Supplementary Materials

Nina Miolane
Stanford University
nmiolane@stanford.edu

Susan Holmes
Stanford University
susan@stat.stanford.edu

Abstract

These are the supplementary materials to the paper: "Learning Weighted Riemannian Submanifolds with Variational Autoencoders and Riemannian Variational Autoencoders".

Section 1 provides elements of Riemannian geometry. Section 2 presents the proof of Proposition 2. Section 3 highlights the difference between a Riemannian VAE and a VAE trained on the tangent space on the input manifold. Section 4 provides the computations for the theoretical 1D example of the paper. Sections 5 and 6 provide details on the experiments on synthetic data and real brain connectomes.

1. Elements of Riemannian Geometry

1.1. (Sub)manifolds, Weighted (Sub)manifolds

We review some elements of Riemannian geometry. In the main paper, these elements enable to introduce the Riemannian variational autoencoder (rVAE) and its goodness of fit in terms of weighted submanifold learning. We refer to [4, 3] for details on differential geometry.

Definition 1 (Riemannian manifold) *A Riemannian manifold M is a differentiable manifold M equipped with a Riemannian metric g, which is a smoothly varying inner product on the tangent spaces of M.*

We assume the Riemannian manifolds considered are simply connected and complete [4]. As the objective of the rVAE is submanifold learning, we recall the definition of a submanifold.

Definition 2 (Submanifold and embedded submanifold) *A subset N of M is a submanifold of M if it is included in M and is a manifold. The submanifold N of M is called an embedded submanifold of M if there exists a smooth manifold X and a smooth embedding f such that: $N = f(X)$.*

The submanifold N inherits differential and geometric structure from M. In particular, the Riemannian inner product from M restricts to tangent spaces T_xN for $x \in N$ to give a Riemannian metric on N.

The Riemannian metric g of M induces an infinitesimal volume element on each tangent space. Thus, a measure on the manifold M has the expression $dM = \sqrt{\det(g(x))}dx$ in a local coordinate system x. As the submanifold N can be equipped with the Riemannian metric inherited from M, we can define a Riemannian measure on it similarly. Once a measure is defined, we can introduce weighted manifolds and submanifolds.

Definition 3 (Weighted (sub)manifold) *Given a complete d-dimensional Riemannian manifold (M, g) and a smooth probability distribution $\omega : M \to \mathbb{R}$, the weighted manifold (M, ω) associated to M and ω is defined as the triplet:*

$$(M, g, \omega.dM), \tag{1}$$

where dM denotes the Riemannian measure of M.

1.2. Geodesics and Geodesic Subspaces

Consider a Riemannian manifold M. We introduce elements of differential geometry required to perform computations in the generative model of the rVAE.

Let u, v be vector fields on M. The Riemannian metric g on M induces the notion of covariant derivative, when considering the Levi-Civita connection associated with g [4]. Intuitively, the covariant derivative of v in the direction of u, written $\nabla_u v$, represents the change of the vector field v in the u direction.

Let γ be a curve on M parameterized by t, as $\gamma : [0, 1] \mapsto M, t \mapsto \gamma(t)$. Its velocity is: $\dot{\gamma} = \frac{d\gamma}{dt}$. Let u be a vector field $t \mapsto u(t)$ defined along γ. We can define the covariant derivative of u along γ to be $\frac{du}{dt} = \nabla_{\dot{\gamma}} u$.

Definition 4 (Riemannian geodesic) *The curve $\gamma : [0, 1] \mapsto M$ is called a geodesic if it satisfies the equation $\nabla_{\dot{\gamma}}\dot{\gamma} = 0$, where ∇ represents the covariant derivative of the Levi-Civita connection associated to the Riemannian metric of M.*

Intuitively, a geodesic is a curve that is parallel to itself. In this sense, a geodesic generalizes the notion of linear curve on Euclidean spaces. As an example, geodesics on the sphere are the great circles.

Geodesics are the equivalent on Riemannian manifolds of 1D linear subspaces in Euclidean spaces. We define "geodesic subspaces", that generalize to manifolds higher-dimensional subspaces of Euclidean spaces [1].

Definition 5 (Submanifold geodesic at a point) *A submanifold N of M is said to be geodesic at $x \in N$ if all geodesics of N passing through x are also geodesics of M.*

1.3. Riemannian Exp and Log Maps

For any point $x \in M$ and tangent vector at this point $u \in T_xM$, there exits a unique geodesic γ, with initial conditions $\gamma(0) = x$ and $\dot{\gamma}(0) = u$. In general, the existence of the geodesic is only guaranteed locally around 0. As we assume the completeness of the manifold M, the geodesics are defined on $\mathbb{R}$.

Definition 6 (Riemannian exp map) *Let γ be the unique geodesic γ, with initial conditions $\gamma(0) = x$ and $\dot{\gamma}(0) = u$. We define the Riemannian exponential map at x as:*

$$Exp_x(u) = \gamma(1). \tag{2}$$

The exponential map takes a point x and tangent vector u and returns the point at time 1 obtained by "shooting" with u along the geodesic. The exponential map is a local diffeomorphism onto a neighbourhood of x in M [4]. Let $V(x)$ be the largest neighbourhood on which the exponential map at x is a diffeomorphism. We can define its inverse on $V(x)$.

Definition 7 (Riemannian log map) *The Riemannian exponential map has an inverse on the domain $V(x)$, defined as the Riemannian log map at x, and denoted Log_x.*

The domain $V(x)$ is the global bijectivity domain of the exponential map at x. One can show that $V(x)$ is a star-shaped domain delimited by a continuous curve $\mathcal{C}_x$.

Definition 8 (Tangential cut locus and cut locus) *The curve C_x delimiting $V(x)$ is called the tangential cut-locus of x. The image of image by the exponential map at x is called the cutlocus of x.*

We can define the distance from a point x to its cut locus $C = \mathcal{C}_x$.

Definition 9 (Injectivity radius) *For each $x \in M$, the injectivity radius of M at x is defined as:*

$$inj_x(M) = sup\{r : Exp_x \text{ is injective on } Br(0) \subset T_xM\},$$

and the injectivity radius of M is defined as:

$$inj(M) = \inf_{x \in M} inj_x(M). \tag{3}$$

If M is compact, then $0 < inj(M) \leq diam(M)$. But in the general case, we may have $inj(M) = 0$ or $inj(M) = +\infty$.

1.4. Hadamard Manifolds

There is a class of manifolds, called Hadamard manifolds, for which the Riemannian exp map is a global diffeomorphism. Mathematically speaking, Hadamard manifolds are complete Riemannian manifolds with non-positive sectional curvature [4].

Theorem 1 (Cartan-Hadamard theorem) *For simply-connected Hadamard manifolds, the exponential map is a global diffeomorphism from T_xM onto $V(x) = M$. As a consequence, the log map is defined on M for any point $x \in M$.*

On Hadamard manifolds, the exponential map to transform the manifold into a vector space.

1.5. Distance in M and between submanifolds of M

We define the notion of distance between points on the manifold M and introduce the associated notion of distance between weighted submanifolds of M.

Definition 10 (Riemannian distance) *For any point $y \in V(x)$, the Riemannian distance function is given by $d_M(x, y) = ||Log_x(y)||_{g(x)}$ using the norm corresponding to the inner product $g(x)$ at x.*

We recall the notion of Wasserstein distance between probability distributions.

Definition 11 (Wasserstein distance) *The 2-Wasserstein distance, between probability measures μ and ν defined on M, is defined as:*

$$d_2(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{M \times M} d_M(x_1, x_2)^2 d\gamma(z_1, z_2) \right)^{1/2} \tag{4}$$

where $\Gamma(\mu, \nu)$ denotes the collection of all measures on $M \times M$ with marginals μ and ν on the first and second factors respectively.

This notion allows defining a notion of distance between two embedded submanifolds of M, with Riemannian metric associated with their embedding f. We consider the 2-Wasserstein distance associated with the corresponding distributions, interpreted as singular distribution in M. This notion of distance is the evaluation metric used to compare weighted submanifold learning techniques.

1.6. Fréchet Mean

Definition 12 (Fréchet mean) *Consider X a random variable with values on M and probability distribution p on M. The population Fréchet mean is defined as:*

$$\mu = argmin_{y \in M} \int_M d_M^2(x, y) dp(x). \tag{5}$$

Consider a random sample $X_1, ..., X_n$ on M. The sample Fréchet mean is defined as:

$$\bar{x} = argmin_{y \in M} \frac{1}{n} \sum_{i=1}^{n} d_M^2(y, x_i). \tag{6}$$

1.7. Examples of Riemannian manifolds

The hypersphere and the hyperbolic plane are examples of Riemannian manifolds. The n-dimensional hypersphere S^n is defined by its embedding in the $(n+1)$-Euclidean space as:

$$S^n = \left\{ x \in \mathbb{R}^{n+1} : x_1^2 + ... + x_{n+1}^2 = 1 \right\} \tag{7}$$

while the n-dimensional hyperbolic space H_n is defined by its embedding in the $(n+1)$-dimensional Minkowski space, as:

$$H_n = \left\{ x \in \mathbb{R}^{n+1} : -x_1^2 + ... + x_{n+1}^2 = -1 \right\}. \tag{8}$$

Both are Riemannian manifolds of constant curvature, which is positive for S^n and negative for H_n [4].

2. Proof of Proposition 1

We begin by introducing notations. For $T \in \mathbb{R}_+^*$, we denote μ_T the standard multivariate normal in $\mathbb{R}^L$ truncated at a distance T from the origin. We write B_T the closed ball of $\mathbb{R}^L$ of radius T, which is a compact subset of $\mathbb{R}^L$ as it is bounded and closed in $\mathbb{R}^L$.

We write $C(\mathbb{R}^L, \mathbb{R}^D)$ the set of continuous functions from $\mathbb{R}^L$ to $\mathbb{R}^D$, and we write $C(B_T, \mathbb{R}^D)$ the set of continuous functions from B_T to $\mathbb{R}^D$.

2.1. Preliminaries

We first introduce a lemma that generalizes a result from [2] to functions in D dimensions.

Lemma 1 *Consider K a compact subsect of $\mathbb{R}^L$. For any $\epsilon > 0$ and any continuous function $f \in C(K, \mathbb{R}^D)$, there exists a neural network represented by the function f_θ, parameterized by θ such that:*

$$\sup_{z \in K} ||f(z) - f_\theta(z)||_2^2 < D\epsilon. \tag{9}$$

Proof 1 *Take $\epsilon < 0$. Take $d \in [1, ..D]$ and consider f_d the d-th component of $f \in C(K, \mathbb{R}^D)$. We have: $f_d \in C(K, \mathbb{R})$. We use the results from [2] to show that there exists a neural network that approximates f_d with precision ϵ.*

In our notations, L is the dimension of the input space (as opposed to k in [2]), z is an element of the input space $\mathbb{R}^L$ (as opposed to x in [2]), and K is the number of units in the hidden layer (as opposed to n in [2]). We define:

$$M_L^{(K)}(\psi) = \left\{ h : \mathbb{R}^L \mapsto \mathbb{R} | h(z) = \sum_{k=1}^{K} \beta_k \psi(a_k' z - \theta_k) \right\}, \tag{10}$$

the space of functions represented by a fully connected neural network with one output unit and one hidden layer with n hidden units, and:

$$M_L(\psi) = \bigcup_{K=1}^{+\infty} M_L^{(K)}(\psi), \tag{11}$$

the space of functions represented by a fully connected neural network with one output unit and any number of hidden units. The ψ represents the common activation function of all units. We assume that ψ is continuous, bounded, and nonconstant, for example, the sigmoid activation function.

From Theorem 2 from [2], $M_L(\psi)$ is dense in $C(K, \mathbb{R})$. As $f_d \in C(K, \mathbb{R})$, there exists a function $h_d \in M_L(\psi)$ such that:

$$\sup_{z \in K} |f_d(z) - h_d(z)| < \epsilon. \tag{12}$$

We fix the functions $h_1, ..., h_D$ that approximate the D components of the function f. We define the function $h : \mathbb{R}^L \mapsto \mathbb{R}^D$ defined by $h = (h_1, ..., h_D)$. This function approximates f:

$$\begin{aligned}
&\sup_{z \in K} ||f(z) - h(z)||_2^2 \\
&\quad = \sup_{z \in K} \sum_{d=1}^{D} |f_d(z) - h_d(z)|^2 \\
&\quad \leq \sum_{d=1}^{D} \sup_{z \in K} |f_d(z) - h_d(z)|^2 \\
&\quad \leq \sum_{d=1}^{D} \sup_{z \in K} |f_d(z) - h_d(z)| \sup_{z \in K} |f_d(z) - h_d(z)| \\
&\quad \leq D\epsilon^2.
\end{aligned}$$

The function f can, therefore, be approximated by a neural network with precision ϵ, and specifically by the neural network represented by h and obtained by juxtaposing the neural networks defining the h_d, for $d = 1, .., D$.

2.2. Proof

We turn to the proof of Proposition 1, which we recall here. In what follows, functions with images in M are denoted with a subscript M, while functions with images in a tangent space of M have no subscript.

Proposition 1 *Let (N_T, ν_T) be a weighted Riemannian submanifold of M, embedded in a submanifold L of M homeomorphic to $\mathbb{R}^L$ and for which there exists an embedding f^M that verifies: $\nu_T = f^M * \mu_T$. Let assume the existence of $\mu \in M$ such that $N \subset V(\mu)$, where $V(\mu)$ is the maximal domain of global bijection of the Riemannian exponential of M at μ. Then, for any $0 < \epsilon < 1$, there exists a Riemannian VAE with decoder represented by a neural network f_θ, parameterized by θ, such that:*

$$d_2(N_T, N_{\theta,T}) < C_{T,D}^2 D\epsilon^2, \tag{13}$$

where d_2 is the 2-Wasserstein distance and $N_{\theta,T} = (f_\theta(\mathbb{R}^L, f_\theta * \mu_T))$, and $C_{T,D}$ a constant that depends on T and D.

Proof 2 *Let (N_T, ν_T) be a weighted Riemannian submanifold of M verifying the assumptions. We write $L = f^M(\mathbb{R}^L)$ the embedded Riemannian submanifold of M, homeomorphic to $\mathbb{R}^L$, that contains N_T: $N_T \subset L = f^M(\mathbb{R}^L)$. Without loss of generality, we can restrict N_T to the support of the singular probability distribution ν_T. We write $N_T = f^M(B_T)$. As B_T is compact and f^M is continuous by definition of an embedding, we observe that N_T is compact.*

As $N_T \subset V(\mu)$, we can define the function:

$$\begin{aligned} f : B_T &\mapsto T_\mu M \\ z &\mapsto f(z) = Log_\mu f^M(z). \end{aligned}$$

We have $f \in C(B_T, \mathbb{R}^D)$, where B_T is a compact subset of $\mathbb{R}^L$. From Lemma 1, there exists a neural network represented by f_θ parameterized by θ, such that:

$$\sup_{z \in B_T} ||f(z) - f_\theta(z)||_2^2 < D\epsilon. \tag{14}$$

We fix this neural network and corresponding function f_θ.

We introduce the weighted submanifold $(N_{\theta,T}, \nu_{\theta,T}$ associated to the neural network represented by f_θ:

$$N_{\theta,T} = (f_\theta^M(B_T), f_\theta^M|_{B_T} * \mathcal{N}(0,1)), \tag{15}$$

which corresponds to the generative model defining the Riemannian VAE. We show that $N_{\theta,T}$ can approximate N_T with a precision ϵ, in terms of the 2-Wasserstein distance.

We compute the squared 2-Wasserstein distance between N_T and $N_{\theta,T}$. By definition, the squared 2-Wasserstein distance between weighted submanifolds is the 2-Wasserstein distance between the probability measures defined on them. Here, the probability distributions are ν_T and $\nu_{\theta,T}$, so that we write:

$$\begin{aligned} &d_2(N_T, N_{\theta,T})^2 \\ &\quad = \inf_{\gamma \in \Gamma(\nu_T, \nu_{\theta,T})} \int_{M \times M} d_M(x_1, x_2)^2 d\gamma(x_1, x_2), \end{aligned}$$

where $\Gamma(\nu, \nu_{\theta,T})$ is the collection of all measures on $M \times M$ with marginals ν and $\nu_{\theta,T}$ on the first and second factors respectively. As ν and $\nu_{\theta,T}$ have support on N and N_θ respectively, this can equivalently be written:

$$\begin{aligned} &d_2(N_T, N_{\theta,T})^2 \\ &\quad = \inf_{\gamma \in \Gamma(\nu_T, \nu_{\theta,T})} \int_{N_T \times N_{\theta,T}} d_M(x_1, x_2)^2 d\gamma(x_1, x_2). \end{aligned}$$

We define a change of variables to compute the integral defining d_2. The function f^M is injective on its image by definition of an embedding. Without loss of generality, we assume that the function f_θ^M is injective on its image, i.e. that its differential is of full rank. If the differential is not of full rank, we approximate it by a full rank matrix as the space of invertible matrices is dense in the space of matrices, with a precision such that Equation 14 still holds. We consider the homeomorphism:

$$((f^M)^{-1}, (f_\theta^M)^{-1}) : N_T \times N_{\theta,T} \mapsto B_T \times B_T, \tag{16}$$

which is a "global chart" of the submanifold $N_T \times N_{\theta,T}$ of $M \times M$.

We write the integral defining d_2 in the chart defined by Equation 16. In other words, we perform the change of variables $(z_1, z_2) = ((f^M)^{-1}(x_1), (f_\theta^M)^{-1}(x_2))$ to represent the point (x_1, x_2). We get:

$$\begin{aligned} &d_2(N_T, N_{\theta,T})^2 \\ &= \inf_{\gamma \in \Gamma((f^M)^{-1} * \nu_T, (f_\theta^M)^{-1} * \nu_{\theta,T})} \\ &\qquad \int_{B_T \times B_T} d_M(f^M(z_1), f_\theta^M(z_2))^2 d\gamma(z_1, z_2) \\ &= \inf_{\gamma \in \Gamma(\mu_T, \mu_T)} \\ &\qquad \int_{B_T \times B_T} d_M(f^M(z_1), f_\theta^M(z_2)) d\gamma(z_1, z_2), \end{aligned}$$

where the probability density γ is expressed in this chart:

$$d\gamma(z_1, z_2) = \gamma(z_1, z_2) dz_1 dz_2, \tag{17}$$

and we have $(f_\theta^M)^{-1} * \nu_{\theta,T} = \mu_T$ by construction.

We use the definition of the inf to get an upper bound on d_2. From the definition of the Wasserstein distance, the notation $\Gamma(\mu_T, \mu_T)$ refers to the collection of measures on $B_T \times B_T$ with marginals μ_T on the first and second factors

respectively, for the Euclidean measure of $\mathbb{R}^L$. We consider the following element $\tilde{\gamma}$ of $\Gamma(\mu_T, \mu_T)$:

$$\tilde{\gamma}(z_1, z_2) = \mu_T(z_1)\delta_{z_1=z_2} = \mu_T(z_2)\delta_{z_1=z_2}. \quad (18)$$

By the property of the inf, we get the upper bound:

$$\begin{aligned} d_2(N_T, N_{\theta,T})^2 &\leq \int_{B_T \times B_T} d_M(f^M(z_1), f_\theta^M(z_2))^2 d\tilde{\gamma}(z_1, z_2) \\ &= \int_{B_T} d_M(f^M(z), f_\theta^M(z))^2 d\mu_T(z). \end{aligned}$$

We define the compact K of $T_\mu M$ as the compact $f(B_T)$ extended in L_2 norm by a distance $\sqrt{D}$ in all directions. We have, for any $z \in B_T$: $f(z) \in f(B_T) \subset K$ and $f_\theta(z) \in K$ following from inequality 14 since $\epsilon < 1$. The Riemannian exponential $Exp(\mu, \bullet)$ is continuous on the compact K, and therefore Lipshitz on K. There exists a constant $C_{T,D}$ such that for any $y_1, y_2 \in K$, we have:

$$d_M\left(Exp(\mu, y_1), Exp(\mu, y_2)\right) < C_{T,D}||y_1 - y_2||_2. \quad (19)$$

We apply this to $y_1 = f(z)$ and $y_2 = f_\theta(z)$ for any $z \in B_T$, and we take the square:

$$\begin{aligned} d_M^2\left(Exp(\mu, f(z)), Exp(\mu, f_\theta(z))\right) \\ < C_{T,D}^2||f^M(z) - f_\theta(z)||_2^2, \end{aligned}$$

which can be rewritten:

$$d_M^2\left(f^M(z), f_\theta^M(z)\right) < C_{T,D}^2||f^M(z) - f_\theta(z)||_2^2.$$

By definition of the sup, as an upper bound, we get:

$$\begin{aligned} d_M^2\left(f(z), f_\theta(z)^M\right) &< C_{T,D}^2 \sup_{z \in B_T} ||f(z) - f_\theta(z)||_2^2 \\ &\leq C_{T,D}^2 D\epsilon^2, \end{aligned}$$

where the last inequality comes from Equation 14. We integrate on both sides using the measure μ_T, and write:

$$\begin{aligned} d_2(N_T, N_{\theta,T})^2 &\leq \int_{\mathbb{R}^L} d_M(f^M(z), f_\theta^M(z))^2 d\mu_T(z) \\ &\leq \int_{\mathbb{R}^L} C_{T,D}^2 D\epsilon^2 d\mu_T(z) \\ &\leq C_{T,D}^2 D\epsilon^2. \end{aligned}$$

Therefore, for any T, any weighted submanifold N_T verifying the assumptions can be approximated by a submanifold $N_{\theta,T}$ generated by the model of a Riemannian VAE.

3. "Important remark" of Section 4.3

Section 4.3 of the paper highlights that the Riemannian VAE (rVAE) learning procedure does not boil down to projecting the manifold-valued data onto some tangent space of M and subsequently learning with a Euclidean VAE. We provide details about this statement here.

Consider a dataset $x_1, ..., x_n$ on the manifold M and a point $\mu \in M$, such that $x_i \in V(\mu)$ for all $i = 1, .., n$. Consider using a (Euclidean) VAE to fit a submanifold to the data $\log_\mu(x_1), ..., \log_\mu(x_n)$ in $T_\mu M \simeq \mathbb{R}^D$. The VAE makes the implicit assumption that the underlying distribution is:

$$\log_\mu(X)|Z \sim \mathcal{N}(f_\theta(Z), \sigma^2 \mathbb{I}_D), \quad (20)$$

and maximizes the associated (lower bound of the) log-likelihood.

In contrast, consider using a rVAE to fit a submanifold to the data $x_1, ..., x_n$. The rVAE makes the implicit assumption that the underlying distribution is:

$$X|Z \sim \mathcal{N}^M(f_\theta^M(Z), \sigma^2 \mathbb{I}_D), \quad (21)$$

where $f_\theta^M(z) = \text{Exp}(\mu, f_\theta(z))$ and maximizes the associated (lower bound of the) log-likelihood.

The generative models (20) and (21) are not equivalent. Taking the Riemannian log at μ of the data generated with model (21) is not equivalent to generating data with model (20). Figure 1 shows the difference in terms of the 2D histograms for generative models on the tangent space $T_\mu S^2$ of a 2D sphere S^2 with pole μ.

Figure 1 (a) shows the function $z \in \mathbb{R} f_\theta(z) T_\mu S^2 \simeq \mathbb{R}^2$. Figure 1 (b) shows the difference of 2D histograms between data generated from the Euclidean VAE model and data generated from the rVAE model and subsequently projected (in the sense of the Riemannian Logarithm) at the tangent space of their Fréchet mean. The sample size for both histograms is $n = 10^7$, with 80 bins in both dimensions.

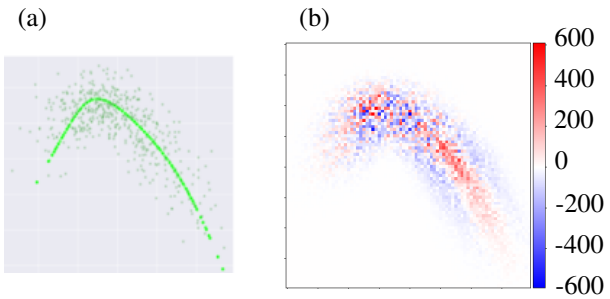

Figure 1. Comparaison of samples generated from the Euclidean VAE model and the "projected" Riemannian VAE model. (a) Underlying true submanifold in the tangent space of the sphere S^2. (b) Difference of 2D histograms between data generated from the Euclidean VAE model and from the "projected" Riemannian VAE model in the tangent space of the sphere S^2.

The noise's standard deviation coming from the rVAE model increases with the distance from the mean in the tangent space, which is also the origin of the tangent space. Due to the sphere's curvature, Riemannian balls on the sphere get deformed when projected on the tangent space.

The deformation increases with the distance from the origin of the tangent space. Consequently, the Riemannian Gaussian's level sets on the sphere get deformed on the tangent space. This behavior creates a curvature-induced heteroscedasticity of the noise.

4. Computations for Section 5.1

4.1. Model

We consider data generated with the model of probabilistic PCA (PPCA) with $\mu = 0$ [5], *i.e.* a special case of a rVAE model:

$$X_i = wZ_i + \epsilon_i, \tag{22}$$

where: $w \in \mathbb{R}^{D\times L}$, $Z \sim \mathcal{N}(0, \mathbb{I}_L)$ *i.i.d.* and $\epsilon \sim \mathcal{N}(0, \mathbb{I}_D)$ *i.i.d.*. The only parameter of model (22) is $\theta = w$. The true distribution of the data writes:

$$p_\theta(x) = N\left(0, ww^T + \mathbb{I}_D\right). \tag{23}$$

In this specific case, the true posterior of the latent variables z is tractable and writes: $p_\theta(z|x) = N\left(\Sigma w^T x, \Sigma = (\mathbb{I}_L + w^T w)^{-1}\right)$. This model could be learned through the EM algorithm, which converges to a maximum likelihood estimator of w, under regularity assumptions, and computes the associated tractable posterior. However, we train a VAE to illustrate the statistical inconsistency described in the main paper. We chose a variational family of Gaussian distributions with variance equal to 1:

$$\mathcal{Q} = \{\mathcal{N}(\phi, 1) \mid \phi \in \mathbb{R}^L\}, \tag{24}$$

parameterized by the unique parameter ϕ.

4.2. Landscape and results for $D = 1$ and $L = 1$

The VAE learns the parameters w, ϕ of the model, and approximate variational distribution by maximizing the evidence lower-bound (ELBO):

$$\mathcal{L}_1(x, \theta, \phi) \tag{25}$$
$$= l(\theta, x) - \mathrm{KL}\left(q(z|x) \parallel p(z|x)\right) \tag{26}$$

in $\theta = w$ and ϕ, where KL is the Kullback-Leibler divergence. The case $D = 1$ and $L = 1$ allows to compute the argument maxima in closed forms, as well as to represent the landspace of the ELBO as a function of the parameters (w, ϕ) in 2D as in Figure 2. We set $w = 2$ and a sample size $n = 1000$.

Figure 2 shows the landscape the ELBO as a function of (w, ϕ). The vertical green line represents the true value of the parameter: $w = 2$, while the vertical blue lines represent the maximum likelihood estimates (MLE) of w given a sample of size $n = 1000$. The MLE has two solutions, showing unidentifiability of the problem because of the symmetry in

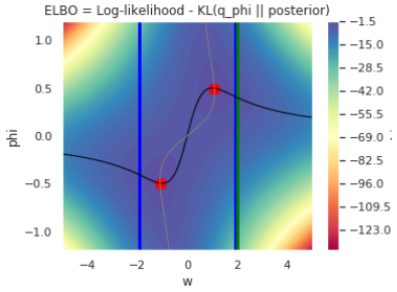

Figure 2. Criteria as functions of the parameters θ and ϕ. From left to right: log-likelihood, negative KL divergence to the true posterior and ELBO (nats).

model (22). The black curve represents the optimal variational distribution $q_\phi(z|x)$, for a given parameter w of the generative model. The grey curve represents the optimal member of the true posterior $p_w(z|x)$, for a given variational posterior $q_\phi(z|x)$. We denote $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n x_i^2$.

The optimal solutions would be $w_{\text{opt}} = w_{\text{mle}} = \sqrt{\hat{\sigma}^2 - 1}$ and $\phi_{\text{opt}} = \phi_{\text{opt}}(w_{\text{mle}})$, i. e. the intersection of the blue line and the black curve, or symmetric solution $w = -w_{\text{mle}}$. In other words, we would like the VAE to converge to the MLE for w, and subsequently computes the optimal variational distribution given this estimate. However, the VAE maximizes the ELBO and converges to one of the two red dots. Specifically, it converges to $w_{\text{elbo}} = \sqrt{\frac{\hat{\sigma}^2}{2} - 1}$ and associated $\phi_{\text{elbo}} = \phi_{\text{opt}}(w_{\text{elbo}})$. The VAE algorithm is, therefore, suboptimal for parameter learning. The KL to the posterior is optimized in both parameters simultaneously. Consequently, the two parameters collapse towards another. In Figure 2, both are attracted to 0. As the ELBO has the negative KL divergence as one of its terms, the parameter w is attracted to this optimal point of the negative KL. Another way of seeing it is as follows: the negative KL takes different values on the black curve, with a maximum close to 0. Thus, on this black curve, (w, ϕ) is attracted to 0.

4.3. Computations for $D = 1$ and $L = 1$

We present the computations for the results of the previous subsection. We compute the two terms of the ELBO, the log-likelihood, and the KL to the posterior. We maximize them in the parameters (w, ϕ).

We first compute the log-likelihood.

Lemma 2 *The log-likelihood of the generative model writes:*

$$L = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(1 + w^2) - \frac{\hat{\sigma^2}}{2(1 + w^2)},$$

where:

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n x_i^2. \tag{27}$$

A maximum likelihood estimator (MLE) for w verifies: $w^2 = \hat{\sigma^2} - 1$. There are two MLEs: $\hat{w} = \sqrt{\hat{\sigma^2} - 1}$ and $\hat{w} = -\sqrt{\hat{\sigma^2} - 1}$. The model is unidentifiable.

Proof 3 *The log-likelihood writes:*

$$
\begin{aligned}
L &= \mathbb{E}_{p_{data}(x)} \left[\log p(x)\right] \\
&= \mathbb{E}_{p_{data}(x)} \left[\log \mathcal{N}(0, 1 + w^2)\right] \\
&= \mathbb{E}_{p_{data}(x)} \left[-\frac{1}{2}\log(2\pi(1+w^2)) - \frac{x^2}{2(1+w^2)}\right] \\
&= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(1+w^2) - \frac{1}{2(1+w^2)}\mathbb{E}_{p_{data}(x)}x^2 \\
&\simeq -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(1+w^2) - \frac{1}{2(1+w^2)}\frac{1}{n}\sum_{i=1}^{n} x_i^2 \\
&= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(1+w^2) - \frac{\hat{\sigma^2}}{2(1+w^2)} \\
&\qquad \textit{where: } \hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2.
\end{aligned}
$$

The MLE for w is computed as follows:

$$
\begin{aligned}
\frac{\partial L}{\partial w} = 0 &\iff \frac{\partial L}{\partial w}\left(-\frac{1}{2}\log(1+w^2) - \frac{\hat{\sigma^2}}{2(1+w^2)}\right) = 0 \\
&\iff \left(-\frac{1}{2}\frac{2w}{1+w^2} + \frac{\hat{\sigma^2} 2.2w}{4(1+w^2)^2}\right) = 0 \\
&\iff \left(1 - \frac{\hat{\sigma^2}}{1+w^2}\right) = 0 \\
&\iff 1 + w^2 = \hat{\sigma^2} \\
&\iff w^2 = \hat{\sigma^2} - 1.
\end{aligned}
$$

Thus, there are two MLEs: $\hat{w} = \sqrt{\hat{\sigma^2} - 1}$ and $\hat{w} = -\sqrt{\hat{\sigma^2} - 1}$.

We turn to the KL between the approximate variational posterior and the true posterior.

Lemma 3 *The Kullback-Leibler divergence between the approximate posterior and the true posterior is:*

$$
\begin{aligned}
&KL\left(q_\phi(z|x) \parallel p_w(z|x)\right) \\
&= KL\left(\mathcal{N}(\phi x, 1) \parallel N\left(\frac{wx}{1+w^2}, \frac{1}{1+w^2}\right)\right) \\
&= -\frac{1}{2}\log(1+w^2) + \frac{1}{2}(1+w^2) + \frac{1}{2}\frac{\left(w - \phi - \phi w^2\right)^2}{1+w^2}x^2 \\
&\qquad - \frac{1}{2}.
\end{aligned}
$$

Furthermore, for a fixed w, the ϕ minimizing the above KL divergence is given by: $\phi = \frac{w}{1+w^2}$.

Proof 4 *The KL divergence between two multivariate Gaussians in L dimensions $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$ is:*

$$
\begin{aligned}
&KL(\mu_1, \Sigma_1 \parallel \mu_2, \Sigma_2) \\
&= \frac{1}{2}\left(\log\frac{|\Sigma_2|}{|\Sigma_1|} - L + tr(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)\Sigma_2^{-1}(\mu_2 - \mu_1)\right).
\end{aligned}
$$

Thus, the KL between the approximate posterior and the true posterior is:

$$
\begin{aligned}
&KL\left(q_\phi(z|x) \parallel p_w(z|x)\right) \\
&= KL\left(\mathcal{N}(\phi x, 1) \parallel N\left(\frac{wx}{1+w^2}, \frac{1}{1+w^2}\right)\right) \\
&= \frac{1}{2}\log\frac{1/(1+w^2)}{1} + \frac{1}{2/(1+w^2)} \\
&\qquad + \frac{\left(wx/(1+w^2) - \phi x\right)^2}{2/(1+w^2)} - \frac{1}{2} \\
&= \frac{1}{2}\log\frac{1}{1+w^2} + \frac{1}{2}(1+w^2) \\
&\qquad + \frac{\left(wx/(1+w^2) - \phi x\right)^2}{2/(1+w^2)} - \frac{1}{2} \\
&= -\frac{1}{2}\log(1+w^2) + \frac{1}{2}(1+w^2) \\
&\qquad + \frac{1+w^2}{2}\left(\frac{wx}{1+w^2} - \phi x\right)^2 - \frac{1}{2} \\
&= -\frac{1}{2}\log(1+w^2) + \frac{1}{2}(1+w^2) \\
&\qquad + \frac{1}{2(1+w^2)}\left(wx - \phi x(1+w^2)\right)^2 - \frac{1}{2} \\
&= -\frac{1}{2}\log(1+w^2) + \frac{1}{2}(1+w^2) \\
&\qquad + \frac{1}{2}\frac{\left(w - \phi - \phi w^2\right)^2}{1+w^2}x^2 - \frac{1}{2}.
\end{aligned}
$$

The expressions of the log-likelihood and the KL to the true posterior can be combined to compute the ELBO.

Lemma 4 *The ELBO writes:*

$$
\begin{aligned}
&\mathcal{L}_1(x, \theta, \phi) \\
&= l(\theta, x) - \mathbb{E}_{p_{data}(x)}\left[KL\left(q(z|x) \parallel p(z|x)\right)\right] \\
&= \frac{1}{2}(1 - \log(2\pi)) - \frac{1}{2}\log\left(\frac{1+w^2}{w^2}\right) \\
&\qquad - \frac{1}{2}w^2 - \frac{\hat{\sigma^2}}{2}\left(\frac{1}{1+w^2} + (1 - \phi w)^2\right),
\end{aligned}
$$

where: $\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$.

Proof 5 *The ELBO writes:*

$$
\begin{aligned}
&\mathcal{L}_1(x,\theta,\phi)\\
&= l(\theta,x) - \mathbb{E}_{p_{data}(x)}\left[KL\left(q(z|x) \parallel p(z|x)\right)\right]\\
&= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(1+w^2) - \frac{\hat{\sigma^2}}{2(1+w^2)}\\
&\qquad - \mathbb{E}_{p_{data}(x)}\left[-\frac{1}{2} - \log w + \frac{1}{2}w^2 + \frac{1}{2}(1-\phi w)^2 x^2\right]\\
&= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(1+w^2) - \frac{\hat{\sigma^2}}{2(1+w^2)}\\
&\qquad + \frac{1}{2} + \log w - \frac{1}{2}w^2 - \frac{1}{2}(1-\phi w)^2 \mathbb{E}_{p_{data}(x)}\left[x^2\right]\\
&= \frac{1}{2}(1-\log(2\pi)) + \log w - \frac{1}{2}\log(1+w^2) - \frac{1}{2}w^2\\
&\qquad - \frac{\hat{\sigma^2}}{2}\left(\frac{1}{1+w^2} + (1-\phi w)^2\right)\\
&= \frac{1}{2}(1-\log(2\pi)) - \frac{1}{2}\log\left(\frac{1+w^2}{w^2}\right) - \frac{1}{2}w^2\\
&\qquad - \frac{\hat{\sigma^2}}{2}\left(\frac{1}{1+w^2} + (1-\phi w)^2\right).
\end{aligned}
$$

Fixing w, the argument minimum of $\mathrm{KL}\left(q_\phi(z|x) \parallel p_w(z|x)\right)$ for ϕ is $\phi_{\text{opt}}(w) = \frac{w}{1+w^2}$. The function $w \to \phi_{\text{opt}}(w)$ is represented by the black curve on Figure 1 of the paper. This is the optimal variational distribution, given a parameter w of the generative model. Fixing ϕ, the argument minimum of $\mathrm{KL}\left(q_\phi(z|x) \parallel p_w(z|x)\right)$ for w is $w_{\text{opt}}(\phi) = \frac{\phi\hat{\sigma^2}}{\phi^2\hat{\sigma^2}+1}$. The function $\phi \to w_{\text{opt}}(\phi)$ is represented by the grey curve on Figure 2. This is the optimal "true" posterior, given a value of the variational posterior.

5. Details on experiments of Section 6

We provide details on the experiments comparing the following methods: VAE, rVAE, and VAE projected.

5.1. Synthetic data

We use a "true" decoder to generate synthetic data. The latent space is $\mathbb{R}$. The decoder has two fully connected layers of dimension 2 each. The model writes:

$$f_\theta(z) = w_2 g(w_1 z + b_1) + b_2, \tag{28}$$

where we choose $w_1 = \begin{pmatrix} 3 \\ -2 \end{pmatrix}$, $b_1 = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$, $w_2 = \begin{pmatrix} 0.05 & -0.5 \\ -0.15 & -0.1 \end{pmatrix}$ and $b_2 = \begin{pmatrix} 0.062 \\ 0.609 \end{pmatrix}$. We use the softplus as the nonlinearity g. We use the Riemannian exponential of the sphere to shoot the submanifold on the sphere. We add noise of variance σ^2, so that we have:

$$X \sim \mathcal{N}^M(\mathrm{Exp}(\mu, f_\theta(z)), \sigma^2), \tag{29}$$

where the Fréchet mean μ is a point on the sphere, the image of $(0,0)$ by the Riemannian exponential.

5.2. Architectures of VAE, rVAE and VAE projected

We train a VAE, a rVAE, and a VAE projected on the generated data.

The rVAE decoder has the same architecture as the decoder that has produced the data, *i.e.* a one-dimensional latent space, and two fully connected layers with 2 units each, and parameters $w_1, b_1, b_2 \in \mathbb{R}^2$ and $w_2 \in \mathbb{R}^{2\times 2}$.

The VAE decoder has an architecture with a one-dimensional latent space, and two fully connected layers with 2 and 3 units respectively, and parameters: $w_1, b_1 \in \mathbb{R}^2$ and $w_2 \in \mathbb{R}^{3\times 2}$, $b_2 \in \mathbb{R}^3$. The VAE projected is the result of the VAE trained on the data and projected back on the sphere. Therefore its architecture is the VAE's.

5.3. Visual comparison of the methods

Figure 3 shows the true submanifold in light green, as well as the result of the training of the PGA, VAE, the Riemannian VAE, and the VAE projected back on the sphere for $n = 10k$ and $\sigma = 0.35$. We observe that PGA is off, as it cannot learn a nongeodesic subspace. The extrinsic VAE learns a submanifold in the embedding space. The rVAE and VAE projected give very similar results.

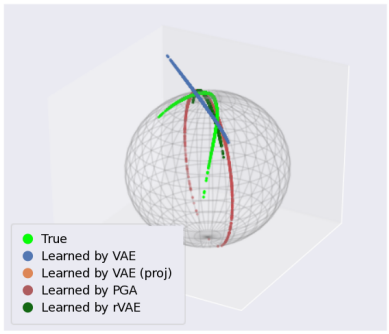

Figure 3. Visual comparison of the submanifold learning methods for $n = 10k$ and $\sigma = 0.35$.

6. Details on experiments of Section 7

6.1. Brain connectomes dataset

We give details on the Human Brain Connectome dataset [6]. The "1200 Subjects release" includes 3T resting-state functional MRI (rs-fMRI) imaging data from 1206 healthy young adult participants. We focus on the sub dataset "R812" which includes 812 subjects with rs-fMRI data reconstructed with the latest reconstruction algorithm.

We choose a parcellation of the brain into $N = 15$ regions or "nodes" and use the subject-specific sets of "node time-series" provided by HCP. Each subject is represented by 15 time-series that correspond to the activation of each

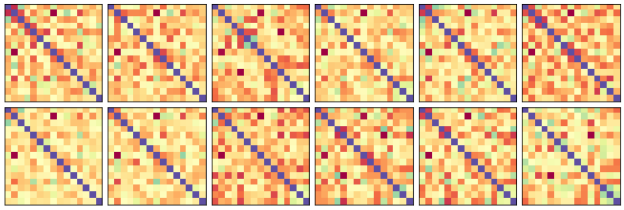

Figure 4. Elements from the dataset of HCP brain connectomes [6], with 15 nodes in each connectome.

of the 15 brain regions through time. From each subject-specific time-series, we build the subject-specific parcellated connectome, which is the 15x15 correlation matrix corresponding to the correlations between nodes. Each of the $n = 812$ subjects is, therefore, represented by a 15x15 symmetric positive definite (SPD) matrix. Elements randomly sampled from this dataset are shown in Figure 4.

6.2. Riemannian geometry of SPD matrices

The space of symmetric positive definite (SPD) matrices in N dimensions is a manifold defined as:

$$\text{SPD}(N) = \left\{S \in \mathbb{R}_{N\times N} : S^T = S, \forall z \in \mathbb{R}^N, z \neq 0, z^T S z > 0\right\}.$$

It has dimension $D = \frac{N(N-1)}{2}$. The manifold $\text{SPD}(N)$ can be equipped with different Riemannian metrics, for example the affine-invariant metric or the Log-euclidean metric. We note that $\text{SPD}(N)$ with either the Log-euclidean metric or the affine-invariant metrics is a Hadamard manifold.

6.3. VAE and rVAE architectures

We present the details of the VAE and rVAE architectures used in the experiments on the HCP dataset [6]. The decoder f_θ is a fully connected neural network. We did not choose a convolutional network as SPD matrices are not images. The layers have dimensions L, L^2 and D, where L is the dimension of the latent space and $D = 120$ the dimension of the data space. In other words:

$$f_\theta(x) = w_2^f g(w_1^f x + b_1^f) + b_2^f, \tag{30}$$

for $b_1^f \in \mathbb{R}^{L^2}$, $w_1^f \in \mathbb{R}^{L\times L^2}$, $b_2^f \in \mathbb{R}^D$, $w_2^f \in \mathbb{R}^{L^2\times D}$. We use ReLU as the activation function g. For the Riemannian VAE, we add the Riemannian exponential map Exp_μ at μ after the last layer, where μ is chosen to be the identity matrix $\mu = \text{Id}_{N\times N}$. As a consequence, Log_μ and Exp_μ correspond to the matrix logarithm and the matrix exponential.

The encoder is a fully connected neural network, with three layers of dimensions $D = 120$, L^2 and L. In other words:

$$\mu_\phi(x) = w_2^\mu g(w_1 x + b_1) + b_2^\mu,$$
$$\log \sigma_\phi^2(x) = w_2^\sigma g(w_1 x + b_1) + b_2^\sigma,$$

for $b_1 \in \mathbb{R}^{L^2}$, $w_1 \in \mathbb{R}^{D\times L^2}$, $b_2^\mu, b_2^\sigma \in \mathbb{R}^L$ and $w_2^\mu, w_2^\sigma \in \mathbb{R}^{L^2\times L}$. We use ReLU as the nonlinearity g.

6.4. VAE and rVAE training

Each SPD matrix is flattened and represented as a vector x of dimension D that corresponds to its upper half. We denote $\log(x)$ the vector representing the upper half of the matrix logarithm of the SPD matrix. The probability density functions of the noise models write in the Euclidean space:

$$p_\theta(x|z) = \frac{1}{\sqrt{(2\pi)^D \sigma^{2D}}} \exp\left(-\frac{||x - f_\theta(z)||^2}{2\sigma^2}\right), \tag{31}$$

and, in the Riemannian space:

$$p_\theta(x|z) = \frac{1}{\sqrt{(2\pi)^D \sigma^{2D}}} \exp\left(-\frac{||\log(x) - f_\theta(z)||^2}{2\sigma^2}\right). \tag{32}$$

The integration constant only depends on σ, as the Riemannian manifold is Hadamard. The losses write respectively:

$$\mathcal{L}^{\text{VAE}}(x^{(i)}) = \frac{1}{2}\sum_{l=1}^{L}\left(1 + \log(\sigma_l^{(i)})^2 - (\mu_l^{(i)})^2 - (\sigma_l^{(i)})^2\right) - \log\sqrt{(2\pi)^D\sigma^{2D}} - \frac{||x - f_\theta(z)||^2}{2\sigma^2},$$

and:

$$\mathcal{L}^{\text{rVAE}}(x^{(i)}) = \frac{1}{2}\sum_{l=1}^{L}\left(1 + \log(\sigma_l^{(i)})^2 - (\mu_l^{(i)})^2 - (\sigma_l^{(i)})^2\right) - \log\sqrt{(2\pi)^D\sigma^{2D}} - \frac{||\log(x) - f_\theta(z)||^2}{2\sigma^2}.$$

References

[1] Thomas Fletcher, Conglin Lu, Stephen M Pizer, and Sarang Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE transactions on medical imaging*, 23(8):995–1005, 2004. 2

[2] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991. 3

[3] Xavier Pennec, Stefan Sommer, and Tom Fletcher. *Riemannian Geometric Statistics in Medical Image Analysis*. Elsevier Ltd, first edit edition, 2019. 1

[4] Mikhail Postnikov. *Riemannian Geometry*. Encyclopaedia of Mathem. Sciences. Springer, 2001. 1, 2, 3

[5] Michael E. Tipping and Christopher M. Bishop. Probabilistic Principal Component Analysis. *Source: Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(3):611–622, 1999. 6

[6] David C. Van Essen, Stephen M. Smith, Deanna Barch, Timothy E. J. Behrends, Essa Yacoub, and Kamil Ugurbil. The WU-Minn Human Connectome Project: An Overview David. *Neuroimage*, 80:62–79, 2013. 8, 9