Linear Independent Component Analysis via Optimal Transport

Ashutosh Jha[3] **Michel Besserve**[1,2] **Simon Buchholz**[1]

[1]Max Planck Institute for Intelligent Systems, Tübingen
[2]Institute of Artificial Intelligence, TU Braunschweig
[3]Methods Center, University of Tübingen

Abstract

Linear Independent Component Analysis (ICA) recovers jointly independent source signals from their linear mixtures. To achieve this, classical ICA algorithms attempt to maximize non-Gaussianity, measured by negentropy, which is linked to independence by information theory. Because exact negentropy optimization is intractable, they rely on proxy contrast functions, such as fourth-order cumulants, and parametric log-likelihoods. We propose instead to measure non-Gaussianity using the squared L_2-Wasserstein distance (W_2^2) to a standard Gaussian. We show that the Wasserstein distance between a standard normal distribution and linear projections of the data is maximized when the projection recovers an independent component. Based on this observation, we propose the OT-ICA algorithm, which finds this projection by gradient-based optimization. Empirical evaluation on simulated data shows that OT-ICA outperforms proxy-based methods for different distributions of the latent variables. Application to EEG artifact removal and econometric price discovery confirm OT-ICA can be used for applied ICA tasks without distributional assumptions.

1 INTRODUCTION

Independent Component Analysis (ICA) [Jutten and Herault, 1985, Comon, 1994] recovers mutually independent source signals $\mathbf{Z}$ from a linear mixture $\mathbf{X} = \mathbf{AS}$. Identifiability holds under the following conditions.

Theorem 1 (Linear ICA [Comon, 1994])**.** *Let $\mathbf{X} \in \mathbb{R}^d$ be observed data and $\mathbf{Z} \in \mathbb{R}^d$ latent sources. Assume (i) $\mathbf{X} = \mathbf{AZ}$ with $\mathbf{A}$ invertible, (ii) $z_1, \dots, z_d$ are mutually independent, and (iii) at most one z_i is Gaussian. Then $\mathbf{A}$ is identifiable up to column permutation and scaling.*

Algorithms for learning $\mathbf{A}$ from data generally rely on the property that linear mixtures of independent non-Gaussian sources are more Gaussian than their constituents [Hyvärinen et al., 2001], indeed, in the limit of many sources, convergence to a Gaussian follows from the central limit theorem. A rigorous statement of this property was given in Cardoso [2022].

Theorem 2 (Independence and Non-Gaussianity [Cardoso, 2022])**.** *Let $Y \in \mathbb{R}^d$ be whitened ($\mathrm{Cov}(Y) = \mathbf{I}$). The mutual information $\mathscr{I}(Y)$, which measures statistical dependence among components $Y_1, \dots, Y_d$, satisfies:*

$$\mathscr{I}(Y) = G(Y) - \sum_i G(Y_i) \tag{1}$$

where $G(\cdot)$ denotes non-Gaussianity defined as the minimal KL divergence to a Gaussian distribution.

This result implies that minimizing $\mathscr{I}(Y)$ is equivalent to maximizing total marginal non-Gaussianity $\sum_i G(Y_i)$. However, evaluating $G(Y_i)$ requires the marginal density. Classical algorithms therefore replace it with proxy contrasts: FastICA [Hyvärinen et al., 2001, Amari et al., 1996] uses logcosh negentropy approximations, JADE [Cardoso and Souloumiac, 1993] fourth-order cumulants, InfoMax [Bell and Sejnowski, 1995] a parametric log-likelihood, and Picard [Ablin et al., 2018] an adaptive nonlinearity with a provably converging solver. These proxies exhibit shortcomings such as evaluating to zero on specific non-Gaussian distributions or inducing singular Hessian matrices that stall the fixed-point solver shown in Appendix E.

In this paper we use Optimal Transport (OT) [Monge, 1781, Kantorovich, 1942] to define a contrast. This allows us to apply OT to linear ICA and extends the long list of applications of optimal transport techniques in machine learning such as generative modelling [Arjovsky et al., 2017], normality testing [del Barrio et al., 1999], and distribution comparison [Peyré and Cuturi, 2019]. Here we propose to quantify non-Gaussianity as the cost of transporting the (empirical)

projection of the observations to a fixed standard Gaussian reference.[1] We justify this choice in Theorem 3 below which implies that (in the population setting) this contrast is maximized when an independent component is recovered. This then motivates our OT-ICA algorithm, which relies on gradient ascent on the optimal transport cost.

2 THE OT-ICA FRAMEWORK

We consider the standard linear ICA setting $\mathbf{X} = \mathbf{AZ}$ where $\mathbf{A} \in \mathbb{R}^{d\times d}$ is an invertible matrix. By preprocessing the observations $\mathbf{X}$ through centering and whitening, we can assume without loss of generality that $\mathbb{E}(\mathbf{X}) = 0$ and $\mathrm{Cov}[\mathbf{X}] = \mathbf{I}_{d\times d}$ (see Appendix A for details). Then we can constrain the unmixing matrix $\mathbf{B}$ to the Orthogonal Group $\mathrm{O}(d)$ (the set of $d \times d$ real matrices satisfying $\mathbf{BB}^\top = \mathbf{I}$), and seek $\mathbf{B} \in \mathrm{O}(d)$ such that $\mathbf{BX} = \mathbf{BAZ}$ has independent components, or equivalently, $\mathbf{B}$ inverts $\mathbf{A}$ up to a scaled permutation matrix. This can be done by iteratively seeking optimal projection directions $\mathbf{b} \in \mathbb{R}^d$, corresponding to each row of $\mathbf{B}$, maximizing a non-Gaussianity function $G(\mathbf{b}^\top\mathbf{X})$ of the projected data.

We will use the square of the L_2-Wasserstein distance W_2 [Villani, 2003], defined as a minimum-cost coupling between probability distributions (see Appendix A):

$$W_2^2(\mu,\nu) \triangleq W_2(\mu,\nu)^2 = \inf_{\pi\in\Pi(\mu,\nu)} \int \|x-y\|^2 d\pi(x,y), \quad (2)$$

where $\Pi(\mu,\nu)$ denotes (for measures μ,ν on $\mathbb{R}^d$) the set of all measures π on $\mathbb{R}^{d+d}$ with first marginal μ and second marginal ν.

OT-ICA for a single component. We consider the contrast given by $W_2^2\big(\mathbb{P}_{\mathbf{b}^\top\widetilde{\mathbf{X}}}, \Gamma\big)$, where $\Gamma \equiv \mathcal{N}(0,1)$ denotes the standard Gaussian. Since W_2 induces a proper metric on probability distributions [Villani, 2003], this contrast is nonzero on all non-Gaussian distributions.

OT-ICA maximizes this contrast over the direction $\mathbf{b}$

$$\widehat{\mathbf{b}} = \arg\max_{\mathbf{b}\in\mathbb{R}^d, |\mathbf{b}|=1} W_2^2\big(\mathbb{P}_{\mathbf{b}^\top\widetilde{\mathbf{X}}}, \Gamma\big). \quad (3)$$

The following bound, due to Johnson and Samworth [2005], establishes that the projection $\widehat{\mathbf{b}}^\top\mathbf{X}$ from (3) indeed recovers an independent component; our contribution is identifying it as an admissible ICA contrast.

Theorem 3 (W_2^2 ICA Contrast; Johnson and Samworth 2005)**.** *Let $\mathbf{Z} = (Z_1,\dots,Z_d)^\top$ be centered independent sources with unit variance and at most one Z_i Gaussian. Assume they all have a smooth and strictly positive density. Then the following bound holds for any normalized weight vector α with at least two nonzero components*

$$W_2^2(\alpha\cdot\mathbf{Z},\Gamma) < \sum_{i=1}^d \alpha_i^2 W_2^2(Z_i,\Gamma). \quad (4)$$

The proof of this result is in Appendix B. We find the following immediate corollary 1.

Corollary 1. *Under the assumptions of Theorem 3 we have*

$$\mathrm{argmax}_{|\alpha|=1}\, W_2^2(\alpha\cdot\mathbf{Z},\Gamma) \subset \{e_1,\dots,e_d\}, \quad (5)$$

i.e., the contrast is maximal only for independent components.

This corollary implies that (3) is maximized by $\widehat{\mathbf{b}}$ such that $\widehat{\mathbf{b}}^\top\mathbf{X} = \widehat{\mathbf{b}}^\top\mathbf{AZ} = \mathbf{Z}_i$ for some i, i.e., when the projection recovers an independent component. OT-ICA optimizes the Wasserstein contrast $W_2^2\big(\mathbb{P}_{\mathbf{b}^\top\widetilde{\mathbf{X}}},\Gamma\big)$ through gradient ascent. This requires a fully differentiable implementation of the Wasserstein distance. This is straightforward in the one dimensional case, where an optimal coupling is given by matching the percentiles of the distributions, such that the Wasserstein distance is given by

$$W_2^2(\mu,\nu) = \int_0^1 |F_\mu^{-1}(t) - F_\nu^{-1}(t)|^2\, dt \quad (6)$$

where F_μ denotes the CDF of μ. In particular, the optimal coupling for finitely many samples simply matches the ordered values from the two empirical distributions.

OT-ICA for all components. After whitening the data, the projection directions associated to all independent components are mutually orthogonal. While components can be extracted iteratively by searching the shrinking orthogonal complement one by one, Theorem 3 extends to the following corollary 2 which justifies the joint estimation of the entire unmixing matrix $\mathbf{B} \in \mathrm{O}(d)$.

Corollary 2. *Under the assumptions of Theorem 3, for any orthogonal matrix $\mathbf{B} \in \mathrm{O}(d)$ with rows $\mathbf{b}_i^\top$, the total marginal contrast satisfies:*

$$\sum_{i=1}^d W_2^2(\mathbf{b}_i^\top\mathbf{Z},\Gamma) \le \sum_{j=1}^d W_2^2(Z_j,\Gamma), \quad (7)$$

where equality holds if and only if $\mathbf{B}$ is a signed permutation matrix.

This corollary 2 (proof in Appendix B) theoretically justifies *symmetric joint optimization* [Hyvärinen et al., 2001]. Rather than iterative deflation, all d rows of $\mathbf{B}$ are optimized simultaneously by ascending the total contrast sum, restoring the $\mathrm{O}(d)$ constraint via symmetric decorrelation

[1] Concurrent and independent work [Laplante et al., 2026] also proposes W_2^2-to-Gaussian as an ICA contrast. It was published on arXiv on July 14, 2026, after this paper was already accepted at TPM, UAI 2026 on July 04, 2026.

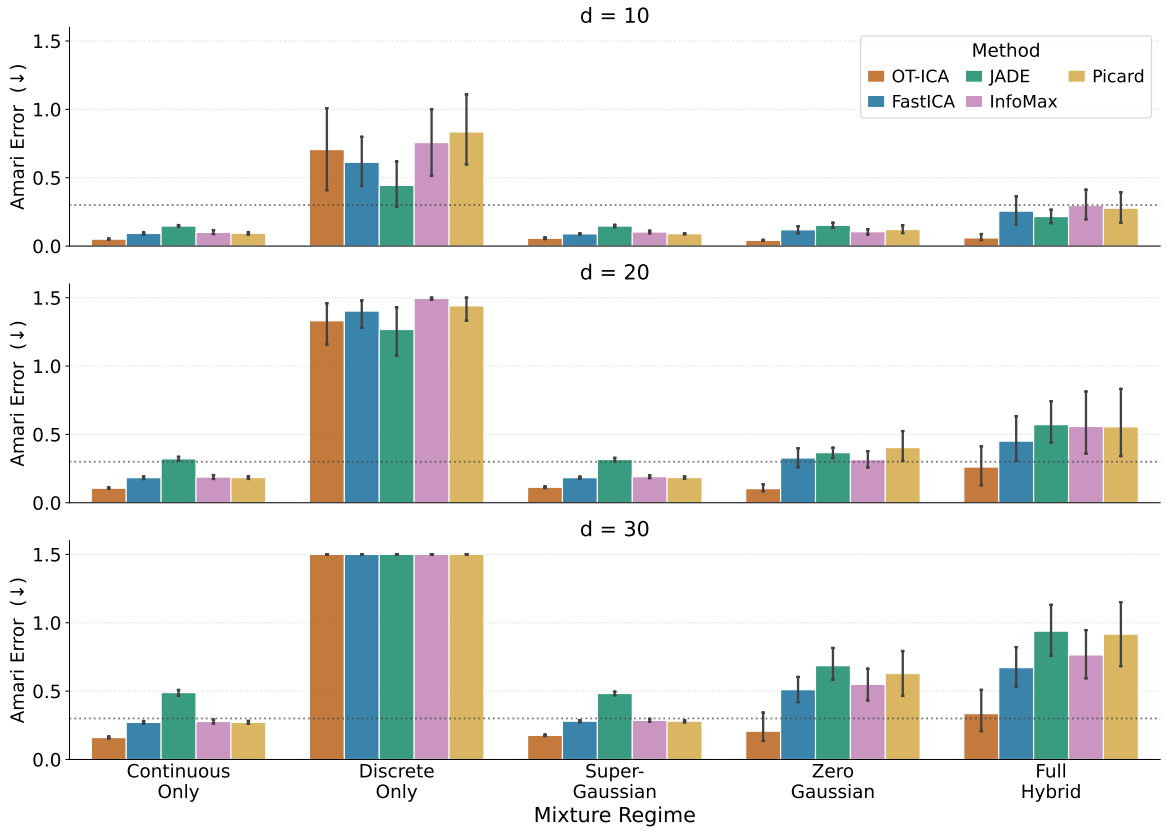

Figure 1: *Mean Amari error (↓ better, clipped at* 1.5*) across five mixture regimes and* $d \in \{10, 20, 30\}$ *(*$N = 10{,}000$*, 10 trials, 95% CI). Dotted line marks* $E = 0.3$*, the good-separation threshold.*

retraction at every step. The full details of the resulting algorithm can be found in Appendix C and pseudocode is given in Algorithm 1.

Let us here briefly mention three important algorithmic components utilized in practice: (i) **Analytical Gaussian Targets** [Fournier and Guillin, 2015], replacing sampled quantile targets of the standard normal distribution with closed-form bin expectations to eliminate approximation noise; (ii) **Riemannian Gradient Retraction** [Absil et al., 2008], projecting gradients onto the tangent space of Orthogonal Group $\mathrm{O}(d)$ and retracting via symmetric decorrelation; and (iii) **Gaussian Dithering** [Schuchman, 1964], smoothing discrete CDFs via narrow Gaussian convolution.

3 SYNTHETIC EXPERIMENTS

We first evaluate our method on synthetic data[2]. Separation quality is measured by the Amari Performance Index (E) [Amari et al., 1996] (defined in Appendix D), which measures the deviation of the global transfer matrix $\mathbf{P} = \mathbf{B}_{\text{est}}\mathbf{A}$ from a generalized permutation matrix ($E < 0.1$: near-perfect; $E < 0.3$: good separation; $E \geq 0.5$: failure). As a sanity check we first consider separation of four different sources with $N = 10,000$ samples where OT-ICA achieves $E = 0.016$ against FastICA's 0.019 on this mixture (Table 1); full transfer matrices and recovered densities are in Appendix D.3.

Method	Amari Error
OT-ICA	0.016
FastICA	0.019

Table 1: *Baseline validation on a 4D mixture of Laplace, Uniform, Student-*$t(3)$*, and Beta*$(0.5, 0.5)$ *sources (*$N = 10{,}000$*).*

We now study the empirical performance of the proposed algorithm systematically. We use FastICA [Hyvärinen et al., 2001], JADE [Cardoso and Souloumiac, 1993], InfoMax [Bell and Sejnowski, 1995], and Picard [Ablin et al., 2018] as baselines and consider five mixture regimes at $d \in \{10, 20, 30\}$, $N = 10{,}000$, 10 trials per condition. The results can be found in Figure 1.

Continuous and heterogeneous regimes. OT-ICA achieves the lowest Amari error on all four continuous and mixed-type configurations at every dimension (Table 4, Appendix D). On *Continuous Only* sources, OT-ICA reduces error 40–45% over FastICA across all three dimensions. The margin is larger on *Full Hybrid* mixtures (Laplace, Gaussian, Uniform, Student-t, Beta combined): OT-ICA attains $E = 0.059, 0.261, 0.335$ at $d = 10, 20, 30$ against

[2]Code to reproduce the experiments is available under `https://github.com/ashutoshjha3103/ot_in_linear_ica`

FastICA's 0.255, 0.451, 0.671, a factor of 2–4×.

Discrete sources. On Discrete only mixtures, OT-ICA underperforms JADE at $d = 10$ (0.706 vs. 0.443) and $d = 20$ (1.331 vs. 1.267); note both values exceed $E = 0.5$, placing all methods in failure territory where differences reflect degree of failure rather than separation quality (a d-dimensional random orthogonal matrix yields $E \approx (d-1)/d$, e.g. ≈ 0.97 at $d = 30$; Appendix D). Table 5 shows W_2^2 retains a clear non-Gaussianity signal on count data, exceeding logcosh resolution by more than $10\times$ on standard Poisson($\lambda = 3$), so the failure is not a contrast deficiency. Step-function CDFs create gradient plateaus that stall the Riemannian solver (Appendix D.5).

OT-ICA's gradient-based Riemannian solver requires more time than FastICA's fixed-point Newton iterations, with the gap widening as d increases; this cost is offset on heterogeneous mixtures where proxy methods fail to separate independent sources (Appendix F).

4 APPLICATIONS

Structural Identification in Price Discovery. In fragmented financial markets, the Information Share (IS) [Hasbrouck, 1995, Baillie et al., 2002] quantifies each venue's contribution to price discovery from the structural decomposition $u_t = \mathbf{B}\epsilon_t$, with u_t the VECM reduced-form residuals and ϵ_t structural innovations [Engle and Granger, 1987, Johansen, 1991]. The covariance constraint $\mathbf{\Omega} = \mathbf{B}\mathbf{B}^\top$ leaves $\mathbf{B}$ unidentified up to an orthogonal rotation; Cholesky identification depends on an arbitrary market ordering. Factoring $\mathbf{B} = \mathbf{S}\mathbf{C}$ [Zema and Cordoni, 2025] reduces identification to finding $\mathbf{C}$, which OT-ICA recovers from the non-Gaussianity of innovations without ordering assumptions (full model in Appendix G). We validate on a simulated three-market VECM [Zema and Cordoni, 2025]: $\mathbf{B}_{\text{true}} = \text{diag}(\sqrt{0.12}, \sqrt{0.24}, \sqrt{0.64})$ encodes true IS values $[0.12, 0.24, 0.64]$ by construction, with Student-t innovations enabling ICA identification. IS estimates match true values to within 0.002 across 500 Monte Carlo runs (see Table 2).

Market / Source	True IS	Estimated IS (Mean)	Std. Dev.
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 2: *IS recovery across 500 Monte Carlo runs without imposing Cholesky ordering.*

EEG Artifact Removal. The skull acts as a linear volume conductor [Hyvärinen et al., 2001]: ocular blink artifacts mix with neural signals before reaching scalp electrodes. Blink artifacts are impulsive and super-Gaussian; their high W_2^2 to Gaussian separates them from neural components without specifying a density model (dataset, preprocessing, and configuration in Appendix H).

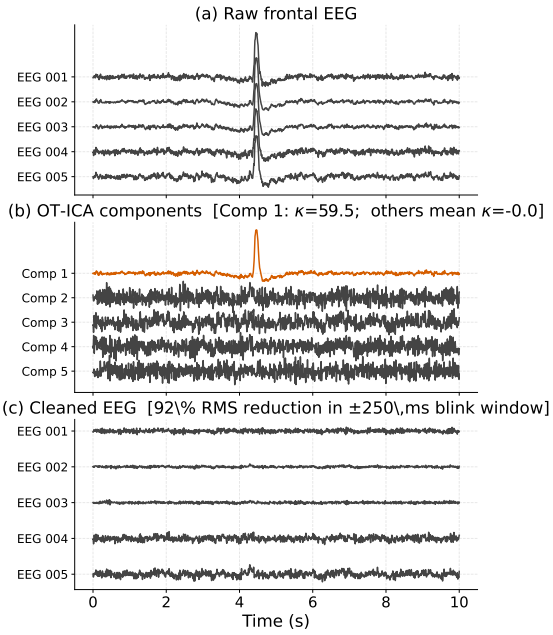

Figure 2: *OT-ICA on five frontal EEG channels (MNE sample dataset). Panel (b): isolated blink component (orange) with excess kurtosis κ in the panel title; remaining components have mean κ near zero. Panel (c): signal reconstructed by zeroing the artifact component, with RMS reduction in the ± 250 ms blink window in the panel title.*

Figure 2 shows OT-ICA concentrating the blink artifact into a single component with excess kurtosis substantially above the remaining four. Zeroing that component and inverting the total unmixing matrix produces the RMS reduction quantified in the panel title, without tuning a density model.

5 CONCLUSION

OT-ICA replaces proxy contrast functions with the W_2^2 metric computed by quantile sorting; Theorem 3 guarantees the contrast is maximized at independent components, and OT-ICA outperforms proxy-based algorithms on all continuous and heterogeneous distributions, with the margin widening where proxy contrasts fail to extract independent components. Price discovery and EEG artifact removal confirm application to ICA tasks without distributional assumptions. On purely discrete sources, all methods fail as d increases ($E > 0.5$); for OT-ICA the W_2^2 contrast retains signal throughout but step-function CDFs stall the solver, and per-iteration cost grows with d (Appendix F), motivating Sinkhorn distances [Cuturi, 2013, Peyré and Cuturi, 2019] and amortized transport as future directions, alongside extending the distribution-free contrast to causal discovery [Shimizu et al., 2006, Liang et al., 2023, Buchholz et al., 2022].

References

Pierre Ablin, Jean-François Cardoso, and Alexandre Gramfort. Faster independent component analysis by preconditioning with Hessian approximations. *IEEE Transactions on Signal Processing*, 66(15):4040–4049, 2018.

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

Shun-Ichi Amari, Andrzej Cichocki, and Howard H Yang. A new learning algorithm for blind signal separation. In *Advances in neural information processing systems*, pages 757–763, 1996.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 06–11 Aug 2017.

Richard T Baillie, G Geoffrey Booth, Yiuman Tse, and Tatyana Zabotina. Price discovery and common factor models. *Journal of Financial Markets*, 5(3):309–321, 2002.

Anthony J Bell and Terrence J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

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

Simon Buchholz, Michel Besserve, and Bernhard Schölkopf. Function classes for identifiable nonlinear independent component analysis. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

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

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

Jean-François Cardoso and Antoine Souloumiac. Blind beamforming for non-Gaussian signals. *IEE Proceedings F – Radar and Signal Processing*, 140(6):362–370, 1993.

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

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Eustasio del Barrio, Juan A Cuesta-Albertos, Carlos Matrán, and Jesús M Rodríguez-Rodríguez. Tests of goodness of fit based on the L_2-Wasserstein distance. *Annals of Statistics*, 27(4):1230–1239, 1999.

Robert F Engle and Clive WJ Granger. Co-integration and error correction: Representation, estimation, and testing. *Econometrica: Journal of the Econometric Society*, 55(2): 251–276, 1987.

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

Joel Hasbrouck. One measure of price discovery in fragmented markets. *The Journal of Finance*, 50(4):1175–1199, 1995.

Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.

Søren Johansen. Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models. *Econometrica: Journal of the Econometric Society*, 59(6): 1551–1580, 1991.

Oliver Johnson and Richard Samworth. Central limit theorem and convergence to stable laws in mallows distance. *Bernoulli*, 11(5):829–845, 2005.

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

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

Harold W Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2): 83–97, 1955.

Félix Laplante, Christophe Ambroise, and Pierre Humbert. Contrast-free ica and causal inference via wasserstein distances to the gaussian, 2026. URL `https://arxiv.org/abs/2607.12832`.

Te-Won Lee, Mark Girolami, and Terrence J Sejnowski. Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural Computation*, 11(2):417–441, 1999.

Wendong Liang, Armin Kekić, Julius von Kügelgen, Simon Buchholz, Michel Besserve, Luigi Gresele, and Bernhard Schölkopf. Causal component analysis. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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

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

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

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Cédric Villani. *Topics in Optimal Transportation*, volume 58. American Mathematical Society, 2003.

Sebastiano Michele Zema and Francesco Cordoni. A unifying non-gaussian information share approach to price discovery. *Available at SSRN 5234231*, 2025.

Supplementary Material for:
Linear Independent Component Analysis via Optimal Transport

A BACKGROUND AND MOTIVATION

A.1 CENTERING, WHITENING, AND THE ORTHOGONAL GROUP $\mathrm{O}(d)$ SEARCH SPACE

The ICA model $\mathbf{X} = \mathbf{AS}$ has d^2 free parameters in the mixing matrix $\mathbf{A}$. Two lossless preprocessing steps reduce this to a $\frac{d(d-1)}{2}$-dimensional rotational search.

Centering. Subtracting the mean, $\mathbf{X} \leftarrow \mathbf{X} - \mathbb{E}[\mathbf{X}]$, ensures the data has zero mean. If $\mathbf{X} = \mathbf{AS}$, centering $\mathbf{X}$ is equivalent to centering $\mathbf{S}$, leaving the mixing geometry intact.

Whitening. Given the eigendecomposition $\mathrm{Cov}(\mathbf{X}) = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$, the whitening transform $\mathbf{Q} = \mathbf{\Lambda}^{-1/2}\mathbf{V}^\top$ yields $\widetilde{\mathbf{X}} = \mathbf{QX}$ with $\mathrm{Cov}(\widetilde{\mathbf{X}}) = \mathbf{I}$ [Hyvärinen et al., 2001]. After whitening, the unmixing operation $\mathbf{B}\widetilde{\mathbf{X}}$ must satisfy $\mathrm{Cov}(\mathbf{B}\widetilde{\mathbf{X}}) = \mathbf{BB}^\top = \mathbf{I}$, so $\mathbf{B}$ lies in Orthogonal Group $\mathrm{O}(d)$. The original mixing matrix decomposes as $\mathbf{A} = \mathbf{Q}^{-1}\mathbf{B}$, so ICA on whitened data is lossless [Cardoso, 2022].

A.2 THE CLT MOTIVATION FOR NON-GAUSSIANITY

The Central Limit Theorem (CLT) provides the following intuition: a normalized sum of independent non-Gaussian variables converges to a Gaussian as the number of terms grows [Hyvärinen et al., 2001]. In ICA, any linear mixture $\mathbf{b} \cdot \mathbf{S}$ with multiple nonzero weights is therefore *more* Gaussian than the individual sources. Recovering an independent source means finding the projection that is *least* Gaussian, motivating the maximization of non-Gaussianity as the ICA objective.

A.3 INFORMATION GEOMETRY: CONNECTING INDEPENDENCE AND NON-GAUSSIANITY

Following Cardoso [2022], ICA can be understood geometrically via two Pythagorean identities. Let $G(\cdot) = D_{\mathrm{KL}}(\cdot \,\|\, \mathcal{N}(\mathrm{Cov}\,\cdot))$ denote non-Gaussianity and $C(Y) = D_{\mathrm{KL}}(\mathcal{N}(\mathrm{Cov}\,Y) \,\|\, \mathcal{N}(\mathrm{diag}\,\mathrm{Cov}\,Y))$ denote correlation.

Product manifold identity. For any product distribution $Q = \prod_i q_i$:

$$D_{\mathrm{KL}}(P_Y \| Q) = \mathscr{I}(Y) + \sum_i D_{\mathrm{KL}}(P_{Y_i} \| q_i). \tag{8}$$

Gaussian manifold identity. For any Gaussian $\mathcal{N}(\Sigma)$:

$$D_{\mathrm{KL}}(P_Y \| \mathcal{N}(\Sigma)) = G(Y) + D_{\mathrm{KL}}(\mathcal{N}(\mathrm{Cov}\,Y) \| \mathcal{N}(\Sigma)). \tag{9}$$

Evaluating the KL divergence from P_Y to $\mathcal{N}(\mathrm{diag}\,\mathrm{Cov}\,Y)$ via both paths yields Cardoso's identity:

$$\mathscr{I}(Y) + \sum_i G(Y_i) = G(Y) + C(Y). \tag{10}$$

After whitening, $\mathrm{Cov}(Y) = \mathbf{I}$ forces $C(Y) = 0$, giving $\mathscr{I}(Y) = G(Y) - \sum_i G(Y_i)$. Since $G(Y)$ is invariant under orthogonal rotation, minimizing $\mathscr{I}(Y)$ (independence) equals maximizing $\sum_i G(Y_i)$ (total marginal non-Gaussianity). This is Theorem 2.

A.4 OPTIMAL TRANSPORT AS THE NON-GAUSSIANITY CONTRAST

Classical ICA algorithms approximate $G(Y_i)$ with proxies. OT-ICA instead measures it directly as the cost of transporting the empirical marginal to a standard Gaussian.

Definition 1 (Coupling and Push-Forward)**.** *A* coupling *of distributions μ and ν is a joint distribution π on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and ν; the set of all such couplings is $\Pi(\mu, \nu)$. A measurable map $T\colon \mathbb{R}^d \to \mathbb{R}^d$* pushes forward *$\mu$ to ν (written $T_{\#}\mu = \nu$) if $\nu(B) = \mu(T^{-1}(B))$ for every Borel set B.*

The p-Wasserstein distance is the minimum transport cost over all couplings:

$$W_p(\mu,\nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int \|x-y\|^p \, d\pi(x,y) \right)^{1/p}. \tag{11}$$

The structure of the optimal coupling π varies qualitatively depending on whether the marginals are discrete, continuous, or mixed (Peyré and Cuturi [2019]). In 1D, Brenier's theorem guarantees the optimal coupling is induced by the monotone quantile map $T = F_\nu^{-1} \circ F_\mu$, so $W_2^2(\mu,\nu) = \int_0^1 |F_\mu^{-1}(t) - F_\nu^{-1}(t)|^2 \, dt$.

Taking $\nu = \Gamma \equiv \mathcal{N}(0,1)$, the quantity $W_2^2(\mathbb{P}_{\mathbf{b}^\top \widetilde{\mathbf{X}}}, \Gamma)$ measures how far the projection $\mathbf{b}^\top \widetilde{\mathbf{X}}$ is from Gaussian. Unlike negentropy proxies, W_2^2 is a true metric, geometry-aware, and computable by sorting without density estimation. OT-ICA maximizes this over $\mathbf{b}$ in Orthogonal Group $\mathrm{O}(d)$; Theorem 3 shows the maximum is attained at pure source directions.

B MATHEMATICAL PROOFS FOR OT-ICA BOUNDS

B.1 PROOF OF THEOREM 3: W_2^2 ICA CONTRAST

Theorem 3 (W_2^2 ICA Contrast; Johnson and Samworth 2005)**.** *Let $\mathbf{Z} = (Z_1, \ldots, Z_d)^\top$ be centered independent sources with unit variance and at most one Z_i Gaussian. Assume they all have a smooth and strictly positive density. Then the following bound holds for any normalized weight vector α with at least two nonzero components*

$$W_2^2(\alpha \cdot \mathbf{Z}, \Gamma) < \sum_{i=1}^d \alpha_i^2 W_2^2(Z_i, \Gamma). \tag{12}$$

Strategy. We first establish a weak upper bound on $W_2^2(\alpha \cdot \mathbf{Z}, \Gamma)$ (used as an intermediate step), then show the bound is strict for any proper mixture under the non-Gaussianity conditions, proving Theorem 3.

Step 1: Upper Bound (intermediate).

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:

$$X = \sum_{i=1}^d \alpha_i T_i(N_i). \tag{13}$$

To compare this mixture to a Gaussian distribution, we construct a target variable Y using the same underlying noise vector $\mathbf{N}$:

$$Y = \sum_{i=1}^d \alpha_i N_i. \tag{14}$$

Given that the weight vector has unit norm ($\sum_i \alpha_i^2 = 1$), the resulting variable Y follows the standard normal distribution, $Y \sim \Gamma$. The pair (X, Y) creates a valid coupling. Because W_2^2 is the infimum over all valid couplings, the optimal distance is less than or equal to the cost of our constructed coupling:

$$W_2^2(\alpha \cdot \mathbf{Z}, \Gamma) \leq \mathbb{E}\left[|X-Y|^2\right] = \mathbb{E}\left[\left(\sum_{i=1}^d \alpha_i \big(T_i(N_i) - N_i\big) \right)^2 \right]. \tag{15}$$

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

$$|X-Y|^2 = \sum_{i=1}^d \alpha_i^2 \big(T_i(N_i) - N_i\big)^2 + \sum_{i \neq j} \alpha_i \alpha_j \big(T_i(N_i) - N_i\big)\big(T_j(N_j) - N_j\big). \tag{16}$$

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

$$\mathbb{E}\left[|X-Y|^2\right] = \sum_{i=1}^{d} \alpha_i^2 \mathbb{E}\left[\left(T_i(N_i) - N_i\right)^2\right]. \tag{17}$$

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:

$$W_2^2(\alpha \cdot \mathbf{Z}, \Gamma) \leq \sum_{i=1}^{d} \alpha_i^2 W_2^2(Z_i, \Gamma). \tag{18}$$

Step 2: Strict Inequality (proof of Theorem 3). 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 requires the gradients to be parallel at every point in the probability space:

$$\nabla X(\mathbf{N}) = \lambda(\mathbf{N}) \nabla Y(\mathbf{N}). \tag{19}$$

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 differentiate both sides to compute their Hessian matrices. On the LHS, the cross-derivatives $\frac{\partial^2 X}{\partial N_i \partial N_j}$ are zero, yielding a diagonal matrix of $\alpha_i T_i''(N_i)$. On the RHS, we differentiate the product $\lambda(\mathbf{N})\mathbf{b}$, yielding a Rank-1 outer product matrix:

$$\underbrace{\begin{bmatrix} \alpha_1 T_1''(N_1) & 0 & \cdots & 0 \\ 0 & \alpha_2 T_2''(N_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \alpha_d T_d''(N_d) \end{bmatrix}}_{\text{Hessian of } X \text{ (Diagonal)}} = \underbrace{\begin{bmatrix} \alpha_1 \frac{\partial \lambda}{\partial N_1} & \cdots & \alpha_1 \frac{\partial \lambda}{\partial N_d} \\ \vdots & \ddots & \vdots \\ \alpha_d \frac{\partial \lambda}{\partial N_1} & \cdots & \alpha_d \frac{\partial \lambda}{\partial N_d} \end{bmatrix}}_{\text{Outer Product } \mathbf{b}(\nabla\lambda)^\top \text{ (Rank-1)}} \tag{20}$$

Examining any off-diagonal entry ($i \neq j$), the LHS dictates it must be zero, establishing $\alpha_i \frac{\partial \lambda}{\partial N_j} = 0$. Because we assume a mixture where multiple weights are non-zero ($\alpha_i \neq 0$), it follows that 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 indicates the optimal transport maps are linear functions of the form:

$$T_i(x) = \lambda x + c. \tag{21}$$

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. 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.

$$W_2^2(\alpha \cdot \mathbf{Z}, \Gamma) < \sum_{i=1}^{d} \alpha_i^2 W_2^2(Z_i, \Gamma). \tag{22}$$

□

B.2 PROOF OF COROLLARY 2: SYMMETRIC JOINT MAXIMIZATION

Proof. Let $\mathbf{B} \in \mathrm{O}(d)$ be an orthogonal matrix with rows $\mathbf{b}_i^\top$. Because $\mathbf{B}$ is orthogonal, its squared elements sum to 1 across both rows ($\sum_{j=1}^{d} B_{ij}^2 = 1$) and columns ($\sum_{i=1}^{d} B_{ij}^2 = 1$). By the intermediate step of Theorem 3, the contrast for each row projection is bounded by:

$$W_2^2(\mathbf{b}_i^\top \mathbf{Z}, \Gamma) \leq \sum_{j=1}^{d} B_{ij}^2 W_2^2(Z_j, \Gamma). \tag{23}$$

Summing this inequality over all d rows yields an upper bound on the total marginal contrast:

$$\begin{aligned}\sum_{i=1}^{d} W_2^2(\mathbf{b}_i^\top \mathbf{Z}, \Gamma) &\leq \sum_{i=1}^{d}\sum_{j=1}^{d} B_{ij}^2 W_2^2(Z_j, \Gamma) \\ &= \sum_{j=1}^{d}\left(\sum_{i=1}^{d} B_{ij}^2\right) W_2^2(Z_j, \Gamma) \\ &= \sum_{j=1}^{d} W_2^2(Z_j, \Gamma).\end{aligned} \tag{24}$$

The maximum possible sum equals the unweighted sum of the pure sources' contrasts. Because Theorem 3 establishes that the inequality is strict whenever a row mixes multiple sources, the global equality holds if and only if every row $\mathbf{b}_i$ contains exactly one non-zero entry (which must be ± 1 due to the unit norm constraint). Thus, the sum is maximized if and only if $\mathbf{B}$ is a signed permutation matrix. □

C ALGORITHMIC ENHANCEMENTS AND METHODOLOGY

This appendix explains the additional ingredients used in the OT-ICA algorithm.

C.1 ANALYTICAL GAUSSIAN TARGETS

Discrete approximation of the target Gaussian distribution introduces approximation noise. To eliminate this noise, we replace point-sampled targets with an analytical formulation, computing the expected value of a standard normal variable within each discrete quantile bin.

Definition 2 (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 CDF. The analytical target value T_i for the i-th sorted sample is:*

$$T_i = N \int_{z_{i-1}}^{z_i} x\phi(x)\, dx = N\big(\phi(z_{i-1}) - \phi(z_i)\big). \tag{25}$$

C.2 RIEMANNIAN GRADIENT AND RETRACTION

Because the input data is whitened, the unmixing matrix $\mathbf{B}$ must remain in Orthogonal Group $\mathrm{O}(d)$, satisfying $\mathbf{B}\mathbf{B}^\top = \mathbf{I}$. Computing the standard Euclidean gradient $\mathbf{G} = \nabla_{\mathbf{B}} W_2^2$ points into the unconstrained ambient space.

Definition 3 (Riemannian Gradient on Orthogonal Group $\mathrm{O}(d)$)**.** *Given the Euclidean gradient $\mathbf{G}$, the Riemannian gradient resulting from the projection onto the tangent space $T_{\mathbf{B}}\mathrm{O}(d)$ of Orthogonal Group $\mathrm{O}(d)$ is given by [Absil et al., 2008]:*

$$\nabla_{\mathrm{O}(d)}\mathbf{B} = \mathbf{G} - \frac{1}{2}\big(\mathbf{G}\mathbf{B}^\top + \mathbf{B}\mathbf{G}^\top\big)\mathbf{B}. \tag{26}$$

Definition 4 (Symmetric Decorrelation Retraction)**.** *To map the matrix back to the orthogonal surface after a tangent step ($\mathbf{B}_{step}$), we compute the overlap covariance $\mathbf{C} = \mathbf{B}_{step}\mathbf{B}_{step}^\top$ and apply the inverse square root:*

$$\mathbf{B}_{new} = \mathbf{C}^{-1/2}\mathbf{B}_{step}. \tag{27}$$

C.3 CONTINUOUS SMOOTHING VIA GAUSSIAN DITHERING

Applying continuous optimal transport directly to discrete mixtures introduces non-smooth step-functions. Adapted from signal processing [Schuchman, 1964], we inject continuous, zero-mean Gaussian noise with variance $\sigma^2_{\text{dither}} = 0.01$ into the 1D projected components prior to sorting. This decorrelates deterministic quantization errors and convolves the discrete empirical distribution with a Gaussian kernel.

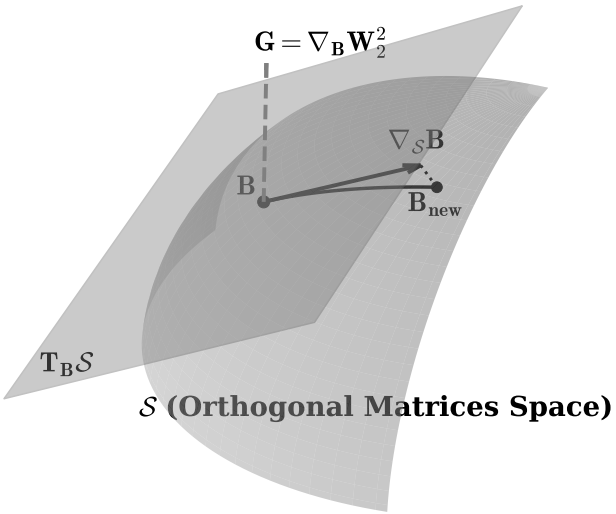

Figure 3: *Geometric sketch of optimization on Orthogonal Group $\mathrm{O}(d)$. The Euclidean gradient $\mathbf{G}$ is projected onto the tangent space to yield the Riemannian gradient $\nabla_{\mathrm{O}(d)}\mathbf{B}$. A retraction maps the estimate back to the orthogonal surface.*

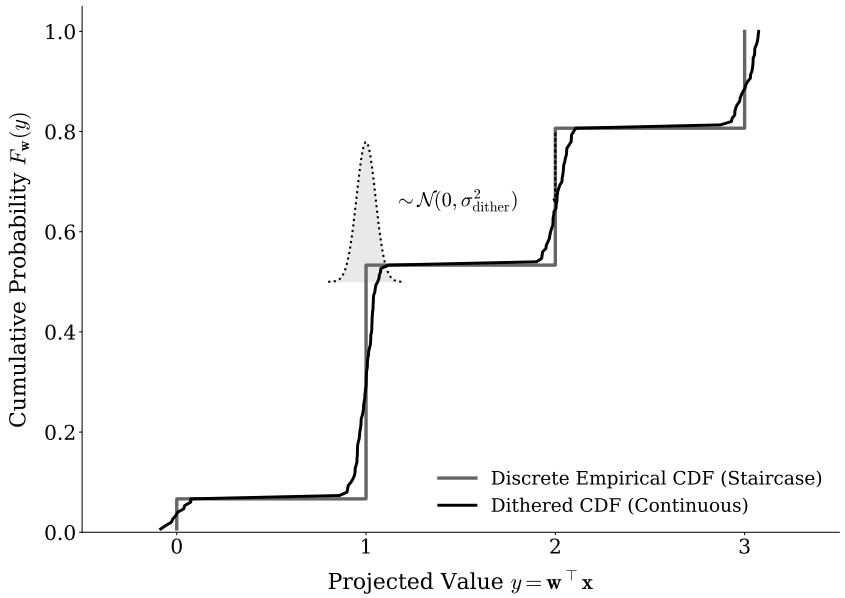

Figure 4: *The non-differentiable staircase of a discrete empirical CDF (gray) is convolved with a narrow Gaussian kernel via dithering, yielding a continuous, differentiable curve (black).*

C.4 TOTAL COMPLEXITY AND STATISTICAL EFFICIENCY

For a dataset of dimension d with N samples, K random restarts (batched into a tensor $\mathbf{B}_{\text{batch}}$), and T optimization iterations, the per-iteration complexity is:

$$\mathcal{O}(K \cdot (d \cdot N \log N + d^3)) \tag{28}$$

The $N \log N$ term represents the sorting of projections, while the d^3 term represents the cost of the symmetric decorrelation used for retraction.

C.5 THE OT-ICA ALGORITHM

The mutual orthogonality of ICs in the whitened space admits an iterative extraction strategy in which each component is found in the orthogonal complement of previously extracted ones. We implement this as *symmetric joint optimization*: all rows of $\mathbf{B}$ are updated simultaneously under the $\mathrm{O}(d)$ constraint via symmetric decorrelation retraction. Symmetric optimization is preferred because it eliminates error accumulation across extraction steps and ensures mutual orthogonality of all components is maintained at every iteration rather than enforced only by projection.

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 N_{batch}, dither noise σ_{dither}
Ensure: Estimated unmixing matrix $\mathbf{B}_{\text{final}} \in \mathbb{R}^{d \times d}$

Phase 1: Preprocessing & Target Generation
Center the data: $\mathbf{X} \leftarrow \mathbf{X} - \mathbb{E}[\mathbf{X}]$
Whiten the data: $\widetilde{\mathbf{X}} \leftarrow (\mathbf{X}\mathbf{X}^\top)^{-1/2}\mathbf{X}$
Compute analytical target $\mathbf{T} \in \mathbb{R}^{N_{\text{batch}}}$ via analytical Gaussian CDF integration
Phase 2: Initialization
Initialize $\mathbf{B} \in \mathbb{R}^{d \times d}$ randomly from $\mathcal{N}(0, 1)$
Apply symmetric decorrelation: $\mathbf{B} \leftarrow (\mathbf{B}\mathbf{B}^\top)^{-1/2}\mathbf{B}$
Phase 3: Stochastic Riemannian Optimization
for $k = 1$ to K **do**
    Sample stochastic mini-batch $\widetilde{\mathbf{X}}_{\text{batch}} \in \mathbb{R}^{d \times N_{\text{batch}}}$ from $\widetilde{\mathbf{X}}$
    Project data onto candidates: $\mathbf{Y} \leftarrow \mathbf{B}\widetilde{\mathbf{X}}_{\text{batch}}$
    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: $\mathcal{L} = \frac{1}{N_{\text{batch}}} \sum_{i=1}^{d} ||\mathbf{Y}^{(i)}_{\text{sorted}} - \mathbf{T}||_2^2$
    Backpropagate to compute Euclidean gradient: $\mathbf{G} = \nabla_{\mathbf{B}}\mathcal{L}$
    Project to tangent space of Orthogonal Group $\mathrm{O}(d)$: $\nabla_{\mathrm{O}(d)}\mathbf{B} = \mathbf{G} - \frac{1}{2}\big(\mathbf{G}\mathbf{B}^\top + \mathbf{B}\mathbf{G}^\top\big)\mathbf{B}$
    Update matrix along tangent plane: $\mathbf{B}_{\text{step}} = \mathbf{B} + \eta\nabla_{\mathrm{O}(d)}\mathbf{B}$
    Retract to Orthogonal Group $\mathrm{O}(d)$: $\mathbf{B} \leftarrow (\mathbf{B}_{\text{step}}\mathbf{B}_{\text{step}}^\top)^{-1/2}\mathbf{B}_{\text{step}}$
end for
Phase 4: Finalization
$\mathbf{B}_{\text{final}} \leftarrow \mathbf{B}(\mathbf{X}\mathbf{X}^\top)^{-1/2}$
return $\mathbf{B}_{\text{final}}$

D EXPERIMENTAL METHODOLOGY AND METRICS

D.1 THE GENERALIZED PERMUTATION MATRIX AND AMARI INDEX

To empirically evaluate the success of source separation, we quantify the distance between the estimated unmixing matrix $\mathbf{B}_{\text{est}}$ and the true mixing matrix $\mathbf{A}$. Because ICA is subject to scaling and permutation ambiguities, the target recovery is a generalized permutation matrix. Let $\mathbf{P} = \mathbf{B}_{\text{est}}\mathbf{A}$ be the global transfer matrix. For a $d \times d$ matrix $\mathbf{P}$ with elements p_{ij}, the

Amari Performance Index [Amari et al., 1996] measures the structural divergence:

$$E(\mathbf{P}) = \frac{1}{2d}\sum_{i=1}^{d}\left(\sum_{j=1}^{d}\frac{|p_{ij}|}{\max_k |p_{ik}|} - 1\right) + \frac{1}{2d}\sum_{j=1}^{d}\left(\sum_{i=1}^{d}\frac{|p_{ij}|}{\max_k |p_{kj}|} - 1\right). \tag{29}$$

D.2 COMPUTATIONAL RESOURCE REGIMES

Because OT-ICA utilizes stochastic gradient descent while FastICA employs fixed-point Newton iterations, we standardize resources via two defined regimes to ensure fairness:

- **Low Compute Baseline:** OT-ICA is allocated restarts $\min(4 \times d, 150)$ and 150/200 deflation/symmetric phase iterations. FastICA is allocated 50 random restarts and 1,000 maximum iterations per component.
- **High Compute Regime:** OT-ICA is allocated restarts $\min(15 \times d, 600)$ and 300/500 iterations. FastICA is allocated 50 restarts and 5,000 maximum iterations per component.

D.3 METHODOLOGY VALIDATION: GLOBAL TRANSFER MATRICES

We validated OT-ICA on a 4D mixture of Laplace, Uniform, Student-$t(3)$, and Beta$(0.5, 0.5)$ sources with $N = 10{,}000$, mixed by the random matrix $\mathbf{A}$ below. The global transfer matrices $\mathbf{P} = \hat{\mathbf{B}}\mathbf{A}$ for both algorithms are near-permutation, with Amari errors below 0.02.

True Mixing Matrix A

$$\begin{bmatrix} 1.058 & 1.520 & -0.249 & 1.012 \\ 0.042 & -0.486 & 0.388 & -0.523 \\ -0.154 & -1.572 & -0.304 & -1.234 \\ -0.006 & 0.503 & 0.053 & 0.928 \end{bmatrix}$$

OT-ICA Transfer Matrix P

$$\begin{bmatrix} -0.012 & -0.009 & -1.000 & -0.003 \\ 0.000 & 0.003 & -0.001 & 1.000 \\ -0.007 & 1.000 & 0.000 & 0.012 \\ -1.000 & -0.007 & 0.005 & -0.010 \end{bmatrix}$$

FastICA Transfer Matrix P

$$\begin{bmatrix} -1.000 & -0.012 & 0.003 & -0.008 \\ 0.012 & -1.000 & 0.001 & -0.004 \\ -0.002 & -0.010 & -0.001 & -1.000 \\ -0.010 & -0.010 & -1.000 & -0.001 \end{bmatrix}$$

Amari Error: OT-ICA $= 0.0155$ FastICA $= 0.0188$

Table 3: *Global transfer matrices for the 4D validation mixture. Off-diagonal entries are* ≤ 0.012 *in absolute value for both algorithms, confirming near-perfect source recovery.*

D.4 FULL BENCHMARK RESULTS

Table 4 reports mean Amari errors for all five methods across all mixture regimes at $d \in \{10, 20, 30\}$ ($N = 10{,}000$, 10 trials). Values at 1.500 indicate total separation failure (clip ceiling). **Bold** marks the lowest error per row. The threshold $E < 0.3$ separates meaningful separation from failure: for a d-dimensional random orthogonal matrix, the expected Amari error approaches $\frac{d-1}{d}$ (e.g., ≈ 0.97 at $d = 30$), so differences among methods above $E = 0.3$ reflect degree of failure rather than quality of separation; fine-grained comparisons in that region are nonetheless preserved for completeness.

D.5 DISCRETE DISTRIBUTIONS: CONTRAST VS. OPTIMIZATION

The *Discrete Only* failure is mechanistically distinct from the continuous failure modes above. Table 5 compares the non-Gaussianity resolution of W_2^2 against the logcosh proxy negentropy $(\mathbb{E}[G(x)] - \mathbb{E}[G(\nu)])^2$ across standard and highly non-Gaussian parameterizations.

Regime	d	OT-ICA	FastICA	JADE	InfoMax	Picard
Continuous Only	10	**0.050**	0.092	0.146	0.100	0.093
	20	**0.106**	0.183	0.321	0.188	0.184
	30	**0.160**	0.271	0.489	0.277	0.271
Strictly Super-Gaussian	10	**0.057**	0.089	0.146	0.103	0.090
	20	**0.113**	0.183	0.316	0.190	0.184
	30	**0.176**	0.280	0.482	0.286	0.280
Zero Gaussian	10	**0.042**	0.118	0.152	0.105	0.121
	20	**0.103**	0.327	0.366	0.315	0.403
	30	**0.207**	0.510	0.686	0.549	0.628
Full Hybrid	10	**0.059**	0.255	0.216	0.297	0.277
	20	**0.261**	0.451	0.571	0.558	0.555
	30	**0.335**	0.671	0.938	0.765	0.917
Discrete Only	10	0.706	0.613	**0.443**	0.757	0.835
	20	1.331	1.402	**1.267**	1.495	1.440
	30	1.500	1.500	1.500	1.500	1.500
Pure Laplace ($\mu = 0, , b = 1/\sqrt{2}$)	10	**0.077**	0.080	0.136	0.083	0.080
	20	**0.165**	0.170	0.285	0.177	0.170
	30	0.280	**0.261**	0.439	0.272	0.262
Pure Uniform ($-\sqrt{3}, , \sqrt{3}$)	10	**0.038**	0.054	0.051	0.056	0.054
	20	**0.094**	0.116	0.110	0.122	0.116
	30	1.500	0.179	**0.171**	0.187	0.180
Pure Student-t ($\nu = 3$)	10	0.068	0.065	0.134	**0.064**	0.065
	20	0.142	**0.135**	0.284	0.136	0.136
	30	0.219	**0.207**	0.436	0.208	0.208
Pure Chi-square $\chi^2(k = 2)$	10	**0.039**	0.094	0.120	0.087	0.094
	20	**0.085**	0.205	0.255	0.190	0.205
	30	**0.129**	0.315	0.392	0.292	0.316
Pure Exponential ($\lambda = 1$)	10	**0.039**	0.094	0.120	0.087	0.094
	20	**0.085**	0.205	0.255	0.190	0.205
	30	**0.129**	0.315	0.392	0.292	0.316

Table 4: *Mean Amari error (↓ better, clipped at 1.500) for all five methods and mixture regimes ($N = 10{,}000$, 10 trials). OT-ICA achieves the lowest error on all five mixed-source configurations. On pure single-distribution sources, OT-ICA leads on Chi-square and Exponential at all d; FastICA leads on Student-t and Laplace ($d = 30$); OT-ICA fails on Uniform at $d = 30$. On Discrete Only, JADE leads at $d \leq 20$; all methods including OT-ICA saturate the clip ceiling at $d = 30$.*

Distribution	$\mathbf{W_2^2}$	Logcosh Negentropy
Laplace (continuous baseline)	0.0398	0.001214
Binomial (standard, $n = 10$)	0.0344	0.000031
Poisson (standard, $\lambda = 3.0$)	0.0513	≈ 0
Binomial (non-Gaussian, $n = 2$)	0.2022	0.000267
Poisson (non-Gaussian, $\lambda = 0.5$)	0.3384	0.000085

Table 5: *W_2^2 provides $\gg 10\times$ greater non-Gaussianity resolution than logcosh on count-based data. Standard Poisson ($\lambda = 3.0$) evaluates to near-zero under logcosh yet retains a clear W_2^2 signal. The failure to unmix discrete sources therefore cannot originate from an insufficient W_2^2 contrast; the issue is in the optimization landscape.*

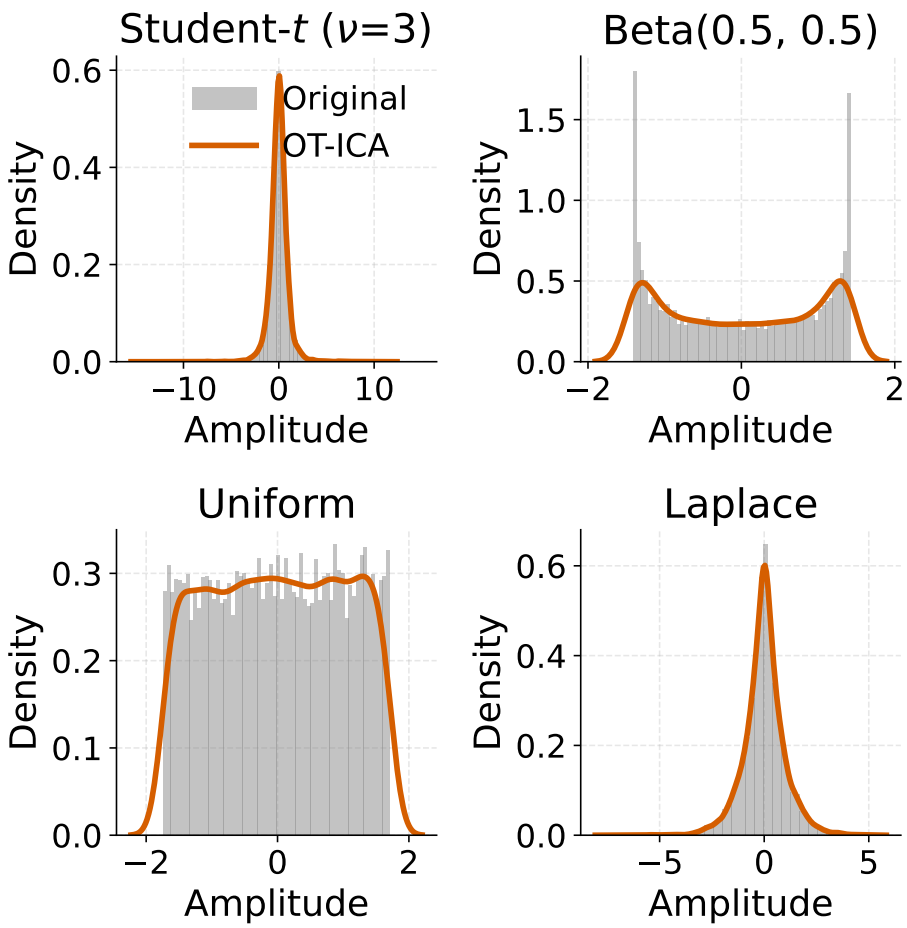

Figure 5: *OT-ICA convergence on four standard continuous source types from the same validation mixture. Estimated latent components are smoothed with kernel density estimation.*

E PROXY CONTRAST BASED ICA ALGORITHMS

This appendix collects a brief overview of the baselines used in the paper.

E.1 FASTICA: NEGENTROPY APPROXIMATION FAILURES

FastICA [Hyvärinen et al., 2001] approximates negentropy using a non-quadratic contrast function $G(\cdot)$, such as the *logcosh* function ($\mathbb{E}[G(x)] - \mathbb{E}[G(\nu)]$, where $\nu \sim \mathcal{N}(0, 1)$). We demonstrate two limitations of this proxy approach.

E.1.1 The Zero Negentropy Condition

If a latent independent source possesses a non-Gaussian distribution such that its expected value under the contrast function equals that of a Gaussian ($\mathbb{E}[G(s_i)] = \mathbb{E}[G(\nu)]$), the approximated negentropy evaluates to zero. In this scenario, the objective function provides no gradient signal, and the algorithm fails to extract the source. As shown in Figure 7, this failure leads to increasing Amari error with increasing dimensions.

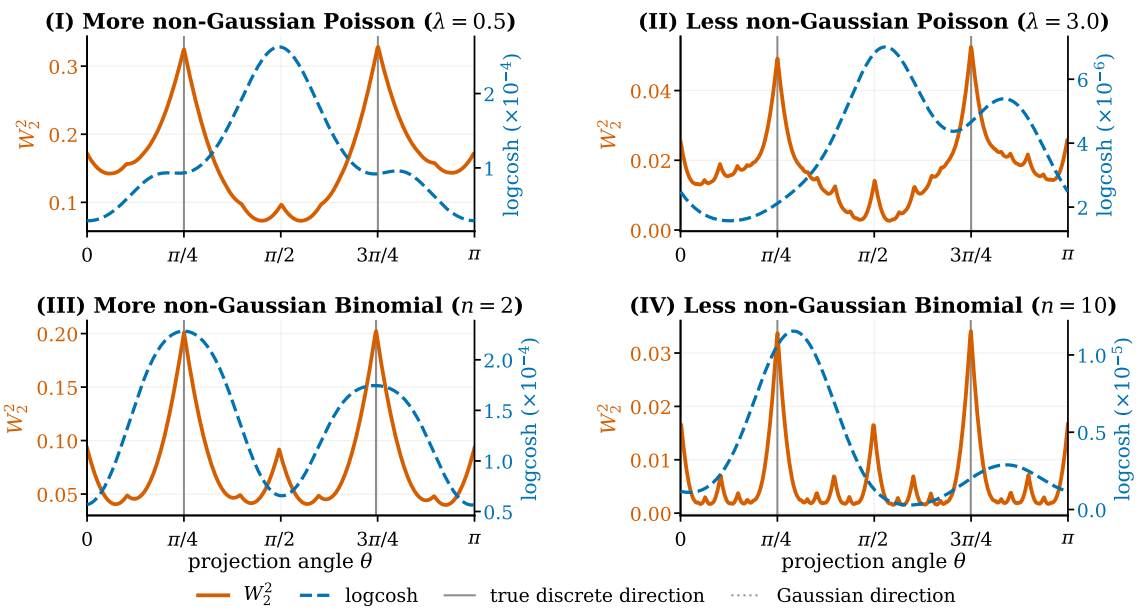

Figure 6: *W_2^2 and logcosh contrast against projection angle for two-source discrete mixtures ($d = 2$, $N = 10{,}000$), evaluated on the whitened data; the true source directions fall at $\pi/4$ and $3\pi/4$ (vertical lines), and both contrasts are π-periodic so the half-period shown is the complete landscape. In every configuration W_2^2 peaks sharply at the true directions, confirming the contrast correctly identifies the sources, while the logcosh proxy's maximum is displaced from them in all but configuration (III) – the one case where FastICA attains its lowest Amari error. Both curves also show fine-grained, jagged local structure away from the peaks, a symptom of the underlying step-function CDFs: these small local plateaus and kinks make the landscape hard for a local gradient-based solver to navigate even where the global maximum is correctly located.*

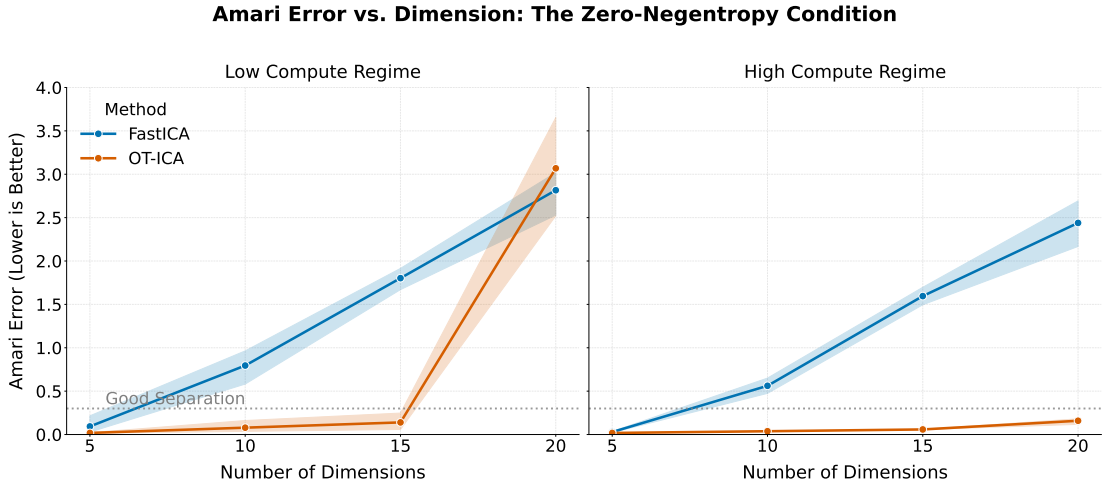

Figure 7: *Amari error comparison for an engineered zero-negentropy distribution across Low and High compute regimes. FastICA's error scales with dimension as the proxy gradient vanishes, crossing the good separation threshold ($E = 0.3$). OT-ICA maintains good separation till higher dimensions depending on compute regime.*

E.1.2 The Vanishing Curvature Condition

FastICA utilizes a Newton fixed-point iteration. Defining $g(x) = G'(x)$ and $g'(x) = G''(x)$, the algorithm approximates the Hessian matrix. For a given component i, the diagonal entry scales according to the expectation $\mathbf{H}_{ii} \approx \mathbb{E}\big[s_i g(s_i) - g'(s_i)\big]$.

The Newton update requires applying the inverse of this Hessian ($\mathbf{H}^{-1}$). If a non-Gaussian source distribution causes the

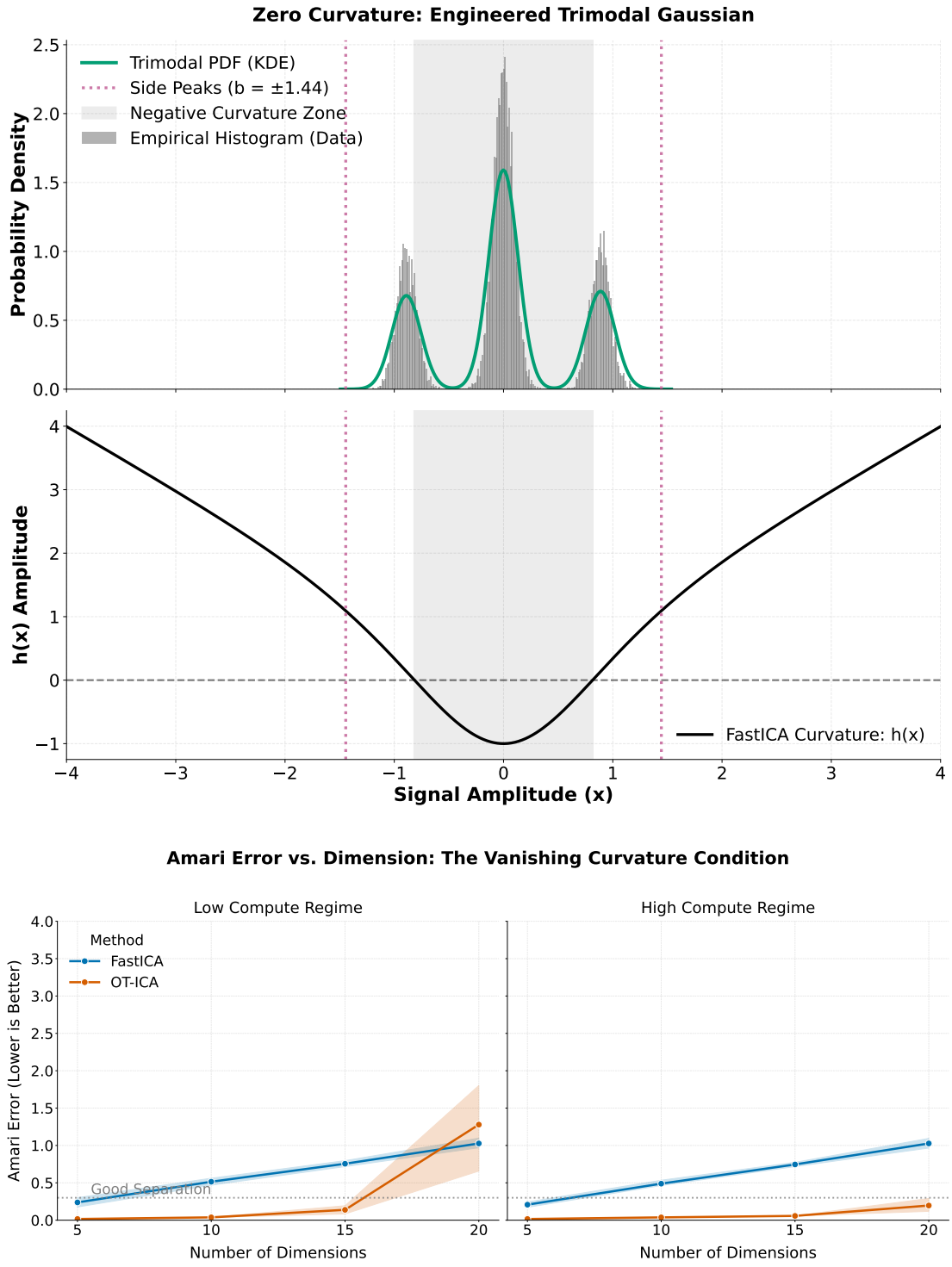

Figure 8: *The Vanishing Curvature failure mode.* ***Top:*** *A trimodal distribution engineered such that negative central curvature and positive outer curvature (side peaks at $b = \pm 1.44$) cancel ($\mathbb{E}[sg(s) - g'(s)] = 0$) while maintaining unit variance.* ***Bottom:*** *Amari error comparison on mixtures containing this distribution. FastICA's Newton solver diverges rapidly as dimension increases across both compute regimes. OT-ICA maintains good separation.*

expectation in the denominator to evaluate to zero ($\mathbb{E}[s_i g(s_i) - g'(s_i)] \to 0$), the inverse becomes undefined [Hyvärinen et al., 2001, Chapter 8]. This analytical divide-by-zero error makes the fixed-point update step intractable. Figure 8 illustrates this failure mode, displaying the specific trimodal distribution that induces zero curvature and the resulting divergence in separation performance compared to our OT-ICA framework.

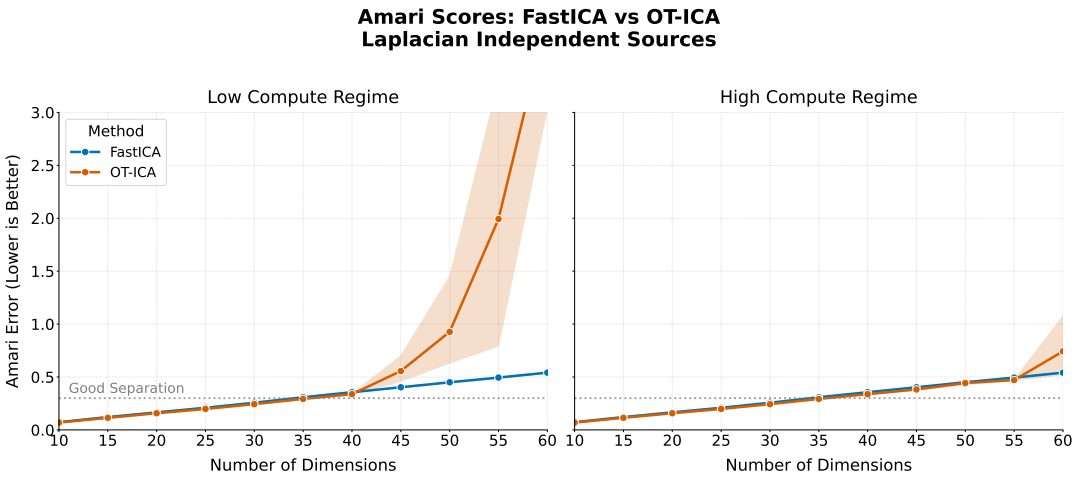

Figure 9: *Amari error across dimensions for Laplacian sources at fixed $N = 10{,}000$ samples. Both OT-ICA and FastICA hit the same curse-of-dimensionality noise ceiling at $d = 40$, a finite-sample sparsity effect shared by all empirical ICA methods at this regime.*

E.2 JADE: SAMPLE COMPLEXITY OF FOURTH-ORDER CUMULANTS

JADE [Cardoso and Souloumiac, 1993] identifies sources by jointly diagonalizing a set of fourth-order cumulant matrices via Jacobi sweeps. Reliable estimation of the full cumulant tensor requires $\mathcal{O}(d^4)$ samples [Cardoso and Souloumiac, 1993, Hyvärinen et al., 2001]. At $d = 30$, this implies approximately $810{,}000$ samples; against the $N = 10{,}000$ available in our benchmark, the cumulant matrices are heavily undersampled. The resulting joint diagonalization converges on estimated tensors reflecting sampling noise rather than true source structure, yielding $E = 0.489$ on continuous sources and $E = 0.938$ on hybrid sources at $d = 30$, both approaching the failure threshold despite JADE's consistency on well-sampled data.

E.3 INFOMAX: SCORE FUNCTION MISSPECIFICATION

InfoMax [Bell and Sejnowski, 1995] maximizes a log-likelihood under a fixed logistic nonlinearity, implicitly treating all sources as sub-Gaussian. On mixtures containing super-Gaussian components (Laplace, Student-t), the score function is misspecified: the gradient points away from the true unmixing direction. Lee et al. [1999] documented this failure mode explicitly, proposing extended InfoMax with online source-type switching. That approach partially addresses the misspecification but requires correct per-component classification, an assumption OT-ICA avoids entirely. In our benchmark, InfoMax produces $E = 0.765$ on hybrid sources and $E = 0.549$ on Zero Gaussian sources at $d = 30$, against OT-ICA's 0.335 and 0.207 respectively, consistent with the misspecified-score explanation.

E.4 PICARD: THE CONTRAST AS THE BINDING CONSTRAINT

Picard [Ablin et al., 2018] replaces InfoMax's gradient ascent with an L-BFGS preconditioned solver on the same log-likelihood objective, achieving provably faster convergence per iteration. In our benchmark, FastICA and Picard produce Amari errors within 0.003 of each other across all continuous configurations at each dimension. This near-identical performance isolates the source of failure: solver speed is not the binding constraint. The shared tanh nonlinearity, and not the optimizer, determines the error floor. Replacing it with an exact, distribution-free contrast reduces error by 40–45% on continuous sources.

F COMPUTATIONAL SCALING AND DIMENSIONALITY LIMITS

We benchmarked OT-ICA against FastICA using a linear mixture of continuous Laplace sources across increasing dimensions with a fixed sample size of $N = 10,000$.

While OT-ICA maintains error rates comparable to FastICA ($E < 0.3$) in lower dimensions, both algorithms exhibit a loss in unmixing quality at $d = 40$. This limit is a manifestation of the curse of dimensionality. The expected error of the empirical Wasserstein distance scales as $\mathcal{O}(N^{-1/d})$. As d grows, a fixed finite sample fails to densely populate the state

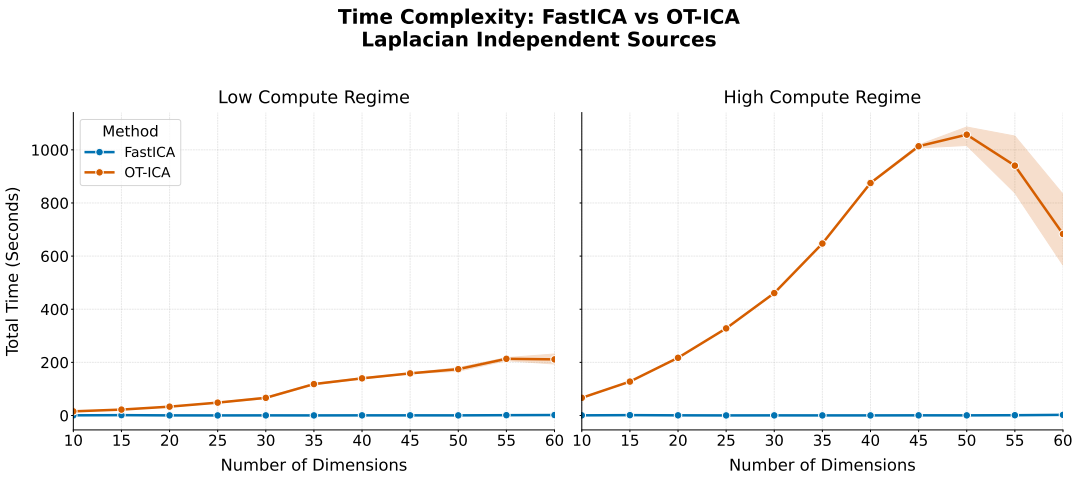

Figure 10: *Total execution time (in seconds) for FastICA and OT-ICA across dimensions.*

space, creating massive empty regions that swallow the gradient signal.

The execution time for OT-ICA scales more steeply than FastICA. FastICA costs $\mathcal{O}(dN)$ per iteration, making it computationally superior for exclusively homogeneous, continuous mixtures that do not trigger its proxy blind spots. OT-ICA incorporates parallel sorting across multiple restarts ($\mathcal{O}(K \cdot dN \log N)$), trading execution speed for reliability on complex topologies.

G PRICE DISCOVERY APPLICATION: DATA GENERATING PROCESS AND IS DEFINITION

G.1 DATA GENERATING PROCESS

We simulate a three-market VECM following the non-Gaussian information-share framework of Zema and Cordoni [2025]. The three markets share a single common efficient price, with long-run impact vector $\psi = (1, 1, 1)^\top$ (each market moves one-for-one with the common trend). The true structural mixing matrix is diagonal:

$$\mathbf{B}_{\text{true}} = \operatorname{diag}\left(\sqrt{0.12},\ \sqrt{0.24},\ \sqrt{0.64}\right), \tag{30}$$

which, as shown below, yields the target Information Shares $[0.12, 0.24, 0.64]$ by construction.

Structural innovations ϵ_t are drawn from Student-t distributions with degrees of freedom $\nu_1 = 5, \nu_2 = 6, \nu_3 = 7$, normalized to unit variance. The VECM reduced-form residuals are $u_t = \mathbf{B}_{\text{true}}\,\epsilon_t$. Heavy tails are a realistic feature of high-frequency financial innovations and are what enables ICA identification: the non-Gaussianity of u_t carries the rotation information that a Gaussian model would lose.

G.2 INFORMATION SHARE DEFINITION

Under Hasbrouck [1995], the Information Share of venue i measures its contribution to price discovery:

$$\text{IS}_i = \frac{(\psi^\top \mathbf{B})_i^2}{\psi^\top \mathbf{B}\mathbf{B}^\top \psi}. \tag{31}$$

With $\psi = (1, 1, 1)^\top$ and diagonal $\mathbf{B}_{\text{true}}$, the numerator equals B_{ii}^2 and the denominator equals $\sum_j B_{jj}^2 = 0.12 + 0.24 + 0.64 = 1$, so $\text{IS}_i = B_{ii}^2 = \{0.12, 0.24, 0.64\}$.

G.3 ESTIMATION PROCEDURE

The reduced-form covariance is $\mathbf{\Omega} = \mathbf{B}\mathbf{B}^\top$. Cholesky or eigendecomposition of $\mathbf{\Omega}$ yields the whitening transform $\mathbf{S} = \mathbf{\Omega}^{1/2}$; the orthogonal rotation $\mathbf{C}$ satisfying $\mathbf{B} = \mathbf{S}\mathbf{C}$ is then identified from the non-Gaussianity of $\mathbf{S}^{-1}u_t$ by OT-ICA. In this

simulation $\boldsymbol{\Omega}$ is treated as known; in practice it is estimated from VECM residuals by OLS. Two economic constraints are imposed post-estimation: the Hungarian algorithm [Kuhn, 1955] resolves the permutation ambiguity by assigning each structural shock to the market for which it explains maximum variance; sign normalization enforces a positive own-market impact. IS estimates are then computed from $\hat{\mathbf{B}} = \mathbf{S}\hat{\mathbf{C}}$ using the Hasbrouck formula above.

H EEG APPLICATION DETAILS

H.1 DATASET AND PREPROCESSING

We use the MNE sample dataset (`sample_audvis_raw.fif`), a publicly available MEG/EEG recording with auditory and visual stimuli. Five frontal EEG channels (EEG 001–005) are extracted and cropped to a 10 s segment (10–20 s from onset). A zero-phase FIR bandpass filter (1–40 Hz, `firwin` design) is applied before ICA to remove DC drift and high-frequency noise. Signals are then z-score normalized per channel: $\tilde{x}_i = (x_i - \mu_i)/\sigma_i$.

H.2 OT-ICA CONFIGURATION

Deflation phase: 50 random restarts, 200 iterations per restart, dither $\sigma = 0.01$. Symmetric refinement phase: 400 gradient steps, learning rate $\eta = 0.05$, mini-batch size 512, dither $\sigma = 0.01$, symmetric decorrelation retraction. No distributional assumptions are imposed; the algorithm operates on the 5-channel whitened signal.

H.3 ARTIFACT IDENTIFICATION AND EVALUATION

Independent components are ranked by excess kurtosis $\kappa = \mathbb{E}[(z-\mu)^4]/\sigma^4 - 3$. The component with the highest κ is designated the ocular artifact; blink artifacts are impulsive (super-Gaussian) and are expected to dominate on frontal channels. Reconstruction quality is evaluated as the RMS amplitude reduction in a ± 250 ms window around the maximum-amplitude blink event, computed as:

$$\text{RMS reduction} = 1 - \frac{\|\mathbf{x}_{\text{clean}}\|_{\text{RMS}}}{\|\mathbf{x}_{\text{raw}}\|_{\text{RMS}}}. \tag{32}$$

The cleaned signal is obtained by zeroing the artifact component in the ICA source space and applying the inverse total unmixing matrix $(\mathbf{W}_{\text{eeg}}\,\mathbf{W}_{\text{white}})^{-1}$.