BEYOND UNIMODAL BASES: PULLBACK GEOMETRY FOR MULTIMODAL DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Data-driven Riemannian geometry provides nonlinear interpolation and geometric representations of high-dimensional data. For these operations to be statistically meaningful, paths between observations should preferentially traverse high-likelihood regions. Existing scalable pullback constructions typically use a unimodal Gaussian latent distribution, assuming that the data reside close to a single manifold. For multimodal data, mapping separated modes or local structures into one Gaussian region can require substantial transport deformation and compromise the resulting geometry.

We introduce a pullback geometry for data supported on mixtures of manifolds. Using a latent Gaussian mixture, we define its Riemannian metric as the matrix square of the responsibility-weighted expected component precision. The metric is smooth and positive definite and recovers the existing Gaussian construction in the single-component limit. For structured overlapping mixtures, we establish conditions under which the log-density is concave along geodesics, providing a formal connection between the proposed geometry and paths through high-likelihood regions, and derive the corresponding local curvature relations.

We instantiate this geometry in a normalizing flow with adaptive mixture learning, allowing the number of active components to emerge from the data and supporting component-wise reconstruction and local effective-dimension estimation. Experiments on synthetic geometric data, a controlled multi-view image setting with a known reference trajectory, and MNIST show reduced transport distortion, competitive path support, close reference-trajectory recovery, and improved interpolation realism. These results extend scalable pullback geometry beyond datasets that reside close to a single manifold while retaining tractable and interpretable local structure.

1 INTRODUCTION

A data-driven Riemannian geometry equips a dataset with learned notions of distance, paths, and local structure, enabling geodesic interpolation, analysis of intrinsic structure, and low-dimensional representation of curved data (Arvanitidis et al., 2018; Diepeveen, 2024). These operations are meaningful only when the geometry reflects the data distribution: geodesics between observations in high-likelihood regions should preferentially traverse other high-likelihood regions, while remaining computationally tractable and stable.

Recent work has scaled data-driven Riemannian geometry through generative modelling (Diepeveen, 2024; Diepeveen et al., 2025). Score-based pullback Riemannian geometry (SBPRG) uses an invertible normalizing flow to map data to a simpler latent distribution, where geometric operations can be computed efficiently and transferred back by pullback. Such a flow defines a diffeomorphism, a smooth invertible map with a smooth inverse, between data and latent space. To retain meaningful geometry, the learned diffeomorphism must also be sufficiently regular; existing constructions encourage it to be approximately locally isometric on the data. Consequently, the flexibility of the flow is restricted, and the choice of latent statistical model directly affects the resulting geometry.

Existing scalable constructions typically use a single Gaussian latent distribution (Diepeveen et al., 2025), effectively assuming that the data reside close to one manifold. For data with multiple modes

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

or local structures, mapping both the observed structures and the low-density regions between them into one Gaussian region can require substantial deformation and compromise the induced geometry.

Figure 1 illustrates this problem on Wavy-X, a synthetic dataset formed by two intersecting curved structures. A single-Gaussian base concentrates these structures into one latent region. A fixed two-component mixture is more flexible but does not alone ensure that the representation or its geodesics reflect the data structure. In contrast, the adaptive mixture recovers two distinct latent directions and produces mapped paths aligned with the observed branches. This motivates jointly learning the multimodal statistical model and its induced geometry.

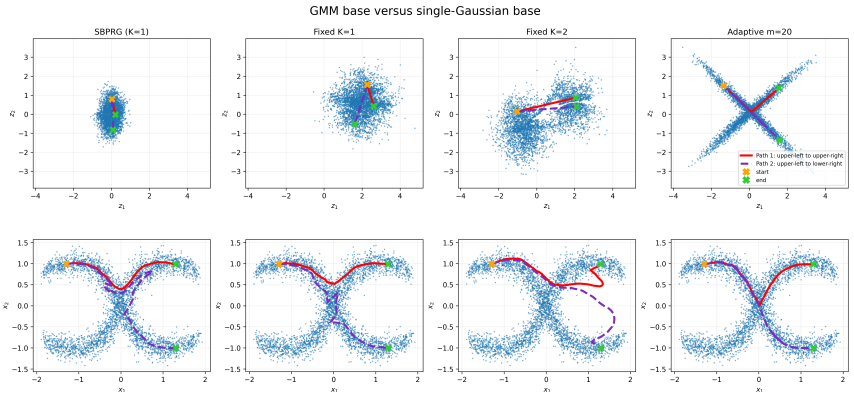


Figure 1: **Motivating Wavy-X example.** Learned latent representations (top) and mapped discrete geodesic approximations (bottom) for two shared endpoint pairs. SBPRG and the fixed single-Gaussian control concentrate the intersecting structures into one latent region. A fixed two-component mixture is more expressive but does not consistently recover paths along the observed branches. The adaptive mixture retains two active components, organizes the data into distinct latent directions, and follows the displayed branches more closely.

Extending scalable pullback geometry to mixtures of manifolds therefore raises three challenges. First, the statistical model must represent multiple modes without forcing them into a single Gaussian latent region. Second, the model must induce a computationally tractable Riemannian metric whose geodesics preferentially pass through high-likelihood regions. Third, because the number of modes is generally unknown, the learning procedure should be agnostic to the number of active components.

We address these challenges through the following contributions:

- We introduce a normalizing flow with an adaptive Gaussian-mixture latent distribution for data supported on mixtures of manifolds.
- We define a responsibility-weighted expected-precision metric that is smooth, positive definite, computationally tractable, and recovers the existing Gaussian construction in the single-component limit. For structured shared-mean mixtures, we establish conditions relating its geodesics to high-likelihood regions and derive local curvature relations.
- We jointly learn the flow and mixture while allowing the number of active components to emerge during training. The responsibilities also support component-wise reconstruction and local effective-dimension estimation. Experiments on synthetic geometric data, a controlled multi-view image dataset, and MNIST (LeCun et al., 1998) evaluate transport distortion, path recovery, and interpolation quality.

Related work is discussed in Appendix A. Section 2 reviews pullback geometry, Section 3 presents the proposed metric and its theoretical properties, and Section 4 describes adaptive learning and geometric computation. Implementation details and additional experiments are provided in the appendix.

2 BACKGROUND

To extend pullback geometry to data supported on mixtures of manifolds, we first recall how the regularity of a normalizing flow and the choice of its latent statistical model determine the resulting data geometry. The central idea of pullback geometry is that, when the learned flow is approximately locally isometric on the data, geometric computations can be performed in a tractable latent space and transferred back to data space with controlled distortion (Diepeveen, 2024; Diepeveen et al., 2025). Consequently, the choice of latent distribution and latent metric directly affects the geometry induced in data space. For a Gaussian latent density $p_Z = \mathcal{N}(\mu, A)$ with covariance $A \succ 0$, the score-based pullback construction uses the constant latent metric

$$g_Z = \big(D\nabla \log p_Z(z)\big)^\top \big(D\nabla \log p_Z(z)\big) = A^{-2}, \tag{1}$$

where $D\nabla \log p_Z$ denotes the Hessian of the log-density.

Its constant form makes geometric operations tractable. However, using a single Gaussian also places the burden of representing multimodal structure on the flow. This creates a tension: statistical expressiveness favours a more flexible transport, whereas the geometric guarantees favour a flow that remains close to a local isometry. Our approach addresses this tension by increasing the expressiveness of the latent base distribution.

A suitable latent metric should be smooth, positive definite, and computationally tractable, while producing geodesics that reflect the latent density. One precise notion of this compatibility is log-concavity of the density along geodesics (Diepeveen et al., 2025). For a geodesic $z(t)$, this requires

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} \log p_Z(z(t)) \leq 0. \tag{2}$$

This condition prevents the log-density from dipping below the linear interpolation of its endpoint values along the geodesic and therefore provides a precise way to study whether paths remain in regions of high likelihood. It cannot hold globally for every multimodal density, particularly when modes are strongly separated. We therefore analyze it for the structured class of overlapping mixtures introduced in Section 3.

3 MULTIMODAL PULLBACK GEOMETRY

We seek a latent model for data supported on mixtures of manifolds and a tractable Riemannian metric whose geodesics preferentially pass through high-likelihood regions. We therefore use a Gaussian-mixture latent distribution. However, the score-based construction does not extend straightforwardly from a single Gaussian to a mixture. Let $z \in \mathbb{R}^d$, let K be the number of components, and let $w_k > 0$, $\sum_k w_k = 1$, with means μ_k and covariances $A_k \succ 0$. For $p_Z(z) = \sum_{k=1}^K w_k \mathcal{N}(z; \mu_k, A_k)$, define the responsibilities and expected precision by

$$\gamma_k(z) = \frac{w_k \mathcal{N}(z; \mu_k, A_k)}{p_Z(z)}, \qquad P(z) = \sum_{k=1}^K \gamma_k(z) A_k^{-1}. \tag{3}$$

The corresponding score-based metric is

$$g_{\mathrm{score}}(z) = \big(D\nabla \log p_Z(z)\big)^\top \big(D\nabla \log p_Z(z)\big) = \big(-P(z) + \mathrm{Cov}_{\gamma(z)}\left[\nabla \log \mathcal{N}(z; \mu_k, A_k)\right]\big)^2. \tag{4}$$

Here, $\mathrm{Cov}_{\gamma(z)}$ denotes covariance over the component index under the responsibilities $\gamma(z)$. Unlike in the single-Gaussian case, this matrix can be singular, producing a degenerate metric. Where responsibilities are shared across components, the covariance term can partially cancel the expected precision. We therefore omit this term and define

$$g_Z(z) = P(z)^2, \tag{5}$$

which is positive-definite. When $K = 1$, or when components are separated far apart but have identical covariance, it reduces to $g_Z = A_1^{-2}$, recovering the metric of Diepeveen et al. (2025). The latter case is important because, for well-separated modes, a path connecting high-density regions

must cross a low-density region, so geodesic log-concavity cannot hold everywhere.[1] Among the possible choices, our proposal recovers prior work and satisfies the desiderata locally within each component.

3.1 STAR-SHAPED DENSITIES

The metric in Equation (5) is defined for general Gaussian mixtures. To analyze its relationship with the latent density explicitly, we consider a structured class in which all component means coincide at the origin and the component precisions share an eigenbasis. This restriction is used only for the following theoretical analysis; the learned experimental mixtures need not satisfy it. The resulting distributions have star-convex superlevel sets, related to the archetypal setting of Diepeveen & Needell (2026).

Theorem 1 (Geodesic log-concavity for Gaussian stars)**.** *For a star with diagonalized covariances, the pointwise log-concavity condition on affine-parametrised geodesic $z(t)$ under $g_Z = P^2$ is*

$$\underbrace{-\dot{z}^\top P \dot{z}}_{\leq 0} + \sum_{1 \leq i < j \leq d} \underbrace{\mathrm{Cov}_{\gamma(z)}[q_{k,i}, q_{k,j}]}_{\textit{precision covariance}} \underbrace{\left(\sqrt{\frac{\rho_i}{\rho_j}} \dot{z}_i z_j - \sqrt{\frac{\rho_j}{\rho_i}} z_i \dot{z}_j \right)^2}_{\geq 0} \leq 0, \tag{6}$$

where $q_{k,i}$ is the i-th diagonal element of A_k^{-1} and $\rho_i = \mathbb{E}_{\gamma(z)}[q_{k,i}] = \sum_k \gamma_k(z) q_{k,i}$.

One immediate consequence is that geodesic log-concavity holds globally if $\mathrm{Cov}_{\gamma(z)}[q_{k,i}, q_{k,j}] \leq 0$ for $i \neq j$. Geodesics along the coordinate axes are also log-concave with respect to the density. Restricting the analysis to a coordinate plane through the origin gives

$$\left(\frac{z_1^2}{\rho_1(z)} + \frac{z_2^2}{\rho_2(z)} \right) \mathrm{Cov}_{\gamma(z)}[q_{k,1}, q_{k,2}] \leq 1. \tag{7}$$

For positive precision covariance, Equation (7) has the form of a generalized ellipse illustrated in Figure 2 with effective radii

$$R_1 = \sqrt{\frac{\rho_1}{\mathrm{Cov}_\gamma[q_{k,1}, q_{k,2}]}}, \qquad R_2 = \sqrt{\frac{\rho_2}{\mathrm{Cov}_\gamma[q_{k,1}, q_{k,2}]}}. \tag{8}$$

Curvature. At the origin, the sectional curvature of the coordinate (i, j)-plane is

$$K_{ij}(0) = \frac{\rho_i + \rho_j}{\rho_i^2 \rho_j^2} \mathrm{Cov}_{\gamma(0)}[q_{k,i}, q_{k,j}], \tag{9}$$

where the ρ_i are evaluated at the origin. Thus, the same precision covariance controls geodesic log-concavity and sectional curvature. At the centre of a Gaussian star, all component scores vanish, so $\nabla P(0) = \nabla g_Z(0) = 0$.

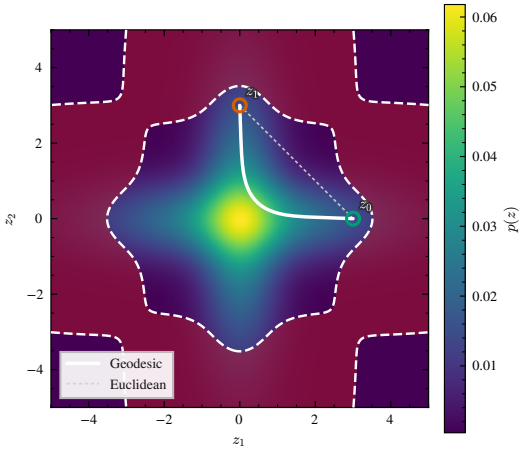


Figure 2: For this $d = 2, K = 3$ mixture of low central covariance (-0.823), a large (non-red) region satisfies geodesic log-concavity; a geodesic follows the arms.

In coordinates normalized such that $P(0) = I$, the local expansion is

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} \log p_Z(z(t)) = -\|u\|_2^2 + \frac{1}{2} \sum_{i,j} C_{ij}^0 u_i^2 z_j^2 + \frac{1}{2} R_0(z, u, u, z) + \mathcal{O}(\|z\|^4 \|u\|^2) \tag{10}$$

where u is the geodesic velocity, $C_{ij}(z) = \mathrm{Cov}_{\gamma(z)}[q_{k,i}, q_{k,j}]$, $C_{ij}^0 = C_{ij}(0)$ and R_0 is the Riemann curvature at the origin. The leading quadratic correction splits into spatial variation of the expected precision and a curvature contribution. The curvature term is generated by cross-coordinate precision covariance.

[1] Viewed via the desiderata, there is no compelling reason a metric should, for example, measure distances between isolated modes to route geodesics between them as a shortest-path algorithm would.

Informing numerical evaluation. We aggregate the sectional curvatures through the Ricci tensor, their standard contraction, to obtain a compact local summary. Its full expression and derivation are provided in Appendix C.3.4. In the diagonal case, it can be computed in $\mathcal{O}(Kd)$ time. Writing $\mathrm{Ric}_\star(0) := \mathrm{Ric}(0)$ for this local structured diagnostic and $r_i = (\mathrm{Ric}_\star(0))_{ii}$, we report the star–Ricci diagnostic

$$\mathcal{L}_\star = \sum_{i=1}^{d} \mathrm{softplus}(r_i + \delta_\star), \tag{11}$$

where $\delta_\star \geq 0$ is a curvature margin; we use $\delta_\star = 0$ in the reported experiments. The quantity is evaluated after training as a local diagnostic. For learned mixtures with untied means, we report it as a structured curvature summary rather than as the exact Ricci curvature of the general learned metric.

Proofs and additional derivations are provided in Appendix C. An additional structured case with separated component means is discussed in Appendix C.4.

4 Stabilized Learning of Multimodal Pullback Geometry

We now address three practical requirements: learning without knowing the mixture complexity, limiting unnecessary transport distortion, and retaining tractable geodesic computation and local dimension reduction. We alternate between updating a normalizing flow and an initially overcomplete Gaussian mixture. Slower mixture updates, damping, and conservative merge–prune operations stabilize training while allowing the number of active components to emerge from the data. Full algorithms and regularizer definitions are provided in Appendix D. We write the learned flow as ϕ_θ and its inverse as ϕ_θ^{-1}.

4.1 Adaptive flow–mixture learning

The flow is updated frequently on minibatches, whereas mixture parameters and structural operations are updated less frequently using aggregated latent statistics. At mixture-update step e, we use temperature-smoothed responsibilities

$$\gamma_k^{(T_e)}(z) = \mathrm{softmax}_k \left(\frac{\log w_k + \log \mathcal{N}(z; \mu_k, A_k)}{T_e} \right), \tag{12}$$

where $T_e > 0$ is annealed to one. Let $\eta_e = \{w_k, \mu_k, A_k\}_{k=1}^{K_e}$ denote the active mixture parameters and η_e^* their responsibility-weighted moment targets. With damping coefficient $\zeta \in (0, 1]$, we update

$$\eta_{e+1} = (1 - \zeta)\eta_e + \zeta \eta_e^*. \tag{13}$$

Here, diagonal variances are blended in log space; full covariances are blended linearly and regularized by a positive diagonal floor.

A symmetric Dirichlet prior regularizes the weights. Its concentration parameter α favours sparse weights when $\alpha < 1$ and balanced weights when $\alpha > 1$. Low-weight components are pruned, and a candidate merge is accepted only when it improves the penalized mixture objective. The MAP-style weight update and complete merge–prune procedure are given in Algorithm 1.

4.2 Training and geometric computation

Training objective. The flow is trained using

$$\mathcal{L}_{\mathrm{train}} = \mathcal{L}_{\mathrm{NLL}} + \lambda_{\mathrm{iso}} \mathcal{L}_{\mathrm{iso}} + \lambda_{\mathrm{metric}} \mathcal{L}_{\mathrm{metric}}, \tag{14}$$

where $\mathcal{L}_{\mathrm{NLL}} = -\mathbb{E}_x \left[\log p_Z(\phi_\theta(x)) + \log |\det D\phi_\theta(x)| \right]$ and $\lambda_{\mathrm{iso}}, \lambda_{\mathrm{metric}} \geq 0$. The isometry term discourages local directional distortion. For MNIST and the BlueWhale rendered-view dataset described in Section 5.3, the metric-smoothness term additionally discourages abrupt changes in $P(z)$ between nearby latent points; otherwise $\lambda_{\mathrm{metric}} = 0$. Both regularizers are defined in Appendix D.2.

For the three-dimensional synthetic experiment (Appendix F.2), MNIST, and BlueWhale, a small mean-shrinkage update moves selected component means toward their weight-averaged centre. This update encourages component overlap and is not an additional term in Equation (14); its definition appears in Appendix D.3.

Metric-based geodesics. Given endpoints $z_0, z_T \in \mathbb{R}^d$, let $z_{0:T} = (z_0, \ldots, z_T)$ be a path with T equal-duration segments. We optimize its interior vertices using

$$\mathcal{E}_{g_Z}(z_{0:T}) = T \sum_{i=0}^{T-1} \left\| P\left(\frac{z_i + z_{i+1}}{2}\right)(z_{i+1} - z_i) \right\|_2^2. \tag{15}$$

Because $g_Z = P^2$, this is the midpoint approximation of the metric energy. The endpoints are fixed, the interior vertices are initialized linearly, and the optimized path is mapped to data space through ϕ_θ^{-1}. Wavy-X uses $T = 80$, whereas MNIST uses $T = 16$. These paths use no density, smoothness, spacing, arclength, or log-concavity penalty. BlueWhale instead uses the support-graph path described in Section 5.3. Algorithm 2 gives the complete continuous solver.

4.3 Component-wise local reconstruction

Let $x_n \in \mathbb{R}^d$ and $r_{nk} = \gamma_k(\phi_\theta(x_n))$. These responsibilities define weighted data-space means $\bar{x}_k$ and covariances C_k^x. If $\nu_{k,1} \geq \cdots \geq \nu_{k,d}$ are the eigenvalues of C_k^x, define

$$d_k = \min\left\{ r : \frac{\sum_{i=1}^{r} \nu_{k,i}}{\sum_{i=1}^{d} \nu_{k,i}} \geq \tau_{\text{var}} \right\}, \tag{16}$$

where $\tau_{\text{var}} \in (0, 1)$ is the variance-retention threshold. The leading d_k eigenvectors define a component-wise linear reconstruction and local effective-dimension estimate. These subspaces are fitted in data space and are not taken from the latent GMM covariances.

5 Experiments

We evaluate whether an adaptive Gaussian-mixture latent base reduces flow distortion and supports data-faithful path computation. The three datasets evaluate complementary aspects of the method: Wavy-X tests transport distortion and path support in two dimensions, BlueWhale provides a metadata-defined reference trajectory, and MNIST evaluates interpolation realism without assuming a ground-truth path. Additional point-cloud and component-wise analyses are reported in Appendices F and G.

5.1 Experimental Setup

Datasets. **Wavy-X** is a two-dimensional synthetic dataset formed by two noisy, sinusoidally perturbed strokes that intersect at the origin and produce four curved half-arms; see Figure 1. Generation details are provided in Appendix E.1.1. **BlueWhale** is a controlled rendered-image dataset whose viewpoint metadata define a reference trajectory through a star-shaped viewpoint family; generation details are given in Appendix E.1.3. **MNIST** is a high-dimensional multimodal benchmark for which we assume no ground-truth interpolation path.

Models and evaluation scope. We compare the adaptive GMM flow with the single-Gaussian score-based pullback Riemannian geometry (SBPRG) baseline (Diepeveen et al., 2025). Here, m is the initial number of components; the fixed-m controls disable adaptation, so $K = m$. The $m = 1$ and $m = 2$ variants separate the effect of using a mixture base from that of adaptive component selection. Straight latent interpolation is included as a path reference where applicable. All models use the same data splits, RealNVP architecture (Dinh et al., 2017), and training budget.

The path procedure depends on the question addressed by each dataset. For Wavy-X, we minimize the expected-precision metric energy in Equation (15) with $T = 80$ segments; MNIST uses the same objective with $T = 16$. BlueWhale instead uses a density-aware, metric-weighted support graph over encoded training samples, allowing direct comparison with its metadata-defined reference trajectory. Construction details are provided in Appendix D.5.

Unless otherwise stated, results are reported for one fixed random seed and should be interpreted descriptively. Star–Ricci quantities are computed only as post-training diagnostics. Although the proposed metric reduces to SBPRG when $K = 1$, our fixed-$m = 1$ control uses the proposed training procedure with component adaptation disabled, whereas SBPRG uses the baseline training procedure. The two models can therefore learn different transport maps.

5.2 WAVY-X: TRANSPORT DISTORTION AND PATH SUPPORT

Wavy-X permits direct inspection of the learned transport and decoded paths. We compare SBPRG, fixed $m = 1$, fixed $m = 2$, and the adaptive GMM initialized with $m = 20$ candidate components. Path 1 joins the upper-left and upper-right half-arms, whereas Path 2 joins the upper-left and lower-right half-arms.

We report test negative log-likelihood (NLL), mean isometry deviation D_{iso}, the 95th percentile (Q95) of the flow-Jacobian condition number κ_J, and the final active-component count K_{active}. We also report the star–Ricci diagnostics $L_\star$ and $\lambda_{\max}(\text{Ric}_\star(0))$, together with path-support deviation. Definitions are provided in Appendix E.2.

Table 1: **Wavy-X quantitative comparison.** Lower is better for NLL, distortion, conditioning, $L_\star$, and support deviation. Paths are densely sampled discrete geodesic approximations under $g_Z = P^2$.

Model	NLL ↓	D_{iso} ↓	Q95 κ_J ↓	K_{active}	$L_\star$ ↓	$\lambda_{\max}(\text{Ric}_\star(0))$	Path 1 support ↓	Path 2 support ↓
SBPRG ($K = 1$)	1.5777	2.9788	40.9951	1	1.3863	0.0000	0.3262	0.3320
Fixed $m = 1$	1.6078	9.1065	28.2596	1	1.3863	≈ 0	0.3396	**0.2905**
Fixed $m = 2$	1.6893	10.2015	28.0523	2	1.3863	≈ 0	1.1756	2.5638
Adaptive $m = 20$	**1.5590**	**1.8943**	**4.1490**	2	$\mathbf{6.47\times10^{-17}}$	**-37.2766**	**0.2794**	0.3158

The adaptive model obtains the lowest NLL, isometry deviation, and upper-tail Jacobian conditioning. It also gives the lowest support deviation for Path 1, while fixed $m = 1$ is best for Path 2. Because support deviation measures proximity to the data cloud, it does not penalize path folding or distinguish between equally supported routes. The fixed-$m = 2$ results show that introducing a mixture base alone does not guarantee better paths.

The adaptive model reduces 20 candidate components to two. Its largest star–Ricci eigenvalue is -37.28, yielding $L_\star \approx 0$. These quantities are structured post-training diagnostics rather than the exact Ricci curvature of the learned untied-mean mixture.

Appendix F.1.2 shows how the same samples are progressively transformed from data space to latent space. Compared with the baselines, the adaptive model better preserves the two intersecting structures while requiring less transport deformation, consistent with the distortion measurements. Component-wise analysis recovers the two crossing arm pairs and assigns each effective dimension one, with 99.41% and 99.29% explained variance; see Appendix F.1.

5.3 BLUEWHALE: RECOVERY OF A DATA-SUPPORTED REFERENCE TRAJECTORY

BlueWhale consists of rendered views with viewpoint metadata defining a reference trajectory through the common junction of two viewpoint arms. Unlike Wavy-X, the encoded observations form three thin branches meeting at a junction that is not locally a smooth manifold. Continuous optimization under a smooth Riemannian metric can therefore cut across the junction or between nearby branches instead of following the observed structure.

We instead compute a shortest path on a mutual k-nearest-neighbour graph whose vertices are encoded training observations. Edge costs combine the proposed metric with a latent-density penalty, producing a density-aware, metric-weighted support-graph path. Construction details are given in Appendix D.5.

We compare the adaptive support-graph path with the decoded SBPRG path and straight latent interpolation. Because the adaptive path consists of observed samples, this experiment evaluates recovery of a data-supported reference trajectory rather than providing a controlled comparison of continuous geodesic solvers.

Figure 3 shows that the adaptive representation admits a support-constrained route aligned with the progressive viewpoint change in the reference sequence. The continuously decoded SBPRG and straight-latent paths contain visibly blurred intermediate frames.

At matched sequence positions, we report mean squared error (MSE), mean absolute error (MAE), mean structural similarity index (SSIM), and worst-step MSE, defined as the largest frame-wise MSE. Metric and sequence alignment details are provided in Appendix E.2.

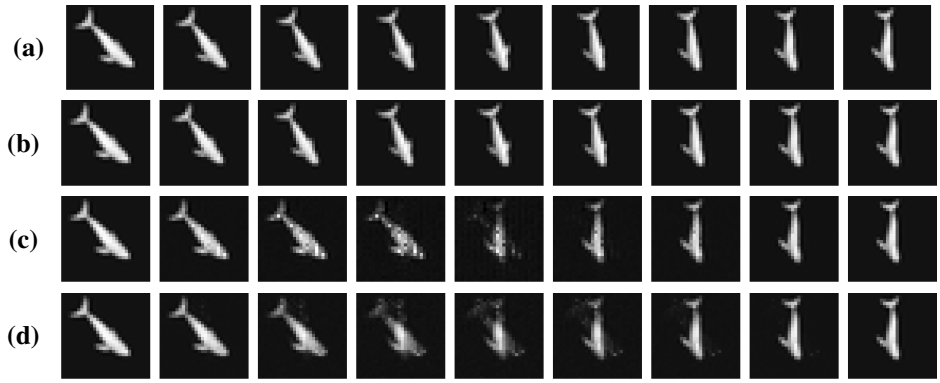


Figure 3: **BlueWhale viewpoint paths.** Image sequences between the same endpoints: **(a)** the metadata-defined reference trajectory; **(b)** the adaptive-GMM support-graph path; **(c)** the decoded SBPRG path; and **(d)** straight latent interpolation. Panel **(b)** consists of observed samples, whereas panels **(c)** and **(d)** are continuously decoded.

Table 2: **BlueWhale path comparison.** Image-space agreement with the metadata-defined reference sequence. The adaptive path is constrained to observed samples and is therefore not directly comparable to the continuously decoded alternatives.

Path	MSE ↓	MAE ↓	Mean SSIM ↑	Worst-step MSE ↓
Adaptive-GMM support graph	0.000306	0.003897	0.993647	0.001257
SBPRG decoded path	0.010547	0.032986	0.835206	0.022814
Straight latent path	0.007188	0.026982	0.859611	0.013496

The support-graph path obtains the lowest MSE and MAE, the highest mean SSIM, and the lowest worst-step MSE relative to the reference sequence. This demonstrates reference-trajectory recovery on the encoded sample support but does not isolate the contribution of the continuous expected-precision path objective. Component-wise reconstruction and effective-dimension results are reported in Appendix G.1.

5.4 MNIST: Interpolation Realism and Perceptual Variation

MNIST provides a multimodal image benchmark without a prescribed ground-truth path. We compare the adaptive GMM, initialized with $m = 30$ candidate components, against SBPRG and straight latent interpolation. Digit labels are not used to train the generative models; they are used only to select held-out different-digit endpoint pairs and to train the independent evaluation classifier. Paths use $T = 16$ segments.

We measure interior-sample realism using the Fréchet distance (Feature FD) between all interior path samples and held-out MNIST images in the feature space of the evaluation classifier. We also report perceptual path length (PPL), the accumulated feature-space distance between consecutive samples, and relative perceptual-distance variance (PDV_{rel}), which measures variation in perceptual step size. Definitions are given in Appendix E.2.

Table 3: **MNIST path interpolation.** Classifier-feature Fréchet distance is the primary measure of interior-sample realism. The adaptive GMM improves this measure at the cost of greater perceptual path length and step-size variation.

Method	Feature FD ↓	PPL ↓	PDV_{rel} ↓
Straight latent	265.218	104.667	0.166
SBPRG	311.615	**101.823**	**0.120**
Adaptive GMM (ours)	**144.432**	164.663	0.426

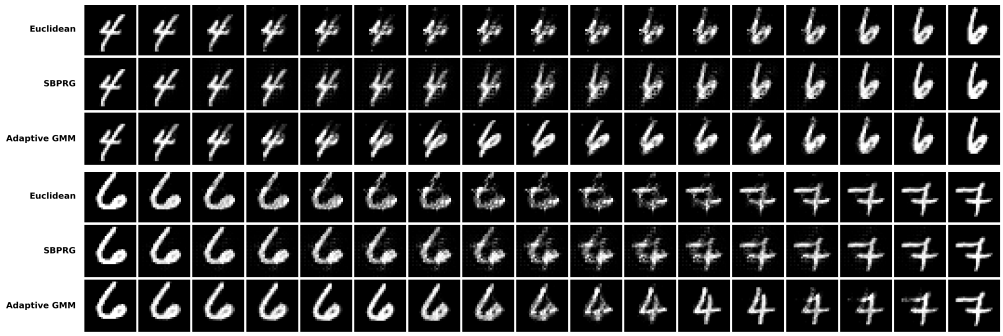


Figure 4: **MNIST path interpolation** ($T = 16$; rows: straight latent / SBPRG / adaptive GMM). **Top:** $4 \to 6$. The adaptive sequence remains recognizable, whereas the alternatives contain visibly noisy intermediate frames. **Bottom:** $6 \to 7$. The adaptive path passes through a digit-4-like region of the learned representation.

The adaptive model obtains the lowest classifier-feature Fréchet distance, indicating that its interior samples are closer to the held-out image distribution in the chosen feature space. Figure 4 is consistent with this result: the adaptive sequences remain recognizable, and the $6 \to 7$ example passes through a digit-4-like region.

The higher PPL and PDV_{rel} show that the adaptive paths are longer and less uniform in perceptual feature space. The results therefore indicate a trade-off: the adaptive geometry favours paths through more realistic intermediate images, whereas SBPRG produces shorter and more uniform perceptual transitions.

The component-wise analysis retains 13 active components with effective dimensions ranging from 17 to 56; reconstruction examples are provided in Appendix G.2.

Summary. In this run, the adaptive model improves transport conditioning and gives the best measured support for one Wavy-X path while remaining competitive for the other. Its BlueWhale representation supports a graph path closely aligned with the metadata-defined reference trajectory. On MNIST, its metric-energy paths improve interior-sample classifier-feature distance at the cost of greater perceptual length and step-size variation. Together, the experiments evaluate complementary aspects of the proposed geometry using dataset-appropriate path procedures.

6 Conclusion

Extracting a tractable geometry from a multimodal statistical model requires more than fitting the data distribution: the induced paths should pass preferentially through high-likelihood regions while permitting stable geometric computation. We addressed this challenge using an adaptive Gaussian-mixture latent distribution and a smooth expected-precision metric that recovers the single-Gaussian construction. For shared-mean Gaussian stars, we related geodesic log-concavity to precision covariance and local curvature.

Synthetic and controlled settings provide a necessary intermediate step toward more complex applications because they allow transport distortion, path support, and reference-path recovery to be evaluated directly. Across Wavy-X, BlueWhale, and MNIST, the results show reduced transport distortion, competitive path support, close reference-trajectory recovery, and improved interpolation realism. The star–Ricci quantity is used here as a post-training diagnostic; studying its role as a regularizer on more complex data, together with more scalable flow architectures, provides a direction for future work.

AI Use Statement

The authors used Claude (Anthropic) for text revision, proofreading, and code documentation. The authors assume responsibility for all content.

Reproducibility statement

The code used to produce all experimental results in this paper will be made publicly available soon. Implementation and evaluation details are provided in the appendix. Measurements are generated with a fixed seed to ensure consistent evaluation and reproducibility.

References

Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: On the curvature of deep generative models. In *International Conference on Learning Representations*, 2018.

Jacob Bamberger, Adam Gosztolai, Pierre Vandergheynst, Michael Bronstein, and Iolo Jones. Riemannian metric matching for scalable geometric modeling of distributions. *arXiv preprint arXiv:2606.14334*, 2026. URL https://arxiv.org/abs/2606.14334.

Willem Diepeveen. Pulling back symmetric Riemannian geometry for data analysis. *arXiv preprint arXiv:2403.06612*, 2024. URL https://arxiv.org/abs/2403.06612.

Willem Diepeveen and Deanna Needell. Manifold learning with normalizing flows: Towards regularity, expressivity and iso-Riemannian geometry. *arXiv preprint arXiv:2505.08087*, 2025. URL https://arxiv.org/abs/2505.08087.

Willem Diepeveen and Deanna Needell. Riemannian archetypal analysis: Interpretable non-linear data analysis on deformed star distributions. *arXiv preprint arXiv:2605.24113*, 2026.

Willem Diepeveen and Melanie Weber. Iso-Riemannian optimization on learned data manifolds. *arXiv preprint arXiv:2510.21033*, 2025. URL https://arxiv.org/abs/2510.21033.

Willem Diepeveen, Georgios Batzolis, Zakhar Shumaylov, and Carola-Bibiane Schönlieb. Score-based pullback Riemannian geometry: Extracting the data manifold geometry using anisotropic flows. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 13746–13773. PMLR, 2025. URL https://proceedings.mlr.press/v267/diepeveen25a.html.

Nat Dilokthanakul, Pedro A. M. Mediano, Marta Garnelo, Matthew C. H. Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with Gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017.

Pavel Izmailov, Polina Kirichenko, Marc Finzi, and Andrew Gordon Wilson. Semi-supervised learning with normalizing flows. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4615–4630, 2020.

Iolo Jones. Manifold diffusion geometry: Curvature, tangent spaces, and dimension. *SIAM Journal on Mathematics of Data Science*, 8(2):257–277, 2026. doi: 10.1137/24M1719219. URL https://doi.org/10.1137/24M1719219.

Dimitris Kalatzis, Johan Ziruo Ye, Jesper Wohlert, and Søren Hauberg. Multi-chart flows. *arXiv preprint arXiv:2106.03500*, 2021.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Hang Shao, Abhishek Kumar, and P. Thomas Fletcher. The Riemannian geometry of deep generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 315–323, 2018.

Peter Sorrenson, Daniel Behrend-Uriarte, Christoph Schnörr, and Ullrich Koethe. Learning distances from data using density-based geometry. *arXiv preprint arXiv:2407.09297*, 2024.

Tianchun Wang, Farzaneh Mirzazadeh, Xiang Zhang, and Jie Chen. GC-flow: A graph-based flow network for effective clustering. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 36157–36173, 2023.

Hanlin Yu, Søren Hauberg, Marcelo Hartmann, Arto Klami, and Georgios Arvanitidis. Learning geometry and topology via multi-chart flows. In *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 2962–2970. PMLR, 2026. URL https://proceedings.mlr.press/v300/yu26c.html.

A Related Work

The central problem is not only to fit a flexible generative distribution, but to extract from it a geometry that supports statistically meaningful and computationally tractable operations. To further put this problem into perspective, it is worthwhile to highlight that the quality of the resulting geometry directly affects downstream tasks such as nonlinear interpolation, dimension reduction, and optimization over learned data manifolds. Existing work addresses complementary parts of this problem: generative-geometric methods construct metrics from learned models, pullback methods improve scalability, and multimodal models increase statistical expressiveness. We review these directions to clarify why their connection remains unresolved.

Data-driven geometry and downstream tasks. Deep generative models can induce Riemannian metrics that account for distortion between latent and data space, providing parametrization-aware distances and nonlinear interpolation paths (Arvanitidis et al., 2018; Shao et al., 2018). Once a reliable geometry has been learned, it can support nonlinear representation learning through Riemannian autoencoders, as well as barycentre computation, clustering, and optimization over learned data manifolds (Diepeveen et al., 2025; Diepeveen & Weber, 2025). These applications depend on the quality of the statistical and geometric models: a metric that assigns low cost to unsupported directions can produce misleading paths, representations, or optimization trajectories. Diffusion-based geometry provides an alternative route to estimating dimension, tangent spaces, curvature, and related quantities directly from samples (Jones, 2026; Bamberger et al., 2026). Together, these works show why a reliable geometric foundation is necessary before downstream manifold operations can be trusted.

Scalable pullback geometry. Early generative-geometric methods demonstrated the usefulness of Riemannian latent spaces but can be expensive to evaluate in high dimensions (Arvanitidis et al., 2018). Recent pullback constructions instead use invertible generative maps to perform geometric computations in a tractable latent space and transfer them to data space (Diepeveen, 2024; Diepeveen et al., 2025). SBPRG is the closest foundation for our work: an anisotropic Gaussian base induces a constant precision metric and supports scalable geodesic computation, dimension estimation, and Riemannian autoencoding. Its statistical model, however, remains unimodal. Consequently, scalability alone does not address data containing multiple modes or local manifold-like structures.

Multimodal modelling and density-aware geometry. Gaussian-mixture priors have been used in variational autoencoders and normalizing flows for clustering, semi-supervised learning, and density estimation (Dilokthanakul et al., 2016; Izmailov et al., 2020; Wang et al., 2023). These methods improve statistical expressiveness but do not by themselves specify the Riemannian metric needed for geometric data analysis. Multi-chart flows use collections of local maps to represent nontrivial structures (Kalatzis et al., 2021; Yu et al., 2026); in contrast, we retain one invertible flow and derive a spatially varying metric from its Gaussian- mixture base.

A complementary line of work modifies distances or geometric mappings to improve their relation to the data distribution. Density-based distances increase the cost of low-probability regions (Sorrenson et al., 2024), while iso-Riemannian methods control transport distortion and the reliability of geometric operations within a global pullback construction (Diepeveen & Needell, 2025). Deformed star distributions provide a closely related structured setting (Diepeveen & Needell, 2026). These works improve transitions between regions or control transport regularity; we address an earlier modelling question: when the data contain multiple modes or local manifold-like structures, what Riemannian metric should be extracted from the underlying multimodal statistical model? We answer this by constructing a positive-definite metric from a Gaussian-mixture base, analyzing its compatibility with the density for a structured class, and learning the active components adaptively.

B Notation and Geometric Background

This section collects the differential- and Riemannian-geometric notation used throughout the paper. We follow the conventions of Diepeveen & Needell (2026), extended here to include the curvature quantities required in Section 3.

Smooth maps and differentials. Let $\mathcal{X}$ and $\mathcal{Z}$ be smooth manifolds of dimension d, with tangent spaces $T_x\mathcal{X}$ and $T_z\mathcal{Z}$. For a smooth map $\phi : \mathcal{X} \to \mathcal{Z}$, its differential at x is the linear map

$$D_x\phi : T_x\mathcal{X} \longrightarrow T_{\phi(x)}\mathcal{Z}. \tag{17}$$

Because the spaces considered in this paper are Euclidean, we identify their tangent spaces with $\mathbb{R}^d$. In coordinates, $D_x\phi$ is the Jacobian matrix of ϕ at x. For a smooth scalar function $h : \mathbb{R}^d \to \mathbb{R}$, we write ∇h for its Euclidean gradient and

$$D\nabla h = \nabla^2 h \tag{18}$$

for its Hessian.

Riemannian metrics and curves. A Riemannian metric g assigns to each point $z \in \mathcal{Z}$ a smoothly varying inner product

$$g_z : T_z\mathcal{Z} \times T_z\mathcal{Z} \longrightarrow \mathbb{R}. \tag{19}$$

In Euclidean coordinates, we identify g_z with a smooth symmetric positive-definite matrix $g(z)$ and write

$$g_z(u, v) = u^\top g(z)v. \tag{20}$$

For a continuously differentiable curve $\gamma : [0, 1] \to \mathcal{Z}$, its Riemannian length and energy are

$$L_g(\gamma) = \int_0^1 \sqrt{\dot\gamma(t)^\top g(\gamma(t))\dot\gamma(t)}\, dt, \tag{21}$$

$$E_g(\gamma) = \frac{1}{2}\int_0^1 \dot\gamma(t)^\top g(\gamma(t))\dot\gamma(t)\, dt. \tag{22}$$

A geodesic between fixed endpoints is a locally length-minimizing curve; under constant-speed parametrization it is equivalently a stationary point of E_g.

Levi–Civita connection and geodesics. Every Riemannian metric determines a unique torsion-free, metric-compatible connection ∇, called the Levi–Civita connection. In local coordinates, its Christoffel symbols are

$$\Gamma^k_{ij} = \frac{1}{2}g^{k\ell}\left(\partial_i g_{j\ell} + \partial_j g_{i\ell} - \partial_\ell g_{ij}\right), \tag{23}$$

where $(g^{ij}) = g^{-1}$ and repeated indices are summed. A smooth curve is a geodesic when

$$\ddot\gamma^k(t) + \Gamma^k_{ij}(\gamma(t))\dot\gamma^i(t)\dot\gamma^j(t) = 0. \tag{24}$$

Pullback geometry. Let $\phi : \mathcal{X} \to \mathcal{Z}$ be a diffeomorphism and let g_Z be a Riemannian metric on $\mathcal{Z}$. The pullback metric on $\mathcal{X}$ is defined by

$$(\phi^* g_Z)_x(u, v) = (g_Z)_{\phi(x)}\left(D_x\phi[u], D_x\phi[v]\right). \tag{25}$$

In matrix form,

$$g_X(x) = D_x\phi^\top g_Z(\phi(x))D_x\phi. \tag{26}$$

Thus, $\phi : (\mathcal{X}, g_X) \to (\mathcal{Z}, g_Z)$ is an isometry by construction. In particular, the length of a curve in data space equals the length of its image in latent space:

$$L_{g_X}(\gamma) = L_{g_Z}(\phi \circ \gamma). \tag{27}$$

Consequently, latent geometric computations can be transferred to data space using ϕ^{-1}. In this paper, ϕ is the learned invertible normalizing flow and

$$g_Z(z) = P(z)^2 \tag{28}$$

is the expected-precision metric.

Riemannian Hessian and geodesic log-concavity. For a smooth scalar function $h : \mathcal{Z} \to \mathbb{R}$, its Riemannian Hessian is

$$\mathrm{Hess}_g\, h(X, Y) = g\left(\nabla_X \operatorname{grad}_g h, Y\right). \tag{29}$$

If γ is a geodesic, then

$$\frac{d^2}{dt^2}h(\gamma(t)) = \mathrm{Hess}_g\, h_{\gamma(t)}\left(\dot\gamma(t), \dot\gamma(t)\right). \tag{30}$$

We say that a positive density p is log-concave along a geodesic γ when

$$\frac{d^2}{dt^2}\log p(\gamma(t)) \le 0 \tag{31}$$

along the path. This is the notion of geodesic log-concavity used in Section 3.

Riemann curvature. We use the curvature convention

$$R(X,Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z. \tag{32}$$

For linearly independent tangent vectors $u, v \in T_z\mathcal{Z}$, the sectional curvature of the plane they span is

$$K_z(u,v) = \frac{g_z(R(u,v)v,u)}{g_z(u,u)g_z(v,v) - g_z(u,v)^2}. \tag{33}$$

For the coordinate plane spanned by the i-th and j-th coordinate vectors, we write this quantity as $K_{ij}(z)$.

The Ricci tensor is the trace of the curvature operator:

$$\mathrm{Ric}_z(u,v) = \mathrm{tr}\left[w \longmapsto R(w,u)v\right]. \tag{34}$$

Equivalently, for any g_z-orthonormal basis $\{e_1, \ldots, e_d\}$,

$$\mathrm{Ric}_z(u,v) = \sum_{a=1}^{d} g_z\left(R(e_a,u)v, e_a\right). \tag{35}$$

The Ricci tensor therefore contracts the collection of sectional curvatures into a symmetric bilinear form with only d^2 coordinate entries. Appendix C.3.4 uses this contraction to obtain a computationally tractable local curvature diagnostic.

Matrix notation. For the curvature convention in Equation (32), we use the associated four-tensor

$$R(X,Y,Z,W) = g(R(X,Y)Z,W). \tag{36}$$

For a square matrix M, we write

$$\mathrm{Sym}(M) = \frac{1}{2}(M + M^\top). \tag{37}$$

The relations $M \succ 0$ and $M \succeq 0$ denote positive definiteness and positive semidefiniteness, respectively. Expectations, variances, and covariances indexed by $\gamma(z)$ are taken with respect to the discrete distribution over mixture components with probabilities $\{\gamma_k(z)\}_{k=1}^K$.

C Geometric Derivations

Here we derive the geometric results used in Section 3.

C.1 Gaussian-mixture score Hessian

Let the mixture be $p_Z(z) = \sum_{k=1}^K w_k \mathcal{N}(z; \mu_k, A_k)$, with $\ell_k(z) = \log \mathcal{N}(z; \mu_k, A_k)$, and define the component score function $s_k(z) = \nabla \ell_k(z) = -A_k^{-1}(z - \mu_k)$. The responsibilities are, by definition,

$$\gamma_k(z) = \frac{w_k e^{\ell_k(z)}}{\sum_j w_j e^{\ell_j(z)}},$$

and hence

$$\nabla \log p_Z(z) = \sum_k \gamma_k(z) s_k(z).$$

Differentiating responsibilities gives

$$\nabla \gamma_k(z) = \gamma_k(z)\left(s_k(z) - \sum_j \gamma_j(z) s_j(z)\right).$$

Since $Ds_k(z) = -A_k^{-1}$,

$$\begin{aligned} D\nabla \log p_Z(z) &= \sum_k \gamma_k(z) Ds_k(z) + \sum_k s_k(z) \nabla \gamma_k(z)^\top \\ &= -\sum_k \gamma_k(z) A_k^{-1} + \mathrm{Cov}_{\gamma(z)}[s_k(z)]. \end{aligned} \tag{38}$$

Writing $P(z) = \sum_k \gamma_k(z) A_k^{-1}$, we obtain

$$D\nabla \log p_Z(z) = -P(z) + \mathrm{Cov}_{\gamma(z)}\left[\nabla \log \mathcal{N}(z; \mu_k, A_k)\right].$$

C.2 VALIDITY OF THE EXPECTED-PRECISION METRIC

Proposition 1. *Suppose $w_k > 0$ and $A_k \succ 0$ for all k. Then $g_Z(z) = P(z)^2$, where $P(z) = \sum_{k=1}^K \gamma_k(z) A_k^{-1}$, defines a smooth Riemannian metric on $\mathbb{R}^d$.*

Proof. As each Gaussian is smooth and strictly positive, $p_Z(z) > 0$ and the responsibilities $\gamma_k(z)$ are smooth. Thus P and $g_Z = P^2$ are smooth. For $v \neq 0$, $v^\top P(z) v = \sum_k \gamma_k(z)\, v^\top A_k^{-1} v > 0$, hence $P(z) \succ 0$. Since $P(z)$ is symmetric, $v^\top g_Z(z) v = v^\top P(z)^2 v = \|P(z)v\|_2^2 > 0$. Hence $g_Z(z) \succ 0$. □

Remark 1 (Single component case)**.** *For $K = 1$ we have $P(z) = A_1^{-1}$, and $g_Z(z) = A_1^{-2}$, which is the single-Gaussian score-based pullback metric.*

C.3 GAUSSIAN STARS

Assume that all component means coincide at the origin and that the component precisions share an eigenbasis. In this basis, $A_k^{-1} = \mathrm{diag}(q_{k,1}, \ldots, q_{k,d})$. Now, define $\rho_i(z) = \mathbb{E}_{\gamma(z)}[q_{k,i}]$, and $C_{ij}(z) = \mathrm{Cov}_{\gamma(z)}[q_{k,i}, q_{k,j}]$. Then $P(z) = \mathrm{diag}(\rho_1, \ldots, \rho_d)$, and $g_{ii}(z) = \rho_i(z)^2$.

Lemma C.1. *For a Gaussian star, $\partial_j \log p_Z(z) = -\rho_j(z) z_j$, and $\partial_j \rho_i(z) = -z_j C_{ij}(z)$. Consequently, $\partial_{ij} \log p_Z(z) = -\rho_i(z)\delta_{ij} + C_{ij}(z) z_i z_j$.*

Proof. The j-th coordinate of the k-th component score is $\partial_j \ell_k(z) = -q_{k,j} z_j$. Averaging over responsibilities gives $\partial_j \log p_Z(z) = -\rho_j(z) z_j$. Moreover, $\partial_j \gamma_k(z) = -\gamma_k(z)\big(q_{k,j} - \rho_j(z)\big) z_j$. Therefore

$$\partial_j \rho_i(z) = \sum_k q_{k,i} \partial_j \gamma_k(z) = -z_j \,\mathrm{Cov}_{\gamma(z)}[q_{k,i}, q_{k,j}],$$

and differentiating $\partial_i \log p_Z = -\rho_i z_i$ gives the last identity. □

C.3.1 GEODESIC LOG-CONCAVITY

Theorem 1 (Geodesic log-concavity for Gaussian stars (restated))**.** *Let $z(t)$ be an affinely parameterized geodesic of $g_Z = P^2$, with $u = \dot{z}$. Then*

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} \log p_Z(z(t)) = -u^\top P(z) u + \sum_{1 \le i < j \le d} C_{ij}(z) \left(\sqrt{\frac{\rho_i(z)}{\rho_j(z)}}\, u_i z_j - \sqrt{\frac{\rho_j(z)}{\rho_i(z)}}\, z_i u_j \right)^2. \tag{39}$$

Proof. For the diagonal metric $g_{ii} = \rho_i^2$, clearly $g^{ii} = \rho_i^{-2}$. The Christoffel symbols are now $\Gamma^\ell_{ab} = \delta_{b\ell} \frac{\partial_a \rho_\ell}{\rho_\ell} + \delta_{a\ell} \frac{\partial_b \rho_\ell}{\rho_\ell} - \delta_{ab} \frac{\rho_a \partial_\ell \rho_a}{\rho_\ell^2}$. Set $f = \log p_Z$. Along geodesics in affine parametrization, $\frac{\mathrm{d}^2}{\mathrm{d}t^2} f(z(t)) = u^\top \nabla^2 f\, u - \partial_\ell f\, \Gamma^\ell_{ab} u_a u_b$. The density Hessian yields

$$u^\top \nabla^2 f\, u = -\sum_i \rho_i u_i^2 + \sum_{a,b} C_{ab} z_a z_b u_a u_b.$$

Using $\partial_\ell f = -\rho_\ell z_\ell$ and $\partial_a \rho_\ell = -z_a C_{\ell a}$, the connection term yields

$$-\partial_\ell f\, \Gamma^\ell_{ab} u_a u_b = -2 \sum_{a,b} C_{ab} z_a z_b u_a u_b + \sum_{a,b} C_{ab} \frac{\rho_a}{\rho_b} z_b^2 u_a^2,$$

and, hence,

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} f(z(t)) = -\sum_i \rho_i u_i^2 + \sum_{a,b} C_{ab} \frac{\rho_a}{\rho_b} z_b^2 u_a^2 - \sum_{a,b} C_{ab} z_a z_b u_a u_b.$$

The diagonal covariance terms cancel. Each off-diagonal pair $i < j$ contributes

$$C_{ij} \left(\frac{\rho_i}{\rho_j} u_i^2 z_j^2 + \frac{\rho_j}{\rho_i} u_j^2 z_i^2 - 2 u_i u_j z_i z_j \right),$$

or, equivalently,

$$C_{ij}\left(\sqrt{\frac{\rho_i}{\rho_j}}\,u_i z_j - \sqrt{\frac{\rho_j}{\rho_i}}\,z_i u_j\right)^2.$$

Substituting gives Equation (39). □

If $C_{ij}(z) \le 0$ for every $i \neq j$, each pairwise correction in Equation (39) is nonpositive; hence the log-density is concave along every geodesic.

In two dimensions, let $C = C_{12}$. For $C > 0$, the correction term is the rank-one quadratic form $C(a^\top u)^2$, where $a = \begin{pmatrix} \sqrt{\rho_1/\rho_2}\,z_2 \\ -\sqrt{\rho_2/\rho_1}\,z_1 \end{pmatrix}$. Satisfying geodesic log-concavity for all velocities u is hence equivalent to $Caa^\top \preceq P$. Since $P \succ 0$, this is again equivalent to $C\,a^\top P^{-1} a \le 1$, which gives

$$\left(\frac{z_1^2}{\rho_1(z)} + \frac{z_2^2}{\rho_2(z)}\right) C_{12}(z) \le 1.$$

The coordinate axes provide an additional case in which strict geodesic log-concavity holds.

Proposition 2 (Coordinate-axis geodesics)**.** *For a diagonalized Gaussian star, each coordinate axis is invariant under the geodesic flow of $g_Z = P^2$. Every nonconstant geodesic contained in the r-th coordinate axis satisfies*

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\log p_Z(z(t)) = -\rho_r(z(t))\dot{z}_r(t)^2 < 0.$$

Proof. Let e_r be the r-th standard basis vector and set $z = se_r$. For $\ell \neq r$,

$$\Gamma^\ell_{rr} = -\frac{\rho_r \partial_\ell \rho_r}{\rho_\ell^2}.$$

Since $\partial_\ell \rho_r = -z_\ell \operatorname{Cov}_{\gamma(z)}[q_{k,r}, q_{k,\ell}]$, we have $\Gamma^\ell_{rr} = 0$ on the r-th axis. A geodesic initialized on the axis with velocity parallel to e_r therefore remains on it. Along this geodesic, write $\rho = \rho_r(s)$. The restricted metric and log-density derivative are $g = \rho(s)^2\,ds^2$, and $\frac{\mathrm{d}}{\mathrm{d}s}\log p_Z(s) = -\rho(s)s$ respectively. The geodesic equation is $\ddot{s} + \frac{\rho'(s)}{\rho(s)}\dot{s}^2 = 0$, Hence we simplify via exact cancellation to

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\log p_Z(s(t)) &= \left[-\rho(s) - s\rho'(s)\right]\dot{s}^2 - \rho(s)s\ddot{s} \\ &= -\rho(s)\dot{s}^2,\end{aligned}$$

and because $\rho(s) > 0$, the log-density is strictly concave on nonconstant coordinate-axis geodesics. □

C.3.2 Sectional curvature

At the common centre, define $R_{ijij}(0) = R_0(e_i, e_j, e_j, e_i)$. Then $\partial_j \rho_i(0) = 0$. All first metric derivative hence vanish at $z = 0$. Differentiating $\partial_j \rho_i = -z_j C_{ij}$ once again yields $\partial_\ell \partial_j \rho_i(0) = -\delta_{j\ell} C_{ij}(0)$. As $g_{ii} = \rho_i^2$, $\partial_j^2 g_{ii}(0) = -2\rho_i C_{ij}$, with ρ_i and C_{ij} taken at the origin. Under the curvature convention $R_{ijij}(0) = -\frac{1}{2}\left(\partial_j^2 g_{ii}(0) + \partial_i^2 g_{jj}(0)\right)$, we now get $R_{ijij}(0) = (\rho_i + \rho_j)C_{ij}$. Since $g_{ii}(0)g_{jj}(0) = \rho_i^2\rho_j^2$, the sectional curvature of the coordinate (i, j)-plane is

$$K_{ij}(0) = \frac{\rho_i + \rho_j}{\rho_i^2 \rho_j^2}\operatorname{Cov}_{\gamma(0)}[q_{k,i}, q_{k,j}].$$

Hence the pairwise precision covariance governing the adjustment to geodesic log-concavity also determines the coordinate sectional curvature at a centre.

C.3.3 LOCAL EXPANSION OF GEODESIC LOG-DENSITY

In coordinates normalized so that $P(0) = I$, abbreviate $C_{ij}^0 = C_{ij}(0)$. Let R_0 denote the Riemann four-tensor evaluated at the origin. As a (centered) Gaussian star is even in z, $C_{ij}(z) = C_{ij}^0 + \mathcal{O}(\|z\|^2)$. Moreover, $\partial_j \rho_i(0) = 0$, and $\partial_\ell \partial_j \rho_i(0) = -\delta_{j\ell} C_{ij}^0$. If we now take the Taylor expansion, $\rho_i(z) = 1 - \frac{1}{2}\sum_j C_{ij}^0 z_j^2 + \mathcal{O}(\|z\|^4)$. This means that the first term of Equation (39) yields

$$-u^\top P(z) u = -\|u\|_2^2 + \frac{1}{2}\sum_{i,j} C_{ij}^0 u_i^2 z_j^2 + \mathcal{O}(\|z\|^4 \|u\|^2).$$

Meanwhile, the pairwise term satisfies

$$\sum_{i<j} C_{ij}(z) \left(\sqrt{\frac{\rho_i}{\rho_j}}\, u_i z_j - \sqrt{\frac{\rho_j}{\rho_i}}\, z_i u_j \right)^2 = \sum_{i<j} C_{ij}^0 (u_i z_j - u_j z_i)^2 + \mathcal{O}(\|z\|^4 \|u\|^2).$$

Since, at $P(0) = I$, $R_{ijij}(0) = 2C_{ij}^0$, we hence obtain $\frac{1}{2} R_0(z,u,u,z) = \sum_{i<j} C_{ij}^0 (u_i z_j - u_j z_i)^2$. Combining expansions results in

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} \log p_Z(z(t)) = -\|u\|_2^2 + \frac{1}{2}\sum_{i,j} C_{ij}^0 u_i^2 z_j^2 + \frac{1}{2} R_0(z,u,u,z) + \mathcal{O}(\|z\|^4 \|u\|^2).$$

The quadratic correction separates into the local variation of the expected precision and the curvature contribution generated by cross-coordinate precision covariance.

C.3.4 RICCI CONTRACTION

For a shared-mean Gaussian mixture centred at the origin, we can derive the Ricci tensor without assuming that the component precisions share an eigenbasis. For brevity, let $Q_k = A_k^{-1}, P = P(0) = \mathbb{E}_{\gamma(0)}[Q_k]$, and $L_k = Q_k - P$. Every component score vanishes at $z = 0$, so $D\gamma_k(0) = 0$. Differentiating, once again, gives, for $u, v \in \mathbb{R}^d$,

$$D^2\gamma_k(0)[u,v] = -\gamma_k(0)\, u^\top L_k v.$$

Also, since $\mathbb{E}_{\gamma(0)}[L_k] = 0$,

$$D^2 P(0)[u,v] = -\mathbb{E}_{\gamma(0)}\left[L_k \,(u^\top L_k v) \right].$$

Because $DP(0) = 0$ and $g = P^2$,

$$D^2 g(0)[u,v] = P\, D^2 P(0)[u,v] + D^2 P(0)[u,v] P.$$

Define $S_k = PL_k + L_k P$. In coordinates, $g_{ij,ab}(0) = -\mathbb{E}_{\gamma(0)}\left[S_{k,ij} L_{k,ab} \right]$. At a point with $\partial_a g_{ij} = 0$, the Ricci tensor is

$$\mathrm{Ric}_{jl} = \frac{1}{2} g^{ik} \left(g_{il,jk} + g_{jk,il} - g_{ik,jl} - g_{jl,ik} \right).$$

Substituting $g^{-1}(0) = P^{-2}$ and the preceding expression for $g_{ij,ab}$, the contribution of a fixed component is

$$\frac{1}{2}\left[\mathrm{Tr}(P^{-2}S_k) L_k + \mathrm{Tr}(P^{-2}L_k) S_k - L_k P^{-2} S_k - S_k P^{-2} L_k \right].$$

For the last few steps here, let $C_k = P^{-1}Q_k - I = P^{-1}L_k$, and, as L_k is symmetric, notice that $PC_k = C_k^\top P$. Further notice the identities $\mathrm{Tr}(P^{-2}S_k) = 2\,\mathrm{Tr}(C_k)$, and $\mathrm{Tr}(P^{-2}L_k) = \mathrm{Tr}(P^{-1}C_k)$. Applying these to the above contraction yields

$$\boxed{\begin{aligned}\mathrm{Ric}(0) = \mathbb{E}_{\gamma(0)}\big[&\mathrm{Tr}(C_k) P C_k + \mathrm{Tr}(P^{-1}C_k) P\, \mathrm{Sym}(C_k) P \\ &- P C_k^2 - \mathrm{Sym}(P C_k^\top C_k) \big].\end{aligned}}$$

For diagonalised component precisions, we get an analogous form, where, by letting $c_{k,i} = \frac{q_{k,i}}{\rho_i} - 1$, $a_k = \sum_j c_{k,j}$, and $b_k = \sum_j \frac{c_{k,j}}{\rho_j}$, the Ricci tensor is

$$\mathrm{Ric}_{ii}(0) = \mathbb{E}_{\gamma(0)}\left[\rho_i c_{k,i} a_k + \rho_i^2 c_{k,i} b_k - 2\rho_i c_{k,i}^2 \right].$$

The vectors c_k, together with a_k and b_k, can be accumulated component-wise, giving all diagonal Ricci entries in $\mathcal{O}(Kd)$ time. This can also be seen from the full general expression by noting that the only operations are matrix multiplications and inverses of diagonal matrices, traces, and sums.

C.4 CHAIN-SHAPED DENSITIES

Suppose that all mixture components share covariance A, with means $\mu_k = \alpha_k v$ along a fixed nonzero direction v. The expected precision is constant, $P(z) = \sum_k \gamma_k(z) A^{-1} = A^{-1}$, so $g_Z = A^{-2}$ and its geodesics are affine lines. Let $\sigma_\alpha^2(z) = \mathrm{Var}_{\gamma(z)}(\alpha_k)$. The component scores satisfy $s_k(z) = -A^{-1}z + \alpha_k A^{-1}v$, and Equation (38) gives

$$\nabla^2 \log p_Z(z) = -A^{-1} + \sigma_\alpha^2(z)(A^{-1}v)(A^{-1}v)^\top.$$

Hence, for a geodesic with velocity u,

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} \log p_Z(z(t)) = -u^\top A^{-1} u + \sigma_\alpha^2(z)(u^\top A^{-1} v)^2.$$

Geodesic log-concavity at z therefore holds in every tangent direction iff

$$\sigma_\alpha^2(z) \sup_{u \neq 0} \frac{u^\top (A^{-1}v)(A^{-1}v)^\top u}{u^\top A^{-1} u} \leq 1.$$

The supremum can be recognized as a generalized Rayleigh quotient. Since the numerator is of rank one,

$$\lambda_{\max} = (A^{-1}v)^\top A (A^{-1}v) = v^\top A^{-1} v = \|v\|_{A^{-1}}^2.$$

Here, $\|v\|_{A^{-1}}^2 := v^\top A^{-1} v$. Thus

$$\mathrm{Var}_{\gamma(z)}(\alpha_k) \leq \|v\|_{A^{-1}}^{-2}.$$

The maximizing generalized eigenvector is parallel to v, so the chain direction is the first direction in which geodesic log-concavity fails as the responsibility-weighted spread of component means increases.

This example shows that multimodality does not necessarily require a non-flat metric. Shared-covariance mixtures induce the same constant metric, yet their geodesic log-concavity can differ because it depends jointly on the density and the metric. The following proposition generalizes the one-dimensional chain construction.

Proposition 3 (Shared-covariance mixtures)**.** *Suppose $A_k = A$ for every component and let $\Sigma_\mu(z) = \mathrm{Cov}_{\gamma(z)}[\mu_k]$. Then $g_Z = A^{-2}$, and geodesic log-concavity holds in every tangent direction at z iff $\lambda_{\max}\left(A^{-1/2}\Sigma_\mu(z)A^{-1/2}\right) \leq 1$.*

Proof. The metric is constant. Moreover, $\nabla^2 \log p_Z(z) = -A^{-1} + A^{-1}\Sigma_\mu(z)A^{-1}$. Hence

$$\sup_{u \neq 0} \frac{u^\top A^{-1}\Sigma_\mu A^{-1} u}{u^\top A^{-1} u} = \lambda_{\max}\left(A^{-1/2}\Sigma_\mu A^{-1/2}\right),$$

and the result follows. □

D TRAINING AND PATH-OPTIMIZATION DETAILS

This section provides the adaptive flow–mixture training procedure, the metric-based geodesic solver, and the definitions of the geometric regularizers and mixture updates used in the experiments.

D.1 ADAPTIVE FLOW–MIXTURE TRAINING

For a proposed merge of components a and b, we use the moment-preserving parameters

$$w_{ab} = w_a + w_b, \tag{40}$$

$$\mu_{ab} = \frac{w_a \mu_a + w_b \mu_b}{w_{ab}}, \tag{41}$$

$$A_{ab} = \frac{1}{w_{ab}} \sum_{k \in \{a,b\}} w_k \left[A_k + (\mu_k - \mu_{ab})(\mu_k - \mu_{ab})^\top\right]. \tag{42}$$

The score S is evaluated with the flow fixed, and the best proposed merge is accepted only when its score improves on the unmerged model by more than δ_{merge}.

Algorithm 1: Stabilized Training of an Adaptive Multimodal Flow

Input: Dataset $\{x_n\}_{n=1}^N$; flow ϕ_θ; number of epochs E; initial mixture size m; Dirichlet parameter α and coefficient λ_{dir}; damping coefficient ζ; responsibility temperature schedule $\{T_e\}_{e=1}^E$; mixture-update interval K_{mix}; structural warmup E_0; pruning threshold ε_{p}; mixture-complexity coefficient β_m; merge-acceptance tolerance δ_{merge}; weight floor ϵ_w; covariance floor ϵ_A; optional mean-shrinkage coefficient s

Output: Flow parameters θ and active mixture parameters $\{w_k, \mu_k, A_k\}_{k=1}^K$

Initialize ϕ_θ and $\{w_k, \mu_k, A_k\}_{k=1}^m$;

for $e = 1, \dots, E$ **do**

Update θ over minibatches by minimizing $\mathcal{L}_{\mathrm{train}}$ from Equation (14), treating the current mixture parameters as fixed;

if $e > E_0$ ***and*** $e \bmod K_{\mathrm{mix}} = 0$ **then**

Compute latent codes $z_n = \phi_\theta(x_n)$;

Compute the temperature-smoothed responsibilities

$$r_{nk}^{(T_e)} = \mathrm{softmax}_k\left(\frac{\log w_k + \log \mathcal{N}(z_n; \mu_k, A_k)}{T_e}\right)$$

and effective counts $N_k = \sum_n r_{nk}^{(T_e)}$;

Compute the MAP-style weight target

$$\widetilde{w}_k = \max\{N_k + \lambda_{\mathrm{dir}}(\alpha - 1), \epsilon_w\}, \qquad w_k^* = \frac{\widetilde{w}_k}{\sum_j \widetilde{w}_j},$$

together with the responsibility-weighted targets

$$\mu_k^* = \frac{1}{N_k}\sum_{n=1}^N r_{nk}^{(T_e)} z_n,$$

$$A_k^* = \frac{1}{N_k}\sum_{n=1}^N r_{nk}^{(T_e)}(z_n - \mu_k^*)(z_n - \mu_k^*)^\top.$$

if $s > 0$ **then**

Apply the mean-shrinkage update in Remark 2 to the selected target means $\{\mu_k^*\}$, using the target weights $\{w_k^*\}$;

Apply the damped weight and mean updates

$$w_k^+ = (1-\zeta)w_k + \zeta w_k^*,$$
$$\mu_k^+ = (1-\zeta)\mu_k + \zeta \mu_k^*.$$

Renormalize $\{w_k^+\}_{k=1}^K$. For diagonal covariances, update the variances in log space:

$$\log \mathrm{diag}\, A_k^+ = (1-\zeta)\log \mathrm{diag}\, A_k + \zeta \log \mathrm{diag}\, A_k^*.$$

For full covariances, set $A_k^+ = (1-\zeta)A_k + \zeta A_k^* + \epsilon_A I$;

Prune components whose weights fall below ε_{p} and renormalize the remaining weights;

Let p_X denote the flow-induced data density. For each candidate merge, evaluate

$$S = -\sum_{n=1}^N \log p_X(x_n) + \beta_m K - \lambda_{\mathrm{dir}}(\alpha - 1)\sum_{k=1}^K \log w_k,$$

where K is the current number of active components;

Accept the best merge only if it decreases S by more than δ_{merge};

return θ *and* $\{w_k, \mu_k, A_k\}_{k=1}^K$;

D.2 Geometric Regularizers

Let $J_\theta(x) = D\phi_\theta(x)$ be the Jacobian of the flow, with x sampled from the empirical minibatch distribution. The stochastic local-isometry regularizer is

$$\mathcal{L}_{\mathrm{iso}} = \mathbb{E}_{x,v}\left[\left(\left\|J_\theta(x)^\top v\right\|_2^2 - 1\right)^2\right], \tag{43}$$

where $v \in \mathbb{R}^d$ is sampled from a standard Gaussian distribution and normalized to unit Euclidean length. Vector–Jacobian products are evaluated without explicitly forming the full Jacobian.

For MNIST and BlueWhale, we additionally use a weak metric-smoothness regularizer. Given minibatch latent codes $\{z_i\}_{i=1}^B$, define

$$\mathcal{L}_{\mathrm{metric}} = \frac{\sum_{i\neq j} a_{ij}\|P(z_i) - P(z_j)\|_F^2}{\sum_{i\neq j} a_{ij}}, \qquad a_{ij} = \exp\left(-\frac{\|z_i - z_j\|_2^2}{h}\right), \tag{44}$$

where h is the median pairwise squared distance within the sampled minibatch. The weighting emphasizes nearby pairs, so the loss discourages abrupt local changes in the expected-precision field.

D.3 Mean-Shrinkage Update

Remark 2 (Mean-shrinkage update)**.** *For selected experiments, we encourage overlap between a subset $\mathcal{I}$ of mixture components by moving their target means slightly toward their target-weighted centre. We first compute*

$$\bar{\mu} = \frac{\sum_{k\in\mathcal{I}} w_k^* \mu_k^*}{\sum_{k\in\mathcal{I}} w_k^*}, \tag{45}$$

and then apply

$$\mu_k^* \leftarrow (1-s)\mu_k^* + s\bar{\mu}, \qquad k \in \mathcal{I}, \tag{46}$$

where $s \in [0,1]$ is the mean-shrinkage coefficient. Depending on the experiment, $\mathcal{I}$ contains either all active components or the two components with the largest mixture weights. This operation modifies the mixture update directly and is not an additional term in the flow-training objective.

D.4 Metric-Based Geodesic Computation

Algorithm 2: Metric-Based Latent Geodesic Computation

Input: Trained flow ϕ_θ and inverse ϕ_θ^{-1}; mixture parameters $\{w_k, \mu_k, A_k\}_{k=1}^K$; endpoints x_A, x_B; number of segments T; optimizer step size η_{geo}; number of iterations M

Output: Data-space path $\{x_i\}_{i=0}^T$

Compute $z_0 = \phi_\theta(x_A)$ and $z_T = \phi_\theta(x_B)$;

Initialize

$$z_i = \left(1 - \frac{i}{T}\right) z_0 + \frac{i}{T} z_T, \qquad i = 1, \dots, T-1;$$

for $m = 1, \dots, M$ **do**

  for $i = 0, \dots, T-1$ **do**

    Set $\bar{z}_i = (z_i + z_{i+1})/2$;

    Compute

$$P(\bar{z}_i) = \sum_{k=1}^{K} \gamma_k(\bar{z}_i) A_k^{-1};$$

  Compute

$$\mathcal{E}_{gZ} = T \sum_{i=0}^{T-1} \|P(\bar{z}_i)(z_{i+1} - z_i)\|_2^2;$$

  Update only $z_1, \dots, z_{T-1}$ by taking a gradient-based optimizer step on $\mathcal{E}_{gZ}$ with step size η_{geo};

Map the optimized vertices to data space: $x_i = \phi_\theta^{-1}(z_i)$ for $i = 0, \dots, T$;

return $\{x_i\}_{i=0}^T$;

Algorithm 2 minimizes only the discrete energy of the proposed metric $g_Z = P^2$. It contains no density modulation or auxiliary smoothness, spacing, Euclidean-arclength, or geodesic log-concavity penalty. Relative to the convention for E_g in Appendix B, the constant factor $1/2$ is omitted; this does not change the minimizing path.

D.5 BlueWhale Support-Graph Path Computation

The encoded BlueWhale observations concentrate around three thin latent curves. Direct continuous path optimization can therefore move between nearby branches through regions containing little or no observed support. For this experiment, we instead compute a data-supported approximation on a neighbourhood graph constructed from the full-dimensional latent codes. The two-dimensional PCA representation is used only for visualization.

Let

$$z_n = \phi_\theta(x_n), \qquad n = 1, \dots, N, \tag{47}$$

denote the encoded training observations. We construct a mutual k-nearest-neighbour graph on $\{z_n\}_{n=1}^N$ using Euclidean distance only to determine local connectivity. In the reported experiment, we use $k = 40$.

For each retained edge (i, j), let

$$\bar{z}_{ij} = \frac{z_i + z_j}{2}, \qquad \Delta z_{ij} = z_j - z_i. \tag{48}$$

Its midpoint approximation to the Riemannian length under the proposed metric $g_Z(z) = P(z)^2$ is

$$\ell_{ij} = \sqrt{\Delta z_{ij}^\top g_Z(\bar{z}_{ij}) \Delta z_{ij}}. \tag{49}$$

To discourage graph transitions through low-density regions, we modulate this length using the latent negative log-likelihood. Let $\beta_{\mathrm{NLL}} \geq 0$ be the density-penalty strength and let $c_{\mathrm{clip}} > 0$ be a clipping threshold. Define

$$q_{\mathrm{NLL}}(z) = \mathrm{clip}\left(\frac{-\log p_Z(z) - m_{\mathrm{NLL}}}{s_{\mathrm{NLL}}}, -c_{\mathrm{clip}}, c_{\mathrm{clip}} \right), \tag{50}$$

where m_{NLL} is the median negative log-likelihood over the encoded training set, s_{NLL} is the corresponding median absolute deviation (clipped below by a small positive numerical floor), and $c_{\mathrm{clip}} = 6$. The edge cost is then

$$C_{ij}^{\mathrm{graph}} = \ell_{ij} \exp\left(\frac{\beta_{\mathrm{NLL}}}{2} q_{\mathrm{NLL}}(\bar{z}_{ij}) \right). \tag{51}$$

We use $\beta_{\mathrm{NLL}} = 0.5$. Thus, edges passing through lower-density regions receive a larger cost, while setting $\beta_{\mathrm{NLL}} = 0$ recovers the graph approximation induced only by the proposed Riemannian metric.

We apply Dijkstra's algorithm to obtain the minimum-cost graph path between the encoded endpoint observations. Because its vertices correspond to observed BlueWhale images, the resulting sequence remains on the sampled data support. We therefore refer to it as a *density-aware, metric-weighted support-graph path*, rather than an exact continuous geodesic of g_Z.

E Experimental Protocol

E.1 Dataset Construction

E.1.1 Wavy-X

Wavy-X is generated from two sinusoidally perturbed strokes intersecting at the origin. For each sample, we first draw a stroke indicator $c \sim \mathrm{Bernoulli}(1/2)$ and a longitudinal coordinate $t \sim \mathrm{Unif}[-4, 4]$. Let u_c denote the unit direction of the selected stroke, n_c its normal direction, and $s_c \in \{-1, 1\}$ the orientation of its sinusoidal perturbation. The two strokes intersect at an angle of $90°$. Samples are generated as

$$x = (t + \epsilon_\parallel) u_c + s_c \, 0.6 \sin(2.5t) n_c + \epsilon_\perp n_c,$$

where

$$\epsilon_\parallel \sim \mathcal{N}(0, 0.04^2), \qquad \epsilon_\perp \sim \mathcal{N}(0, 0.12^2).$$

We generate $N = 8000$ observations using random seed zero.

E.1.2 Tetrahedral Star

The tetrahedral-star dataset consists of four noisy unit-length arms in $\mathbb{R}^3$. Let

$$c_1 = (1, 1, 1)^\top, \quad c_2 = (1, -1, -1)^\top, \quad c_3 = (-1, 1, -1)^\top, \quad c_4 = (-1, -1, 1)^\top, \tag{52}$$

and define the unit arm directions

$$u_k = \frac{c_k}{\|c_k\|_2}, \qquad k = 1, \ldots, 4. \tag{53}$$

For each arm, we independently draw $n_{\mathrm{arm}} = 2000$ longitudinal coordinates $t_{kn} \sim \mathrm{Uniform}[0, 1]$. Let

$$\epsilon_{kn}^{\perp} = (I_3 - u_k u_k^{\top})\xi_{kn}, \qquad \xi_{kn} \sim \mathcal{N}(0, \sigma_{\perp}^2 I_3), \tag{54}$$

and

$$\epsilon_{kn}^{\parallel} = a_{kn} u_k, \qquad a_{kn} \sim \mathcal{N}(0, \sigma_{\parallel}^2), \tag{55}$$

with all random variables sampled independently. Each observation is

$$x_{kn} = t_{kn} u_k + \epsilon_{kn}^{\perp} + \epsilon_{kn}^{\parallel}. \tag{56}$$

We use $\sigma_{\perp} = 0.05$, $\sigma_{\parallel} = 0.02$, and random seed zero, producing $N = 4n_{\mathrm{arm}} = 8000$ observations. Arm identities are retained only for evaluation and visualization.

E.1.3 BLUEWHALE

We construct BlueWhale by rendering a fixed whale mesh under viewpoints parameterized by a three-dimensional star-shaped variable. For each sample, we select one of three unit arm directions u_k, draw $t \sim \mathrm{Uniform}[0, 1]$, and set $\theta = tu_k$. The parameter θ determines the object rotation under fixed camera, lighting, and background settings. Images are rendered with supersampling and downsampled to 28×28 grayscale pixels. We generate 5,000 images and retain the arm identity and θ for evaluation only. These metadata define a reference trajectory between different arms: from the first endpoint to the common origin and then from the origin to the second endpoint.

E.1.4 MNIST

Digit labels are not used to train the MNIST generative models; they are used only to select held-out different-digit endpoint pairs, train the independent evaluation classifier, and organize visualizations. The adaptive model is initialized with $m = 30$ candidate components, and interpolation paths use $T = 16$ segments. Additional metric specifications are reported below.

E.2 EVALUATION METRICS

For Wavy-X, we report negative log-likelihood (NLL) and mean local isometry distortion. The latter is

$$D_{\mathrm{iso}} = \mathbb{E}_x \left[\frac{1}{d} \left\| D\phi_\theta(x)^{\top} D\phi_\theta(x) - I_d \right\|_F \right], \tag{57}$$

where $d = 2$ and ϕ_θ is the learned flow from data to latent space. Flow conditioning is summarized by the 95th percentile, denoted Q95, of

$$\kappa_J(x) = \frac{\sigma_{\max}(D\phi_\theta(x))}{\sigma_{\min}(D\phi_\theta(x))}. \tag{58}$$

We distinguish the active-component count

$$K_{\mathrm{active}} = \sum_{k=1}^{K} \mathbf{1}\{w_k \geq \tau\}, \qquad \tau = 0.01, \tag{59}$$

from the entropy-based effective component count

$$K_{\mathrm{eff}} = \exp\left(-\sum_{k=1}^{K} \bar{w}_k \log \bar{w}_k \right), \qquad \bar{w}_k = \frac{w_k}{\sum_{j=1}^{K} w_j}. \tag{60}$$

The former is integer-valued, whereas the latter measures how evenly the mixture mass is distributed across components.

For a curvature margin $\delta_\star \geq 0$, the star–Ricci diagnostic is

$$L_\star = \sum_{i=1}^{d} \mathrm{softplus}(r_i + \delta_\star), \qquad r_i = \mathrm{Ric}_{ii}(0), \tag{61}$$

and $\lambda_{\max}(\mathrm{Ric}_\star(0))$ denotes the largest eigenvalue of the corresponding local star–Ricci matrix. These quantities are evaluated after training as local diagnostics and are not used to optimize the reported models. We use $\delta_\star = 0$ in all reported experiments. For learned mixtures with unequal means, these are structured star diagnostics rather than the exact Ricci curvature of the learned metric.

After subdividing each optimized latent segment into 20 equal parts and mapping the resulting vertices to data space, let $\{\widetilde{x}_m\}_{m=0}^{M}$ denote the densified path and let $\mathcal{X}$ be the sampled data cloud. Its support deviation is

$$D_{\text{support}} = \frac{1}{M-1}\sum_{m=1}^{M-1}\frac{\min_{x\in\mathcal{X}}\|\widetilde{x}_m - x\|_2}{s_{\mathcal{X}}}, \tag{62}$$

where $s_{\mathcal{X}}$ is the median nearest-neighbour distance within $\mathcal{X}$, excluding each point itself from its nearest-neighbour query. Increasing the subdivision count from 20 to 40 produces essentially unchanged results.

BlueWhale. At matched positions along each candidate and reference sequence, we report mean squared error (MSE), mean absolute error (MAE), and the structural similarity index (SSIM). Worst-step MSE is the largest frame-wise MSE along the sequence. The metadata-defined reference and the procedure used to construct the support-graph path are described in Appendices E.1.3 and D.5.

MNIST. Let $f(x)$ denote the representation produced by the independently trained MNIST classifier. Let (m_p, Σ_p) and (m_r, Σ_r) be the empirical means and covariances of the interior-path and reference-image features, respectively. Their feature Fréchet distance is

$$D_{\text{FD}} = \|m_p - m_r\|_2^2 + \mathrm{tr}\Big(\Sigma_p + \Sigma_r - 2(\Sigma_p^{1/2}\Sigma_r\Sigma_p^{1/2})^{1/2}\Big). \tag{63}$$

For a path $\{x_t\}_{t=0}^{T}$, define the perceptual step lengths

$$\ell_t = \|f(x_{t+1}) - f(x_t)\|_2, \qquad t = 0, \dots, T-1. \tag{64}$$

Perceptual path length and relative perceptual-distance variance are

$$\mathrm{PPL} = \sum_{t=0}^{T-1}\ell_t, \qquad \mathrm{PDV}_{\text{rel}} = \frac{T^{-1}\sum_{t=0}^{T-1}(\ell_t - \bar{\ell})^2}{\bar{\ell}^2}, \qquad \bar{\ell} = \frac{1}{T}\sum_{t=0}^{T-1}\ell_t. \tag{65}$$

Feature FD is our primary measure of interior-sample realism; PPL and $\mathrm{PDV}_{\text{rel}}$ measure perceptual length and step-size uniformity.

E.3 Component-Wise Linear Reconstruction

For observations $x_n \in \mathbb{R}^d$, define $r_{nk} = \gamma_k(\phi_\theta(x_n))$ and $N_k = \sum_n r_{nk}$. The weighted data-space mean and covariance of component k are

$$\bar{x}_k = \frac{1}{N_k}\sum_n r_{nk}x_n, \qquad C_k^x = \frac{1}{N_k}\sum_n r_{nk}(x_n - \bar{x}_k)(x_n - \bar{x}_k)^\top. \tag{66}$$

Let $\nu_{k,1} \geq \cdots \geq \nu_{k,d} \geq 0$ be the eigenvalues of C_k^x. For a variance-retention threshold $\tau_{\text{var}} \in (0,1)$, the component-wise effective dimension is

$$d_k = \min\left\{r : \frac{\sum_{i=1}^{r}\nu_{k,i}}{\sum_{i=1}^{d}\nu_{k,i}} \geq \tau_{\text{var}}\right\}. \tag{67}$$

If $U_k \in \mathbb{R}^{d\times d_k}$ contains the leading eigenvectors, the component-k reconstruction of x_n is

$$\widehat{x}_{nk} = \bar{x}_k + U_k U_k^\top(x_n - \bar{x}_k). \tag{68}$$

Reported oracle reconstructions select, for each observation, the component with the smallest reconstruction error. This is a post-training diagnostic, not an inference-time assignment rule. We use $\tau_{\text{var}} = 0.90$ in the reported analyses.

F Additional Synthetic Results

This section provides additional results for the synthetic point-cloud experiments. Wavy-X illustrates component-wise structure and flow deformation in two dimensions, while the tetrahedral-star experiment extends the evaluation to three dimensions. The local reconstruction analyses are performed after training and are not used to optimize the flow. Latent responsibilities define weighted component memberships, from which we fit component-wise linear subspaces in data space. We report the number of leading directions required to explain a specified fraction of weighted variance as a local effective dimension, rather than as a formal intrinsic dimension.

F.1 Wavy-X

F.1.1 Component-Wise Reconstruction

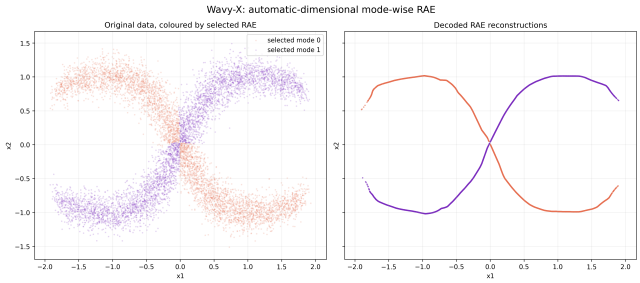


Figure 5: **Component-wise linear reconstruction on Wavy-X.** Original samples and responsibility-weighted linear reconstructions for the two retained components. One principal direction explains 99.41% and 99.29% of their respective weighted data-space variance.

The two retained components have weights 0.5032 and 0.4968, and their dominant directions approximately recover the two crossing arm pairs. Oracle best-subspace selection, which assigns each sample to the component-wise reconstruction with the lowest error, gives mean MSE 1.04×10^{-2} and Q95 MSE 4.17×10^{-2}. These results summarize the dominant component-wise linear structure; they do not establish that the complete Wavy-X distribution is globally linear or one-dimensional.

F.1.2 Layer-Wise Transport

Figure 6 visualizes how matched Wavy-X samples evolve through the layers of each normalizing flow. Blue points show the observations at the input layer, red points show their final latent representations, and black curves trace individual samples through the intermediate layers. The figure complements the numerical distortion measures by showing the different transport deformations produced by the four latent models.

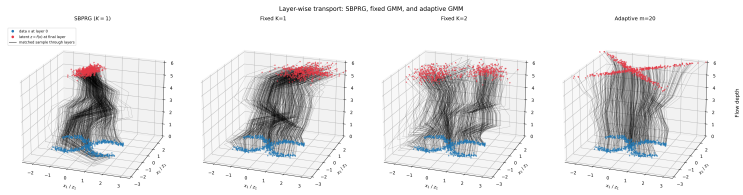


Figure 6: **Layer-wise transport on Wavy-X.** Matched samples are traced from the data space at flow depth zero (blue), through the intermediate flow layers (black), to the final latent representation (red). From left to right: SBPRG with $K = 1$, the fixed $m = 1$ single-Gaussian model, the fixed $m = 2$ Gaussian-mixture model, and the adaptive model initialized with $m = 20$ candidate components. The visualization shows how each learned flow reorganizes the same Wavy-X observations and complements the quantitative transport-distortion comparison.

F.2 THREE-DIMENSIONAL TETRAHEDRAL-STAR EXPERIMENT

We further evaluate the proposed method on a three-dimensional tetrahedral-star dataset consisting of four noisy arms meeting at a common junction. This extends the Wavy-X setting to a higher-dimensional multimodal geometry. We compare SBPRG, fixed one- and two-component mixture bases, and the adaptive mixture-base model. The same endpoint pairs are re-encoded under each model before computing the corresponding metric paths.

Figure 7 shows the resulting paths in latent and data space. The adaptive representation preserves the multi-arm organization, and the mapped paths remain close to the observed support for both selected endpoint pairs.

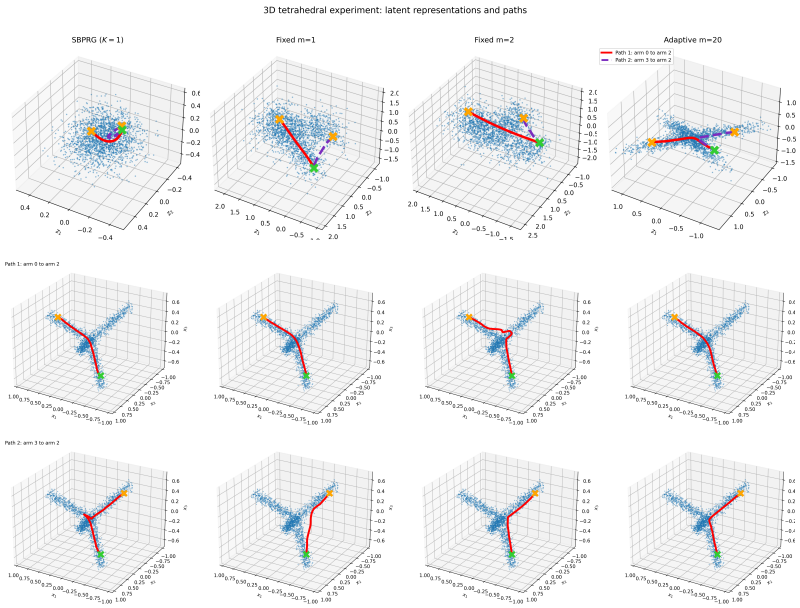


Figure 7: **Tetrahedral-star metric paths. Top:** latent representations and paths. **Bottom:** corresponding paths mapped to data space. Red and purple curves denote the two prescribed endpoint pairs.

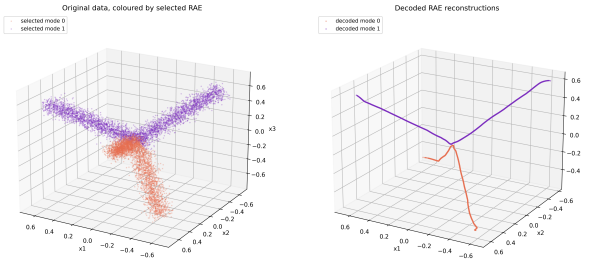
Table 4 reports the quantitative results. In this run, the adaptive model retains approximately two effective components and obtains the best NLL, isometry distortion, and upper-tail Jacobian condition number. It obtains the lowest support deviation for Path 2, while SBPRG obtains the lowest value for Path 1.

Table 4: **Tetrahedral-star quantitative comparison.** Lower is better for NLL, isometry distortion, upper-tail Jacobian conditioning, the star–Ricci diagnostic $L_\star$, and support deviation.

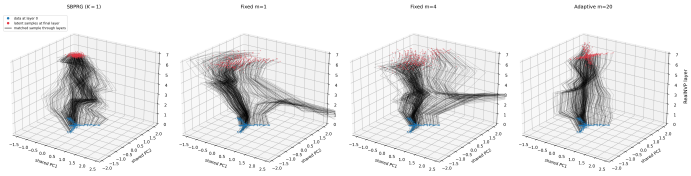
Model	NLL ↓	D_{iso} ↓	Q95 κ_J ↓	K_{eff}	$L_\star$ ↓	$\lambda_{\max}(\text{Ric}_\star(0))$	Support 1 ↓	Support 2 ↓
SBPRG ($K = 1$)	-1.7634	4.6871	61.5166	1.0000	2.0794	0.0000	**0.3970**	0.4137
Fixed $m = 1$	-1.1533	21.3192	12.2771	1.0000	2.0794	≈ 0	0.4329	1.1875
Fixed $m = 2$	-1.1440	23.5916	11.3507	2.0000	2.0794	≈ 0	0.5142	0.6308
Adaptive $m = 20$	**-1.7787**	**2.8724**	**5.9996**	1.9988	$< 10^{-6}$	**-14.8535**	0.4329	**0.4067**

G ADDITIONAL IMAGE-DATA ANALYSIS

The following post-training analyses summarize component-wise linear structure for BlueWhale and MNIST. Component subspaces are fitted in data space using latent responsibilities and do not alter the learned flow.



(a) Component-wise local reconstruction.



(b) Layer-wise transport in a shared PCA projection.

Figure 8: **Additional tetrahedral-star diagnostics.** (a) Component-wise reconstructions summarize the dominant local linear structure. (b) Layer-wise trajectories visualize how each flow transports the data toward its latent base.

G.1 BLUEWHALE

The learned mixture retains three active components with weights 0.2760, 0.3182, and 0.4058. Under the 90% weighted-variance criterion, their effective dimensions are $d_k \in \{2, 3, 2\}$, explaining 96.65%, 91.16%, and 90.70% of the respective weighted data-space variance. Although each generating arm is parameterized by one scalar, its rendered image trajectory is nonlinear in pixel space and may therefore require more than one linear direction.

Oracle best-subspace reconstruction gives mean MSE 5.66×10^{-4} and Q95 MSE 1.46×10^{-3}. Here, oracle selection means that the lowest-error component reconstruction is chosen for each sample; it is a post-training diagnostic rather than an inference-time assignment rule.

G.2 MNIST

Figure 9 shows representative reconstructions obtained by oracle best-subspace selection. The model retains $K_{\text{active}} = 13$ active components, with component-wise effective dimensions ranging from 17 to 56 (mean 36). Digit 1 has the lowest reported reconstruction MSE (4.1×10^{-3}), whereas digit 8 has the highest (1.78×10^{-2}). This difference is consistent with the greater linear complexity of two-loop shapes, although reconstruction errors alone do not identify topology. Residual blur is expected because each reconstruction uses a linear approximation in pixel space. Digit labels are used only to organize the visualization and are not used to fit the component models.

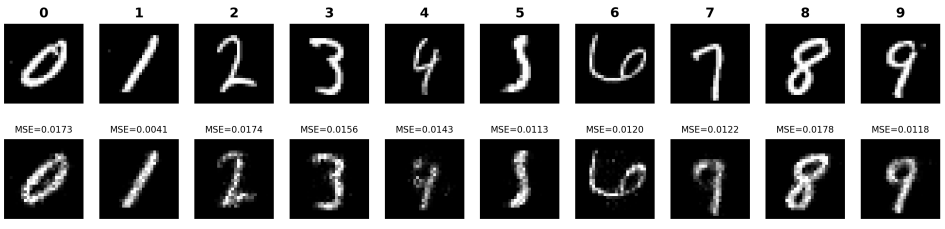


Figure 9: **Component-wise local reconstruction on MNIST.** *Top:* one representative sample per digit class. *Bottom:* oracle best-subspace reconstruction with pixel MSE. The model retains $K_{\text{active}} = 13$ active components from an initial pool of 30; digit labels are not used during training or subspace fitting.