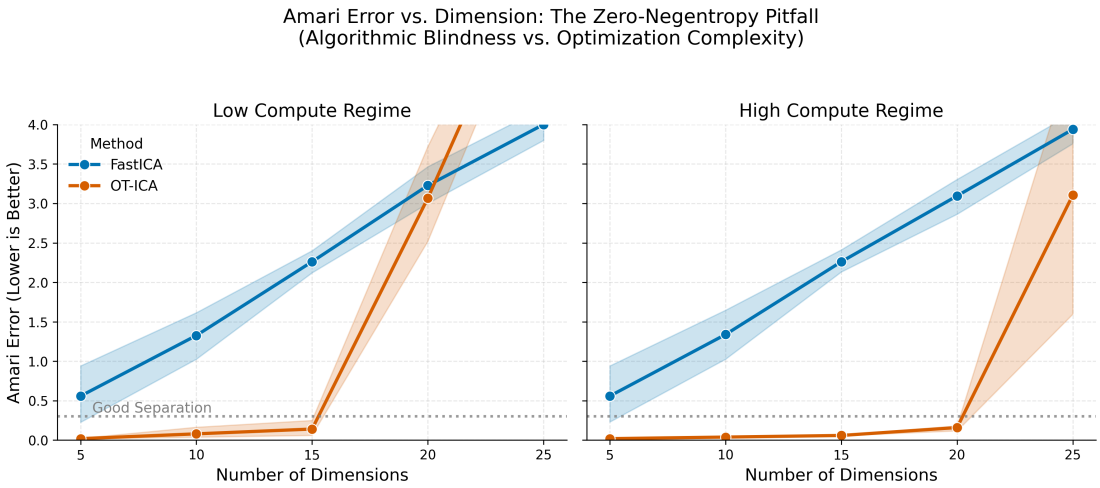

Figure 5.4: *Amari error comparison between FastICA and OT-ICA for the zero-negentropy Trimodal Gaussian distribution. FastICA lacks convergence ($E > 0.5$) across all regimes. OT-ICA unmixes the sources, though it requires the High Compute regime for higher dimensions.*

curvature of the optimization landscape, as formalized by Hyvärinen, Karhunen, and Oja (Hyvärinen et al., 2001, Chapter 8) (Appendix A.1).

For the fixed-point iteration to converge to a true independent component s_i, the expectation governing the second-order Taylor expansion of the iteration must not vanish. Defining $g(x) = G'(x)$ and $g'(x) = G''(x)$, where G is the contrast function, convergence requires:

$$\mathbb{E}\big[s_i g(s_i) - g'(s_i)\big] \neq 0 \tag{5.3}$$

For the standard *logcosh* contrast function (with $a = 1$), this equates to $g(x) = \tanh(x)$ and $g'(x) = 1 - \tanh^2(x)$.

If a non-Gaussian source distribution causes this expectation to evaluate to zero, the denominator of the FastICA update rule approaches zero. The fixed-point iteration fails to converge under these conditions.

5.4.4 Empirical Results: Vanishing Curvature

We engineered a continuous source distribution to test this curvature condition. We analyzed the target function:

$$h(x) = x\tanh(x) - 1 + \tanh^2(x). \tag{5.4}$$

We constructed a Trimodal Gaussian Mixture parameterized by symmetrically spaced peak locations $\{-b, 0, b\}$ and variance σ^2. Solving the integral equation $\int h(x)\text{PDF}(x)dx = 0$ subject to the unit variance constraint ($\mathbb{E}[X^2] = 1$) determines

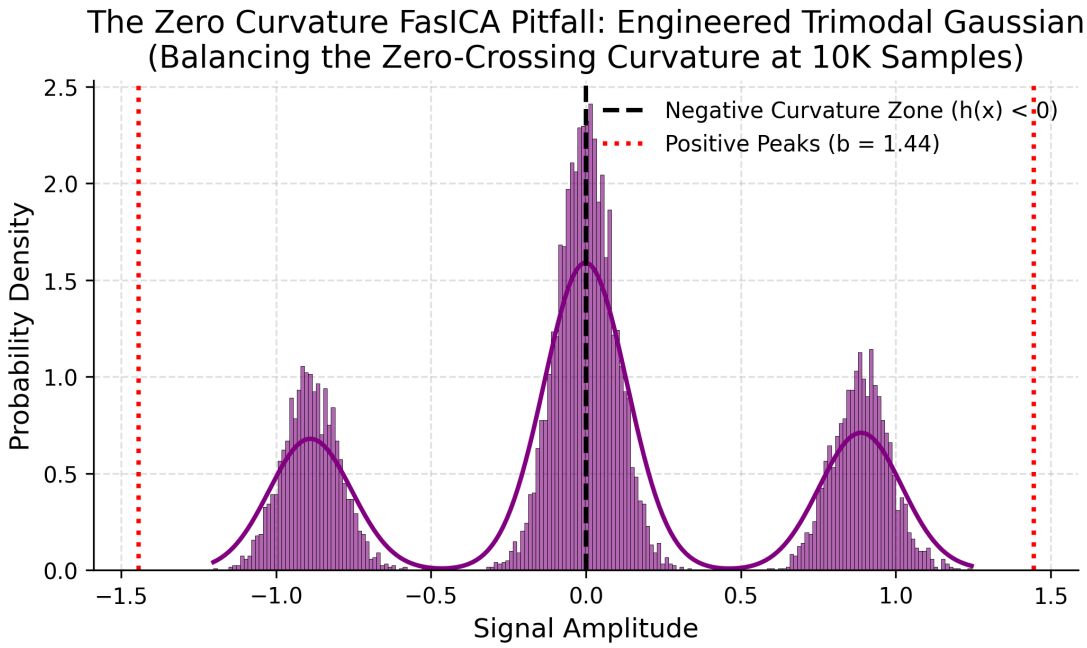

Figure 5.5: *The engineered Trimodal Gaussian Mixture. The central mass generates negative algorithmic curvature, while the side peaks generate positive algorithmic curvature. The parameters b and p are solved such that these regions cancel ($\int h(x)PDF(x)dx = 0$) while maintaining unit variance.*

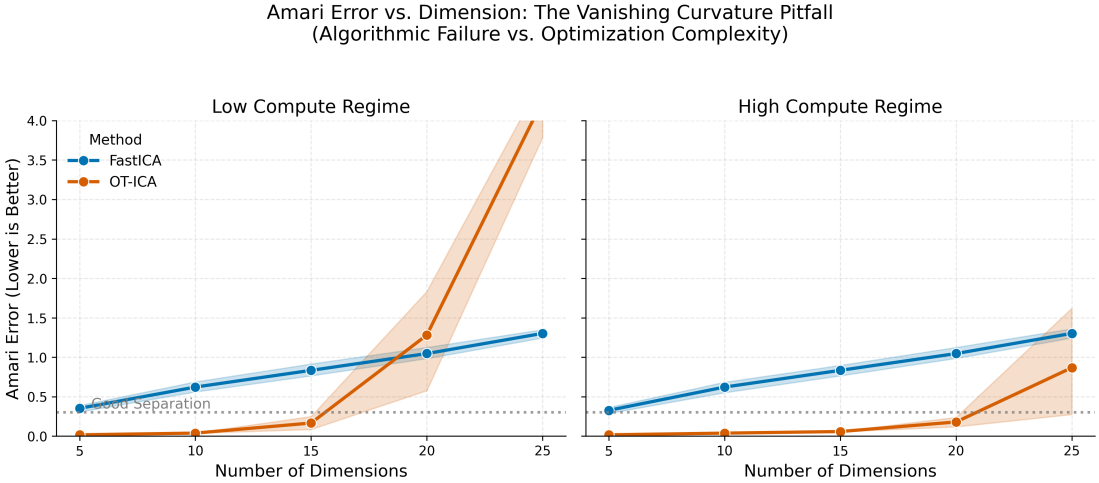

Figure 5.6: *Amari error comparison between FastICA and OT-ICA across dimensions for the engineered Trimodal distribution. FastICA lacks convergence ($E > 1.0$) across all regimes.*

parameters that balance the regions of negative and positive curvature.

This distribution is non-Gaussian, but its theoretical algorithmic curvature evaluates to zero (Figure 5.5). We evaluated FastICA and OT-ICA across dimensions ($d \in [5, 25]$) at $N = 10,000$ under both compute regimes.

FastICA returned unmixing matrices with high Amari errors ($E > 0.3$) across dimensions (Figure 5.6). The High Compute Regime did not improve FastICA's

performance, indicating the non-convergence is structural.

OT-ICA unmixed lower dimensions ($d < 20$) in the Low Compute regime and degraded as d increased. In the High Compute regime, OT-ICA maintained separation ($E < 0.3$) through 20 dimensions. This suggests OT-ICA's performance is limited by the expanding non-convex search space rather than local density approximations.

5.5 The Generalized Hybrid Mixture Stress Test

To evaluate OT-ICA and FastICA on heterogeneous signals, we tested a generalized hybrid mixture.

5.5.1 Experimental Setup

We generated linear mixtures at dimensions $d = 30$ and $d = 40$ with $N = 10,000$. One source was fixed as a standard Gaussian distribution, $\mathcal{N}(0, 1)$. The remaining $(d-1)$ independent sources were randomly drawn from eight distributions: Laplace, Bernoulli, Uniform, Student-t, Poisson, Binomial, Chi-squared, and Exponential.

FastICA was evaluated with a limit of $10,000$ iterations, while OT-ICA utilized its baseline parameters (Table 5.2).

5.5.2 Empirical Results: Hybrid Mixture Stress Test

OT-ICA maintained moderate to good separation ($E < 0.5$) across the dimensions for all but discrete-only mixtures (Figure 5.7).

FastICA exhibited instability in three of the five hybrid mixtures: full hybrid, zero Gaussian, and discrete-only. At $d = 30$, FastICA triggered non-convergence

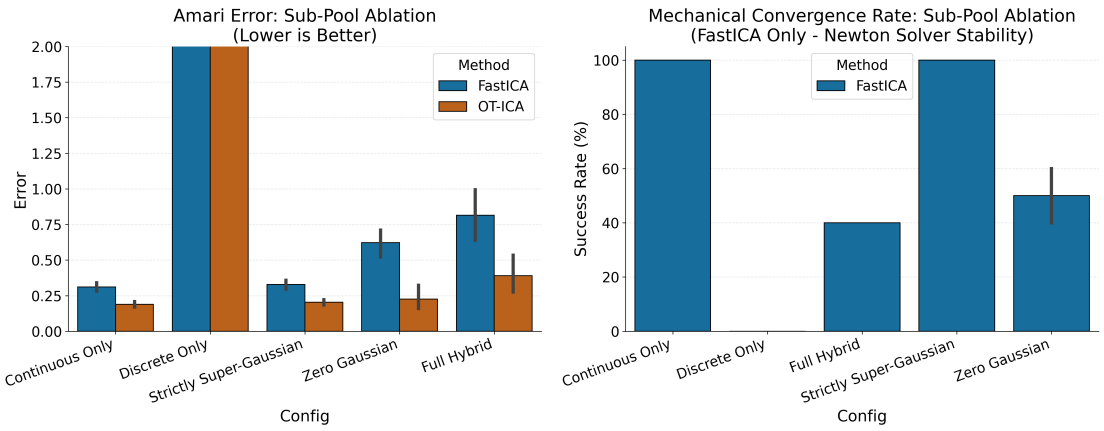

Figure 5.7: *Amari error comparison for the Generalized Hybrid Mixture. OT-ICA maintains stability across the tested dimensions.*

warnings. Providing an extended $10,000$ iteration limit resulted in Amari errors remaining high ($E > 0.5$). This indicates that FastICA's fixed-point update steps may not converge reliably when processing mixtures containing both super-Gaussian and sub-Gaussian signals simultaneously.

5.6 Discrete Only Mixtures: A Unique Challenge

We isolated specific discrete distributions from the discrete-only mixtures to observe their impact on optimization. We generated mixtures containing one continuous Gaussian source and $(d-1)$ sources of a single discrete type (Bernoulli, Poisson, or Binomial).

5.6.1 Empirical Results: Discrete Only Mixtures

FastICA and OT-ICA separated binary Bernoulli mixtures, but both algorithms reported high Amari errors ($E > 1.0$) for multi-step count-based geometries of standard Poisson ($\lambda = 3$) and Binomial ($n = 10$) distributions (Figure 5.8).

5.6.2 Optimization Properties of Discrete Data

To analyze this behavior, we calculated the squared Wasserstein distance (W_2^2) and the Logcosh proxy negentropy to measure the distance between standard normal noise and discrete distributions.

Adjusting the parameters (Poisson $\lambda = 0.5$, Binomial $n = 2$) increases the non-Gaussianity measured by W_2^2 (Table 5.4). The Logcosh proxy evaluates to lower relative values for these discrete structures compared to the continuous Laplace baseline.

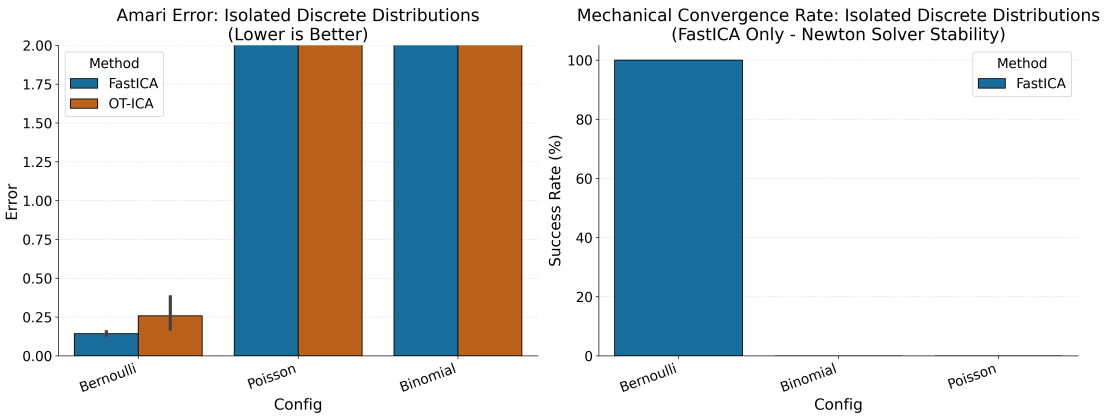

Figure 5.8: *Amari error for mixtures composed entirely of a single discrete distribution type. Both algorithms separate binary (Bernoulli) data but show higher error on count data (Poisson, Binomial).*

Distribution	W_2^2 Distance	Logcosh Proxy Negentropy
Laplace (Continuous Baseline)	0.0398	0.001214
Binomial (Standard, $n = 10$)	0.0344	0.000031
Poisson (Standard, $\lambda = 3.0$)	0.0513	0.000000
Binomial (Harsh, $n = 2$)	0.2022	0.000267
Poisson (Harsh, $\lambda = 0.5$)	0.3384	0.000085

Table 5.4: *Comparison of non-Gaussianity metrics across distributions.*

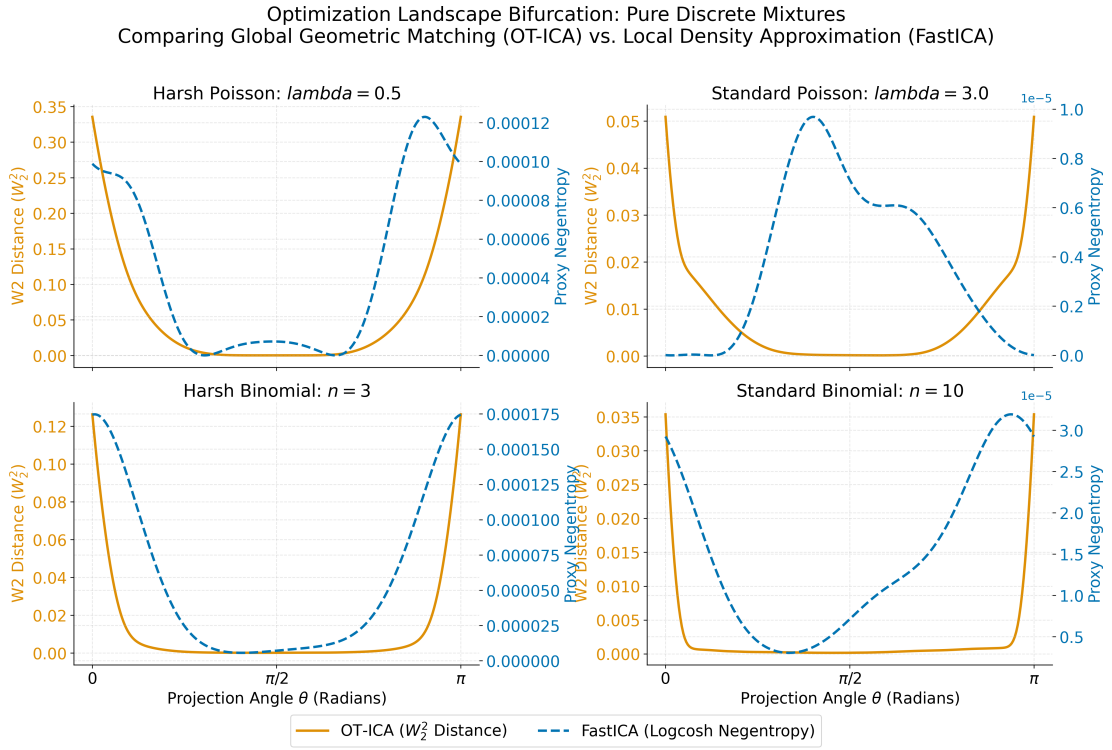

Figure 5.9: *The 1D optimization landscapes for discrete mixtures. The step-function geometry of discrete data affects the gradient properties of both the OT-ICA surface and the FastICA proxy approximation.*

The 1D optimization surfaces were mapped for both metrics (Figure 5.9). For FastICA (dashed blue line), the *logcosh* proxy landscape has low amplitude gradient changes (scale $1e-5$). For OT-ICA (solid orange line), the geometric matching reflects the step-function nature of the empirical CDF, creating flat plateaus where the gradient approaches zero. This non-smooth landscape can stall the L-BFGS solver prior to identifying the components.

5.6.3 Empirical Results: Harsh Discrete Environments

We evaluated the algorithms on the skewed parameterizations.

For the skewed Poisson mixture ($\lambda = 0.5$), FastICA did not converge (Figure 5.10). OT-ICA achieved separation ($E \approx 0.15$), as the skewed nature of the Poisson

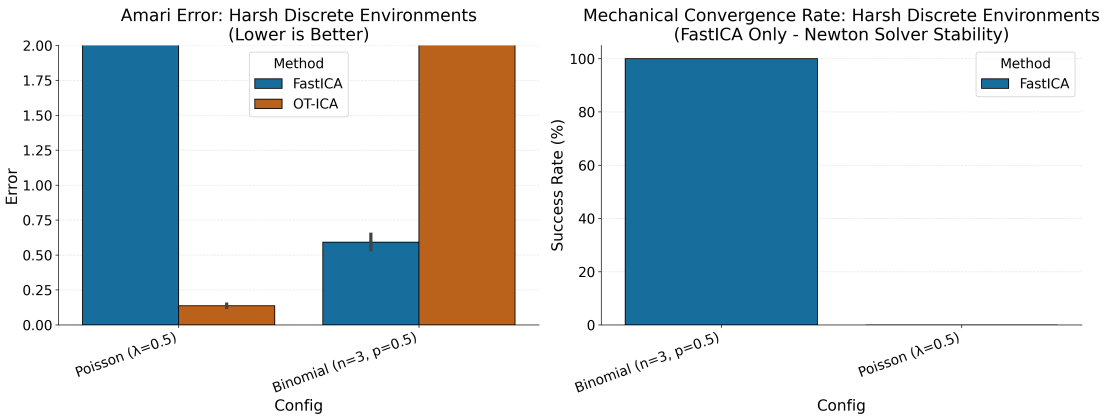

Figure 5.10: *Amari Error for skewed discrete parameterizations. OT-ICA achieves separation on the Poisson distribution but does not converge on the symmetric Binomial.*

distribution creates a narrower plateau at the minimum of the optimization surface (Figure 5.9). For the symmetric Binomial mixture ($n = 3$), OT-ICA got stuck, resulting in an Amari error of $E > 2.0$, due to the wider plateau at the optimum.

5.7 Synthesis: The Case for OT-ICA

The evaluations indicate a distinction between theoretical identifiability and optimization behavior. While statistical non-Gaussianity permits source separation, extracting discrete sources requires navigating a non-smooth optimization surface. As observed in the pure discrete environments, step-functions create zero-gradient plateaus that stall first-order solvers, despite the objective function correctly identifying the non-Gaussianity.

In heterogeneous environments containing both continuous and discrete variables, the continuous distributions smooth the overall joint Cumulative Distribution Function, allowing the gradient solver to utilize the optimal transport metric across the manifold (Figure 5.7).

FastICA's reliance on local negentropy approximations creates specific convergence limitations, such as the zero-negentropy condition and vanishing algorithmic curvature, which affect its performance in generalized hybrid mixtures. Because OT-ICA explicitly maps the full empirical geometry via the CDF, it bypasses these specific parametric approximations. Its performance is instead constrained by computational resources and the non-convexity of the Stiefel manifold. For homogeneous mixtures of standard continuous distributions, FastICA provides a computational advantage. In heterogeneous environments where proxy assumptions do not hold, OT-ICA provides a reliable alternative methodology capable of resolving signals that traditional proxy solvers cannot separate.

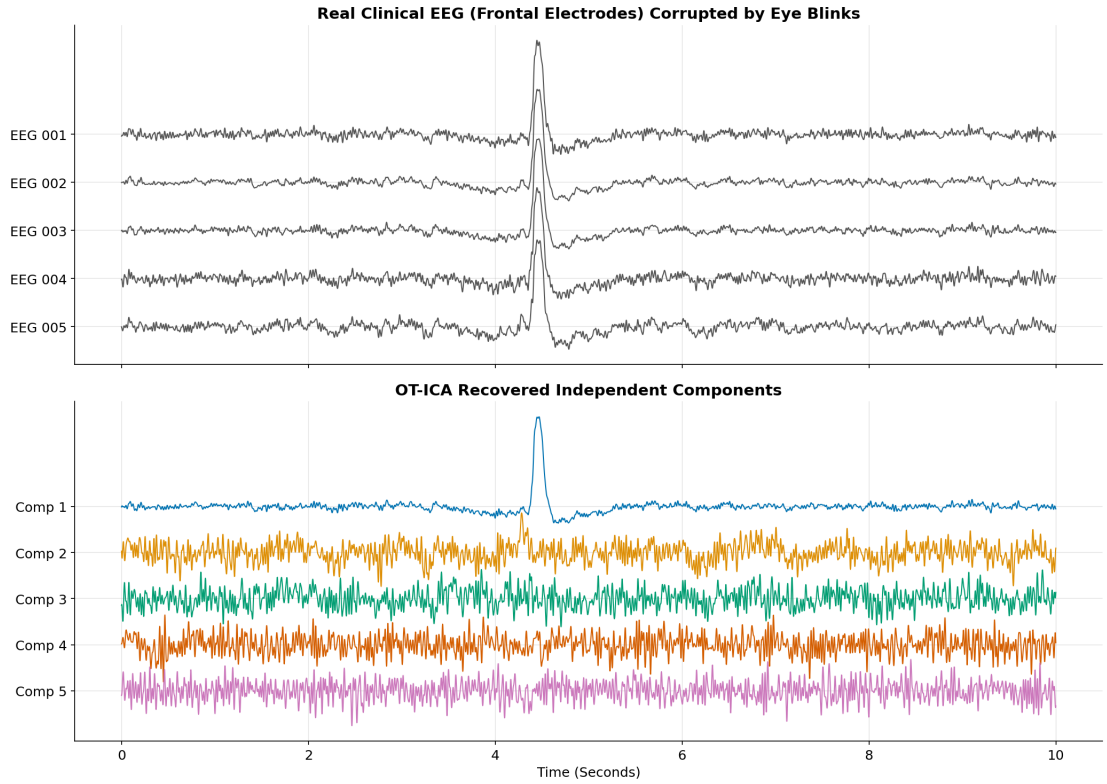

Figure 5.11: *Application of OT-ICA to EEG data.* ***Top (Plot A):*** *Five frontal EEG channels containing eye-blink artifacts.* ***Bottom (Plot B):*** *The independent components recovered by OT-ICA. The ocular artifact is isolated into a single independent component (Comp 1).*

5.8 Application: EEG Artifact Removal

5.8.1 The Challenge of Volume Conduction

In electroencephalography (EEG) recordings, the human skull acts as a linear volume conductor. Electrical signals generated by brain regions and non-brain muscles (such as the eyes) mix linearly before reaching the scalp electrodes. This satisfies the linear mixing model of ICA: $\mathbf{X} = \mathbf{AS}$.

Eye blinks (ocular artifacts) possess high amplitude and mask underlying brain waves. Because eye blinks are sparse and super-Gaussian, ICA is utilized to isolate and remove them. We apply the OT-ICA framework to a clinical dataset to isolate these artifacts from continuous brain wave recordings.

5.8.2 Empirical Results on Clinical EEG Data

We utilized the `mne` library to process a clinical MEG/EEG dataset. The data was band-pass filtered between 1 Hz and 40 Hz, and we isolated a 10-second window (10.0s to 20.0s) containing eye-blinks across five frontal EEG channels. The data was standardized before applying the OT-ICA algorithm.

The raw frontal electrodes contain synchronized high-amplitude spikes (Figure 5.11, Plot A). The OT-ICA framework separated these signals. The ocular artifact is isolated into a single independent component (Comp 1), leaving the underlying continuous signals in the remaining components (Figure 5.11, Plot B).

5.9 Application: Price Discovery and Information Shares

5.9.1 Non-Gaussianity in Econometric Market Microstructure

We apply OT-ICA to price discovery in financial econometrics. In fragmented financial markets, determining which market contributes most to price discovery involves measuring information shares. Standard vector error correction models (VECMs) identify information shares using uncorrelation of price innovations. However, identification relying solely on uncorrelation is statistically equivalent up to an orthogonal transformation (Zema & Cordoni, 2025).

By utilizing the non-Gaussianity of latent market shocks, ICA can resolve this orthogonal ambiguity, providing a measurement of market information shares.

5.9.2 Economic Identification Strategy

To address the permutation and sign ambiguities of ICA, we impose two economic identification constraints.

First, to resolve the permutation ambiguity, we enforce a Dominant Diagonal constraint. The structural shock assigned to Market i must contribute the maximum variance to that market's price innovations. We implement this using the Hungarian Algorithm for optimal bipartite matching.

Second, to resolve the sign ambiguity, we enforce a Positive Impact constraint based on positive spillovers. A structural innovation representing 'good news' is assumed to have a positive initial impact on its own market price. Columns of the unmixing matrix violating this logic are sign-flipped.

5.9.3 Empirical Results on Simulated Market Data

We simulated a fragmented market scenario targeting three markets with pre-defined Information Shares of 12%, 24%, and 64%. Market innovations were generated using Student-t distributions to satisfy the non-Gaussianity requirements. We executed 500 Monte Carlo simulations comprising 5,000 observations each.

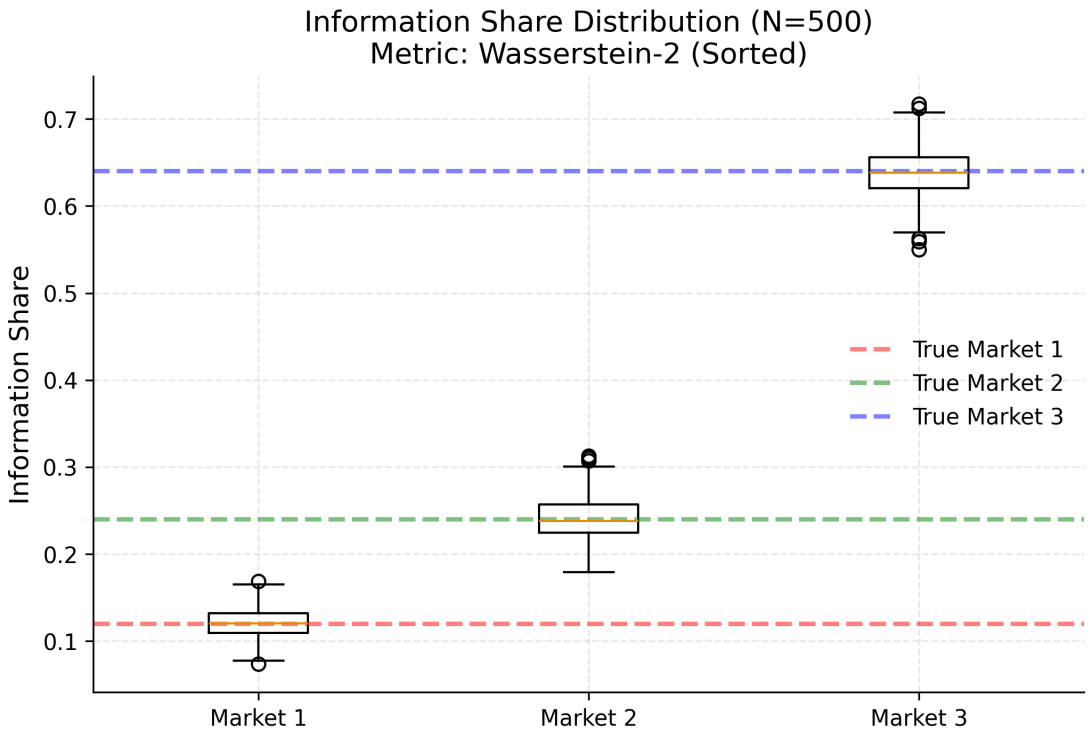

Figure 5.12: *Pairplot demonstrating the recovery of true structural shocks versus estimated shocks from OT-ICA for a single Monte Carlo run. The linear alignment indicates unmixing of the latent market innovations.*

A representative run demonstrates the correlation between the generative shocks and the empirical shocks estimated by OT-ICA (Figure 5.12).

Applying the Dominant Diagonal and Positive Impact constraints, we mapped the statistical components to their corresponding markets and calculated empirical information shares.

The estimated information shares converge to their theoretical targets across all 500 simulations (Table 5.5).

Market / Source	True IS	Estimated IS (Mean)	Standard Deviation
Market 1	0.1200	0.1212	0.0161
Market 2	0.2400	0.2399	0.0245
Market 3	0.6400	0.6389	0.0260

Table 5.5: *Monte Carlo simulation results comparing true Information Shares (IS) against the estimated IS recovered via OT-ICA over 500 runs.*

The empirical density of the estimated information shares is centered around the true parameter values across all simulation runs (Figure 5.13).

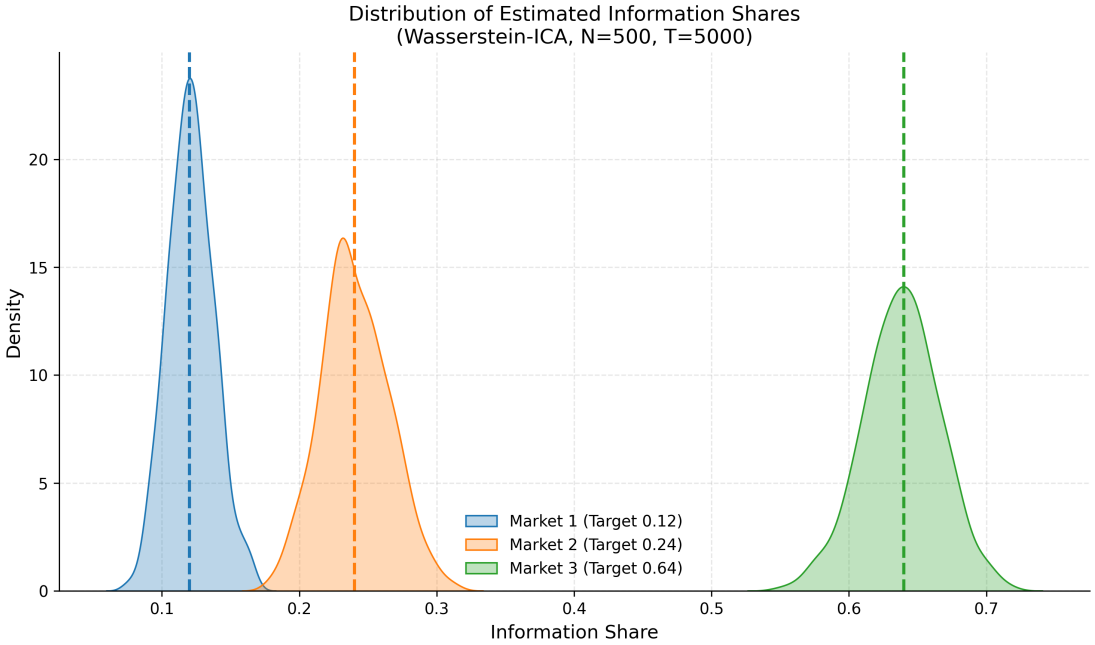

Figure 5.13: *Empirical density distributions of the estimated Information Shares across 500 Monte Carlo simulations. The vertical dashed lines represent the true theoretical values.*

Chapter 6

Conclusion

6.1 Summary of Contributions

This thesis evaluated the Optimal Transport Independent Component Analysis (OT-ICA) framework, transitioning linear source separation from local proxy approximations to global geometric matching. By formulating the search for independent components as a minimization of the squared Wasserstein distance (W_2^2) to a standard normal distribution on the Stiefel manifold, we addressed specific mathematical vulnerabilities present in established proxy algorithms.

Our theoretical analysis established that the squared Wasserstein distance bounds mixture non-Gaussianity, providing a direct geometric objective for extracting latent sources. Empirically, we demonstrated that while Newton-based algorithms like FastICA fail to converge in the presence of structural blind spots, such as zero-negentropy topologies and vanishing curvature, OT-ICA maintains stable convergence. By mapping the full empirical cumulative distribution function rather than relying on point-estimates of density, OT-ICA successfully resolved generalized hybrid mixtures where proxy solvers infinitely oscillate.

We introduced specific algorithmic adaptations to make this optimal transport framework computationally viable. These included exact analytical Gaussian targets to eliminate discretization noise, batched vectorization to handle the non-convex search space, and Gaussian dithering to process the discontinuous optimization landscapes of discrete data. Finally, we demonstrated the practical application of this framework, utilizing OT-ICA to isolate sparse artifacts in clinical EEG recordings and to resolve orthogonal ambiguities in econometric price discovery.

6.2 Limitations

The OT-ICA framework is subject to constraints regarding computational efficiency. Because calculating the exact Hessian of the Wasserstein distance relies on continuous density estimation, OT-ICA must utilize first-order Quasi-Newton methods (L-BFGS) combined with parallel random restarts. This results in execution times that scale exponentially with dimensionality, making it computationally expensive compared to fixed-point iteration solvers.

The framework also encounters a statistical performance ceiling related to the curse of dimensionality. As the dimensionality expands, the non-convex search space of the orthogonal group $O(d)$ becomes increasingly complex. In our empirical evaluations, a fixed sample size of $N = 10,000$ became insufficient to maintain the heuristic threshold for good source separation (Amari error $E < 0.3$) when $d > 40$, demonstrating a statistical resolution limit of the joint distribution.

Finally, while Gaussian dithering smooths step-wise empirical CDFs, highly symmetric and sparse discrete distributions (such as a low parameter Binomial distribution) yield optimization landscapes with flat, zero-gradient plateaus. While OT-ICA navigates hybrid environments where continuous variables smooth the joint distribution, isolating purely discrete, symmetrically clustered sources remains a limitation for first-order gradient solvers.

6.3 Future Work

Future research directions include the integration of entropic regularization (Sinkhorn distances) (?, ?; Peyré & Cuturi, 2019) to replace the Gaussian dithering step. Sinkhorn divergence provides a naturally differentiable approximation of the Wasserstein metric that processes discrete distributions without mathematically altering the underlying data.

Improving computational scaling represents another critical objective. Investigating stochastic second-order methods, or utilizing neural-based transport map approximations (such as Normalizing Flows), could reduce the execution time required to search the Stiefel manifold. Leveraging Amortized Optimal Transport to learn the transport map offline would bypass the $\mathcal{O}(N \log N)$ sorting requirement entirely during runtime.

Integrating the OT-ICA framework into the domain of causal discovery offers another promising avenue. Causal structures are frequently modeled using Linear Non-Gaussian Acyclic Models (LinGAM) (Shimizu, Hoyer, Hyvärinen, & Kerminen, 2006). Because LinGAM relies directly on ICA to identify the independent error terms that dictate causal directionality, applying a globally geometrically aware

metric like OT-ICA could improve causal estimation in complex, heterogeneous datasets where traditional ICA proxies fail. Furthermore, the emerging field of Causal Representation Learning seeks to infer causally related latent variables. The recently proposed Causal Component Analysis (CauCA) (Liang et al., 2023) bridges ICA and causal learning by modeling the causal dependence among latent components under known graphs. Integrating the robust Wasserstein metric of OT-ICA into the CauCA framework could enhance the estimation of unmixing functions and causal mechanisms, particularly when latent variables exhibit complex, heterogeneous non-Gaussianity.

Finally, extending the geometric properties of optimal transport contrast functions to nonlinear Independent Component Analysis represents a compelling frontier. Unsupervised learning of latent variable models via nonlinear ICA requires strict constraints on the function class to ensure identifiability, as explored by Buchholz, Besserve, and Schölkopf (Buchholz, Besserve, & Schölkopf, 2022). The optimal transport metric, which inherently evaluates global structural deformations, may serve to effectively regularize these constrained function spaces, enabling reliable unmixing of nonlinear representations.

Appendix A

Mathematical Proofs and Derivations

A.1 Derivation of the FastICA Fixed-Point Failure Condition

Proof of Theorem 8.1 Denote by $H(\mathbf{w})$ the function to be minimized/maximized, $E\{G(\mathbf{w}^T\mathbf{z})\}$. Make the orthogonal change of coordinates $\mathbf{q} = \mathbf{A}^T\mathbf{V}^T\mathbf{w}$. Then we can calculate the gradient as $\frac{\partial H(\mathbf{q})}{\partial \mathbf{q}} = E\{\mathbf{s}g(\mathbf{q}^T\mathbf{s})\}$ and the Hessian as $\frac{\partial^2 H(\mathbf{q})}{\partial \mathbf{q}^2} = E\{\mathbf{s}\mathbf{s}^T g'(\mathbf{q}^T\mathbf{s})\}$. Without loss of generality, it is enough to analyze the stability of the point $\mathbf{q} = \mathbf{e}_1$, where $\mathbf{e}_1 = (1, 0, 0, 0, \dots)$. Evaluating the gradient and the Hessian at point $\mathbf{q} = \mathbf{e}_1$, we get using the independence of the s_i,

$$\frac{\partial H(\mathbf{e}_1)}{\partial \mathbf{q}} = \mathbf{e}_1 E\{s_1 g(s_1)\} \tag{A.1}$$

and

$$\frac{\partial^2 H(\mathbf{e}_1)}{\partial \mathbf{q}^2} = \mathrm{diag}(E\{s_1^2 g'(s_1)\}, E\{g'(s_1)\}, E\{g'(s_1)\}, \dots). \tag{A.2}$$

Making a small perturbation $\epsilon = (\epsilon_1, \epsilon_2, \dots)$, we obtain

$$\begin{aligned} H(\mathbf{e}_1 + \epsilon) &= H(\mathbf{e}_1) + \epsilon^T \frac{\partial H(\mathbf{e}_1)}{\partial \mathbf{q}} + \frac{1}{2}\epsilon^T \frac{\partial^2 H(\mathbf{e}_1)}{\partial \mathbf{q}^2}\epsilon + o(\|\epsilon\|^2) \\ &= H(\mathbf{e}_1) + E\{s_1 g(s_1)\}\epsilon_1 + \frac{1}{2}\Big[E\{s_1^2 g'(s_1)\}\epsilon_1^2 + E\{g'(s_1)\}\sum_{i>1}\epsilon_i^2\Big] + o(\|\epsilon\|^2) \end{aligned} \tag{A.3}$$

Due to the constraint $\|\mathbf{w}\| = 1$ we get $\epsilon_1 = \sqrt{1 - \epsilon_2^2 - \epsilon_3^2 - \dots} - 1$. Due to the fact that $\sqrt{1-\gamma} = 1 - \gamma/2 + o(\gamma)$, the term of order ϵ_1^2 is $o(\|\epsilon\|^2)$, i.e., of higher order, and can be neglected. Using the aforementioned first-order approximation for ϵ_1

we obtain $\epsilon_1 = -\sum_{i>1} \epsilon_i^2/2 + o(\|\epsilon\|^2)$, which finally gives

$$H(\mathbf{e}_1 + \epsilon) = H(\mathbf{e}_1) + \frac{1}{2}\Big[E\{g'(s_1) - s_1 g(s_1)\}\Big] \sum_{i>1} \epsilon_i^2 + o(\|\epsilon\|^2) \tag{A.4}$$

which clearly proves $\mathbf{q} = \mathbf{e}_1$ is an extremum, and of the type implied by the condition of the theorem.

Proof of convergence of FastICA The convergence is proven under the assumptions that first, the data follows the ICA data model and second, that the expectations are evaluated exactly.

Let g be the nonlinearity used in the algorithm. In the case of the kurtosis-based algorithm, this is the cubic function, so we obtain that algorithm as a special case of the following proof for a general g. We must also make the following technical assumption:

$$E\{s_i g(s_i) - g'(s_i)\} \neq 0, \quad \text{for any } i \tag{A.5}$$

which can be considered a generalization of the condition valid when we use kurtosis, that the kurtosis of the independent components must be nonzero. If this is true for a subset of independent components, we can estimate just those independent components.

To begin with, make the change of variable $\mathbf{q} = \mathbf{A}^T\mathbf{V}^T\mathbf{w}$, as earlier, and assume that $\mathbf{q}$ is in the neighborhood of a solution (say, $q_1 \approx 1$ as before). As shown in the proof of Theorem 8.1, the change in q_1 is then of a lower order than the change in the other coordinates, due to the constraint $\|\mathbf{q}\| = 1$. Then we can expand the terms using a Taylor approximation for g and g', first obtaining

$$\begin{aligned} g(\mathbf{q}^T\mathbf{s}) = g(q_1 s_1) + g'(q_1 s_1)\mathbf{q}_{-1}^T\mathbf{s}_{-1} + \frac{1}{2}g''(q_1 s_1)(\mathbf{q}_{-1}^T\mathbf{s}_{-1})^2 \\ + \frac{1}{6}g'''(q_1 s_1)(\mathbf{q}_{-1}^T\mathbf{s}_{-1})^3 + O(\|\mathbf{q}_{-1}\|^4) \end{aligned} \tag{A.6}$$

and then

$$\begin{aligned} g'(\mathbf{q}^T\mathbf{s}) = g'(q_1 s_1) + g''(q_1 s_1)\mathbf{q}_{-1}^T\mathbf{s}_{-1} \\ + \frac{1}{2}g'''(q_1 s_1)(\mathbf{q}_{-1}^T\mathbf{s}_{-1})^2 + O(\|\mathbf{q}_{-1}\|^3) \end{aligned} \tag{A.7}$$

where $\mathbf{q}_{-1}$ and $\mathbf{s}_{-1}$ are the vectors $\mathbf{q}$ and $\mathbf{s}$ without their first components. Denote by $\mathbf{q}^+$ the new value of $\mathbf{q}$ (after one iteration). Thus we obtain, using the independence of the s_i and doing some tedious but straightforward algebraic manipulations,

$$q_1^+ = E\{s_1 g(q_1 s_1) - g'(q_1 s_1)\} + O(\|\mathbf{q}_{-1}\|^2) \tag{A.8}$$

$$
\begin{aligned}
q_i^+ = & \frac{1}{2} E\{s_i^3\} E\{g''(s_1)\} q_i^2 \\
& + \frac{1}{6} \mathrm{kurt}(s_i) E\{g'''(s_1)\} q_i^3 + O(\|\mathbf{q}_{-1}\|^4), \quad \text{for } i > 1
\end{aligned}
\tag{A.9}
$$

We obtain also

$$
\mathbf{q}^* = \mathbf{q}^+ / \|\mathbf{q}^+\| \tag{A.10}
$$

This shows clearly that under assumption (A.5), the algorithm converges (locally) to such a vector $\mathbf{q}$ that $q_1 = \pm 1$ and $q_i = 0$ for $i > 1$. This means that $\mathbf{w} = ((\mathbf{VA})^T)^{-1}\mathbf{q}$ converges, up to the sign, to one of the rows of the inverse of the mixing matrix $\mathbf{VA}$, which implies that $\mathbf{w}^T\mathbf{z}$ converges to one of the s_i. Moreover, if $E\{g''(s_1)\} = 0$, i.e., if the s_i has a symmetric distribution, as is usually the case, the convergence is cubic. In other cases, the convergence is quadratic. If kurtosis is used, however, we always have $E\{g''(s_1)\} = 0$ and thus cubic convergence. In addition, if $G(y) = y^4$, the local approximations are exact, and the convergence is global.

References

Absil, P.-A., Mahony, R., & Sepulchre, R. (2008). *Optimization algorithms on matrix manifolds*. Princeton University Press.

Amari, S.-I., Cichocki, A., & Yang, H. H. (1996). A new learning algorithm for blind signal separation. In *Advances in neural information processing systems* (pp. 757–763).

Billingsley, P. (1995). *Probability and measure* (3rd ed.). John Wiley & Sons.

Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, *44*(4), 375–417.

Buchholz, S., Besserve, M., & Schölkopf, B. (2022). Function classes for identifiable nonlinear independent component analysis. In *Advances in neural information processing systems* (Vol. 35).

Caffarelli, L. A. (1992). The regularity of mappings with a convex potential. *Journal of the American Mathematical Society*, *5*(1), 99–104.

Cardoso, J.-F. (2022). Independent component analysis in the light of information geometry. *Entropy*, *24*(3), 377.

Comon, P. (1994). Independent component analysis, a new concept? *Signal processing*, *36*(3), 287–314.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory*. John Wiley & Sons.

Fournier, N., & Guillin, A. (2015). On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, *162*(3), 707–738.

Hyvärinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis*. John Wiley & Sons.

Jutten, C., & Herault, J. (1985). Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal processing*, *24*(1), 1–10.

Kantorovich, L. V. (1942). On the translocation of masses. *C. R. (Doklady) Acad. Sci. URSS (N.S.)*, *37*, 199–201.

Liang, W., Kekić, A., von Kügelgen, J., Buchholz, S., Besserve, M., Gresele, L.,

& Schölkopf, B. (2023). Causal component analysis. In *Advances in neural information processing systems* (Vol. 36).

Liu, D. C., & Nocedal, J. (1989). On the limited memory bfgs method for large scale optimization. *Mathematical programming*, *45*(1-3), 503–528.

MacKay, D. J. C. (2003). *Information theory, inference and learning algorithms*. Cambridge University Press.

Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*.

Peyré, G., & Cuturi, M. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, *11*(5-6), 355–607.

Santambrogio, F. (2015). *Optimal transport for applied mathematicians* (Vol. 87). Birkhäuser Basel.

Schuchman, L. (1964). Dither signals and their effect on quantization noise. *IEEE Transactions on Communication Technology*, *12*(4), 162–165.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., & Kerminen, A. (2006). A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, *7*(10).

Villani, C. (2003). *Topics in optimal transportation* (Vol. 58). American Mathematical Society.

Zema, S. M., & Cordoni, F. (2025). A unifying non-gaussian information share approach to price discovery. *Available at SSRN 5234231*.