Learning Riemannian geometry for mixed-effect models using deep generative networks.

Maxime Louis, Benjamin Charlier, Stanley Durrleman

Learning Riemannian geometry for mixed-effect models using deep generative networks.

Maxime Louis [1,2] **Benjamin Charlier**[3,1,2]**, Stanley Durrleman**[1,2]
[1]Inria Paris, Aramis project-team, 75013, Paris, France
[2] Sorbonne Universités, UPMC Univ Paris 06, Inserm, CNRS, Institut du cerveau et de la moelle (ICM)
[3]Institut Montpelliérain Alexander Grothendieck, CNRS, Univ. Montpellier

Abstract

We take up on recent work on the Riemannian geometry of generative networks to propose a new approach for learning both a manifold structure and a Riemannian metric from data. It allows the derivation of statistical analysis on manifolds without the need for the user to design new Riemannian structure for each specific problem. In high-dimensional data, it can learn non diagonal metrics, whereas manual design is often limited to the diagonal case. We illustrate how the method allows the construction of a meaningful low-dimensional representation of data and exhibit the geometry of the space of brain images during Alzheimer's progression.

1 Introduction

There has been a lot of interest into statistical methods for the analysis of manifold-valued data. These methods aim at proposing nonlinear alternatives to usual linear models when the data lies on a Riemannian manifold: the Fréchet mean [13] is an extension of the linear mean, principal geodesic analysis (PGA) [8] is an extension of Principal Component Analysis (PCA), geodesic regression [9] is an extension of linear regression and [25] proposes unsupervised clustering of manifold-valued data. In these approaches, the Riemannian manifold and its metric are fixed and typically encode the known constraints on the data, such as positive constraints or positive-definite matrices. These known constraints do not preclude an even more constrained structure of a particular distribution of points on this manifold, that we would like to estimate.

Consequently, a number of works [26, 17, 12, 19] offer unsupervised methods to perform manifold learning from data. However, only a few of these methods do obtain a parametrization of a submanifold of the space of observations –most of them do it implicitly such as [26, 17, 19]– and almost none of them do estimate a Riemannian metric on the learned manifold: they often use the geometry induced by a standard metric on the space of observations. We argue here that the estimation of a Riemannian metric is an important part of the learning process which complements the manifold learning part. Indeed, there is no reason for the geometry induced by an usual metric on the data to be particularly relevant for machine learning tasks.

To gain a better understanding of the geometry of learning, there has been recent work on deep generative networks and on the geometry of the associated latent spaces [2, 24, 11, 5]. Deep generative models -such as auto-encoders [12] or generative adversarial networks [10]- learn an immersed submanifold of the observation space and a system of coordinates on this submanifold. In these papers, the geometry of the latent space is inferred a posteriori by pulling back the geometry of the observations onto the latent space. In [24] the authors notice, experimentally, that the latent space is approximately flat and linear with this geometry.

Consequently, we propose in this paper to reverse the mechanic. Given a mixed-effect model with a Euclidean structure, we reformulate it on a low-dimensional Euclidean space and use a deep

Preprint. Work in progress.

generative network to learn a mapping from this latent space to the observation space. The image of the network is a low-dimensional immersed submanifold that we equip with the push-forward of the Euclidean metric. We illustrate this procedure on two examples:

- An extension of Probabilistic Principal Component Analysis (PPCA) [27]: we formulate a nonlinear generalization. We show how this formulation is close to Probabilistic Principal Geodesic Analysis (PPGA) [28].
- An unsupervised longitudinal model: we generalize the formulation proposed in [23] to allow inference of a submanifold of the space of observations and of a Riemannian metric on this manifold adapted to the unsupervised task.

In Section 2, we describe the construction of a latent space and of the push-forward of its Euclidean geometry onto the space of observations and we apply it to PPCA and to a longitudinal model. In Section 3, we describe the inference procedure used in the experiments presented in Section 4.

2 Geometrical model

2.1 Geometry of generative networks

We describe a generative neural network as a parametric function $\Psi_w : U \to V$ where U is an open subset of $\mathbb{R}^d$, V is an open subset of $\mathbb{R}^D$ where $d, D \in \mathbb{N}$ with typically $d \ll D$ and w are the neural network weights. In [24], the authors show that $\Psi_w(U)$ is a d-dimensional immersed submanifold of $\mathbb{R}^D$ if the activation functions of each layer are smooth and monotonic and if the weight matrix of each layer has maximal rank. The first condition is easy to enforce and we check the second condition after training. Note that an immersed submanifold is not a submanifold in general, but it is *locally* a submanifold.

A Riemannian metric on a smooth manifold $\mathcal{M}$ is a smoothly varying inner product on the tangent bundle $T\mathcal{M}$. In [24, 5], the authors illustrate how to pull-back a metric on $\mathbb{R}^D$ to the latent space $\mathbb{R}^d$. They note, experimentally, that the induced metric on the latent space is almost flat. Based on this empirical observation, we postulate that there exists a latent space where the pull-back of the Riemannian structure describing the data is Euclidean.

As a consequence, we propose to go the opposite way by pushing-forward the Euclidean metric of a latent space onto a submanifold of the observation space. Doing so enforces by construction the flatness of the latent space during learning. Let g be the Euclidean metric on U, we can define the push-forward of g on $\Psi_w(V)$. For any smooth vector fields X, Y on $\Psi_w(U)$, it is defined as:

$$\Psi_w^*(g)(X, Y) = g((\Psi_w)_*(X), (\Psi_w)_*(Y))$$

where $(\Psi_w)_*(X)$ and $(\Psi_w)_*(Y)$ are the pull-back of X and Y on U defined by $(\Psi_w)_*(X)(f) = X(f \circ \Psi_w^{-1})$ for any smooth function $f : U \to \mathbb{R}$. If $\gamma : [0, 1] \to U$ is a geodesic on (U, g), then $\Psi_w \circ \gamma : [0, 1] \to V$ is a geodesic on $(\Psi_w(U), \Psi_w^*(g))$. Finally, for any p on a manifold $\mathcal{M}$, we recall that the Riemannian exponential map Exp_p is defined on a neighborhood of 0 in $T_p\mathcal{M}$ by $\mathrm{Exp}_p(v) = \gamma(1)$ where γ is the unique geodesic with $\gamma(0) = p$ and $\dot{\gamma}(0) = v$.

Note that the function Ψ_w parametrizes:

- A submanifold $\Psi_w(U)$ of the space of observations
- A metric $\Psi_w^*(g)$ on this submanifold

We show in the next section how to adapt mixed-effect models which assume a linear structure of the observation space into models which consider a similar structure on a submanifold of the observation space described by a latent space and a neural network Ψ_w.

2.2 Generalizing mixed-effect models to Riemannian manifolds.

We consider generative models of the form:

$$y_i = Fx_i + Gu_i + \varepsilon_i \tag{1}$$

where $y_i \in \mathbb{R}^D$ is an observation, $x_i \in \mathbb{R}^f$ is known, u_i is a latent random variable in $\mathbb{R}^g$, F and G are unknown matrices, and $\varepsilon \sim \mathcal{N}(0, \Sigma_\varepsilon)$ is a Gaussian noise. We also suppose a normal prior on the latent variable: $u \sim \mathcal{N}(0, \Sigma_u)$. We propose nonlinear generalizations of two different models of the form (1): PPCA and a longitudinal model.

Generalizing PPCA. As shown in [27], PCA is the Maximum Likelihood (ML) estimate of the PPCA model which writes $y_i = Gu_i + \varepsilon_i$ where u is the latent position with prior $u \sim \mathcal{N}(0, I_d)$, G is a $D \times d$ matrix containing the principal directions and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_D)$. This is a model of the form (1) with no covariates x_i. We propose to generalize the model to:

$$y_i = \Psi_w(u_i) + \varepsilon_i. \tag{2}$$

where Ψ_w is a parametric family of $\mathcal{C}^1$ immersions. PPCA describes each observation as a point on a linear subspace with coordinates u_i while the proposed formulation (2) describes each observation as a point on a submanifold of $\mathbb{R}^D$ with coordinates u_i. G has been removed on the new model, as the neural network is flexible enough to convert the isotropic unit normal distribution of the latent variable u into a distribution on the submanifold $\Psi(\mathbb{R}^d)$ which is close to the data distribution.

To illustrate the similarity with PPGA, we denote $J = \text{Jac } \Psi_w(0)$ the Jacobian of Ψ_w and write, for any observation y_i close to $\Psi_w(0)$:

$$y_i = \Psi_w(u_i) + \varepsilon_i = \text{Exp}_{\Psi_w(0)}((\Psi_w)_*(u_i)) + \varepsilon_i = \text{Exp}_{\Psi_w(0)}(Ju_i) + \varepsilon_i. \tag{3}$$

where the first step is by construction of the Euclidean latent space $\mathbb{R}^d$ and the second is by definition of $(\Psi_w)_*$. If Ψ_w was fixed –which amounts to fixing the manifold structure– the estimation of J by ML in (3) is almost equivalent to the PPGA proposed in [28] (to the detail of the structure of the noise, which we assume isotropic Gaussian in $\mathbb{R}^D$ when it is assumed isotropic Gaussian on $\mathcal{M}$ for the PPGA). One can also note the proximity with a variationnal auto-encoder as described in [15]. Finally, model (2) should be compared to the Locally Adaptive Normal Distribution (LAND) proposed in [1], which proposes to learn a Riemannian metric so that the observations are distributed along a normal distribution for this Riemannian metric. Indeed, after inference of the model (2), the Riemannian metric will have been learned so that the observations are approximately distributed according to a normal distribution on the manifold $(\Psi_w(U), \Psi_w^*(g))$, as is the case for LAND. But our model goes beyond LAND by learning a submanifold of the space of observations and a metric which is not necessarily diagonal, while staying computationally efficient in high dimension.

Generalizing a model for the analysis of longitudinal data. In [23], the authors propose a framework for the analysis of manifold-valued trajectories. They assume an a priori Riemannian geometry on the space of observations. We denote $(y_{ij})_{j=1,\ldots,n_i} \in \mathbb{R}^D$ the observations of the subject i, measured at times $(t_{ij})_{j=1,\ldots,n_i}$. In the case of a linear manifold, the model writes:

$$y_{ij} = v_0 \Phi_i(t_{ij}) + As_i + \varepsilon_{ij} \tag{4}$$

where $v_0 \in \mathbb{R}^D$, $\Phi_i(t) = \exp(\eta_i)(t - \tau_i)$ is the time reparametrization of the individual trajectory: for disease modelling, $\exp(\eta_i) > 0$ controls the pace of progression of the subject and $\tau_i \in \mathbb{R}$ controls the time-shift for the subject progression. A is a $D \times d$ matrix with $d \leq D - 1$ which contains directions orthogonal to v_0, which allow to take into account different positions of the trajectories for different subjects. The latent variables are $u_i = (\eta_i, \tau_i, s_i)$ with priors $\eta_i \sim \mathcal{N}(0, \sigma_\eta)$, $\tau_i \sim \mathcal{N}(0, \sigma_\tau)$ and $s_i \sim \mathcal{N}(0, I_{d-1})$. σ_τ and σ_η are initialized from the data and fixed during the estimation. Ignoring the time reparametrization Φ_i, the model (4) is of the form (1). We use a similar generalization procedure as for (2) and write the new generative model:

$$y_{ij} = \Psi_w \left(e_1 \Phi_i(t_{ij}) + \sum_{l=2}^{d} s_{il} e_l \right) + \varepsilon_{ij} \tag{5}$$

where $(e_1, \ldots, e_d)$ is the canonical basis of $\mathbb{R}^d$. $t \mapsto \Psi_w(e_1 t)$ should be interpreted as the geodesic of mean progression, while the directions $(\Psi_w)_*(e_i)$ for $i \in \{2, \ldots, d\}$ capture variability between subjects. As for PPCA, v_0 and A have been removed from the new model since the neural network is flexible enough to send e_1 onto a main direction of progression and $(e_2, \ldots, e_d)$ onto vector fields which are orthogonal to this direction of projection. Figure 1 summarizes the construction. Note that as in [23], the subject trajectories are obtained by a parallel shift of a mean geodesic in the directions $e_2, \ldots, e_n$ on the manifold described by Ψ_w. This can be interpreted as the Riemannian analog of linear translation.

Both generalizations share the similar idea of keeping the structure of the initial model in a low-dimensional space $\mathbb{R}^d$ and using a neural network to transport this structure onto a submanifold of

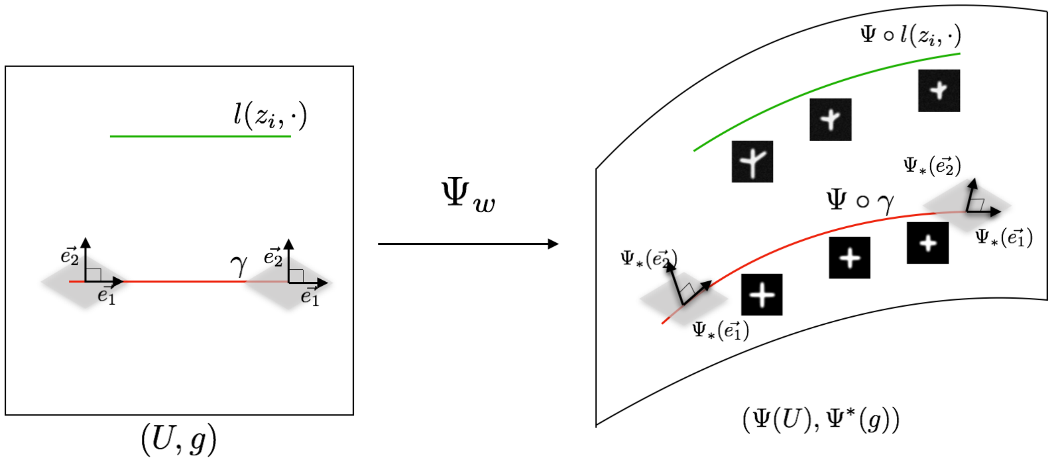

Figure 1: Transporting the Euclidean geometry onto a submanifold of the observation space using the neural network Ψ_w. $t \mapsto \Psi_{(}e_1 t)$ is the geodesic of mean progression.

the space of observations. This general principle of converting the linear structure into an equivalent nonlinear structure can be adapted to numerous models such as Probabilistic Linear Discriminant Analysis [21], Random slopes/intercepts models or linear regression.

Finally, in both cases, Ψ_w being an isometry between $(\mathbb{R}^d, g)$ and $(\Psi_w(\mathbb{R}^d), \Psi_w^*(g))$, any computation of Riemannian exponential, logarithm or parallel transport can be done in $\mathbb{R}^d$ before being push-forward to the observation space. Thus, these operations, which can be costly in a Riemannian setting, are inexpensive in our setting. Besides, the analysis of the data can be done directly in the latent space $\mathbb{R}^d$ whose Euclidean structure is faithful to the nonlinear geometry of the data.

3 Inference

We describe here the inference procedure for models (2) and (5). The inference consists in finding the Maximum A Posteriori (MAP) of a directed probabilistic model with latent variables u. A possibility is to use variationnal Bayes approaches such as [15]. An alternative is to use the Expectation-Maximization (EM) algorithm. The E step requires the computation of integrals of the form $\int_u \log\left(p(y|u, \theta)\right) p(u, \theta_k) \mathrm{d}u$ which are intractable in our case, so we resort the the Stochastic Approximation EM (SAEM) [6] which alternates:

- *Simulation.* For each observation y_i, generate u_i, a realization of the hidden variable under the posterior density $p(u|y_i, \theta_k)$.
- *Approximation.* Update $Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k(\sum_i p(y_i|u_i, \theta) - Q_{k-1}(\theta))$
- *Maximization.* Set $\theta_{k+1} = \mathrm{argmax}_\theta Q(\theta)$.

where during L burn-in iterations $\gamma_k = 1$ and then for all $k > L$, $0 \leq \gamma_k \leq 1$, $\sum_{i=1}^{\infty} \gamma_k = \infty$ and $\sum_{i=1}^{\infty} \gamma_k^2 < \infty$. Once again, this procedure is intractable in our case, since the maximization step cannot be computed at a reasonnable cost. We therefore replace the Approximation step by simply setting $Q(\theta) = \sum_i p(y_i|u_i, \theta)$, which can be optimized by stochastic gradient descent. This amounts to keeping only the burn-in phase of the SAEM ($\gamma_k = 1$) which is, as we noted empirically, the most important phase with respect to space exploration of the individual variables u_i.

Simulation-Expectation To perform the Simulation step, we use the symmetric Hasting-Metropolis sampler [20], a Markov Chain Monte Carlo method. We run 25 iterations of the Markov Chain for each simulation, to limit samples correlation.

Maximization The maximization can be performed by stochastic gradient descent on Q with respect to the neural network weights w. We run ten epochs of gradient descent at each maximization, using Adam [14]. The noise variance σ_ε can be updated using a closed-form expression derived from the log-likelihood for both models (2) and (5). Performing the gradient descent repeatedly with slightly changing latent variables seems to have a stabilizing effect on the neural network learning: it fills the latent space around each subjects and can be interpreted as a data augmentation procedure.

Algorithm 1 Inference procedure for models (2) and (5)

```
input  : Observations $(y_i)_i$, fixed effects $(\beta_i)_i$, initial parameters $\theta^0 = (w^0, \sigma_\varepsilon^0)$ and samples $u_i^0$.
output : Estimation of $\theta_{\text{MAP}}$ and of samples $u_i$ distributed following $p(u \mid (y_i), \theta_{\text{MAP}})_i$.
repeat
    Simulation-Expectation: Draw a candidate $u_i^c$ from the proposal distribution $p_b(.|u_i^k, y_i)$ for all $i$
    Compute the acceptance ratio $\tau = \frac{p(y_i|\theta^k, u_i^c) p(u_i^c|\theta^k)}{p(y_i|\theta^k, u_i^k) p(u_i^k|\theta^k)}$ and accept $u_i^c$ with probability $\min(\tau, 1)$.
    Maximization: Update $\sigma_\varepsilon$ using the closed form expression.
    Update $w^k$ by stochastic gradient descent:
    for $epoch \leftarrow 0$ to $10$ do
        for batch $b$ in $(y_i, u_i)_i$ do
            Compute $Q = \sum_{y,u \in b} p(y|u, \theta^k)$
            Update $w$ by stochastic gradient descent $w = w - \alpha \cdot \nabla_w Q$ or using Adam.
        end
    end
until convergence;
```

4 Experiments

The network architectures used in the experiments which follow are standard and given in the supplementary materials. We only use $\mathcal{C}^1$ transfer functions to obtain a family of $\mathcal{C}^1$ functions.

A python code for the experiments and the data sets will be made available upon publication of the paper. Note that both the E and the M steps are massively parallelisable: the samples for each subjects are independent and the neural network training can easily be performed on multiple CPUs or GPUs. 200 iterations of the proposed method take an average of 1h30 on 4 CPUs for the longitudinal model, and 10 minutes for the generalized PPCA on a subset of 1000 digits from MNIST.

4.1 Generalized PPCA on MNIST

We estimate the model (2) on 1000 randomly selected digits from MNIST [18]. The mean squared error (MSE) for different dimensions d are shown on Figure 3. The proposed approach consistently beats PCA on both train and held-out sets. To analyze the effect of the Riemannian metric learning, we compare four dimension reduction methods: the first two PCA components, Isomap [26] on the raw images, t-SNE [19] and MDS [17] performed on the latent positions u of the images after estimation. MDS uses the Euclidean distances on the latent positions u of the observations, which amounts by construction to using the geodesic distances between the observations. It is therefore a direct representation of the learned Riemannian manifold structure. Figure 2 shows the results. All methods are run with default parameters. Interestingly, although the learning is unsupervised, the manifold geometry did capture class distributions in a much more clustered way than the other methods. This underlines the importance of estimating a metric on the manifold and of not relying on usual metrics which may not be adapted to the geometry of the data set.

4.2 Generalized longitudinal model

4.2.1 Recovering a synthetic geometry

We perform experiments of model (5) on a synthetic data set. Each observation is a 64×64 gray-level image of a cross fully described by 3 parameters: the angles of the right and left arm and the length of the arms. The arm angles are drawn along a zero-centered normal distribution. A mean progression scenario is prescribed for the arm lengths. Figure 5 shows synthetic subjects. Note that each image is defined by a finite number of parameters with respect to which it varies smoothly: hence the generated set of images belongs to a 3-dimensional submanifold of the set of 64×64 grey level images.

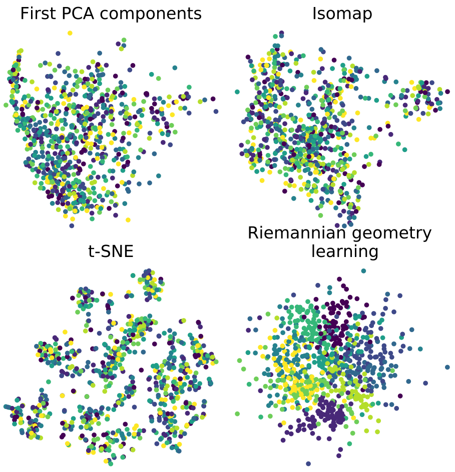

Figure 2: PCA, Isomap, t-SNE and Riemannian geometry learning on MNIST. Colors indicate labels.

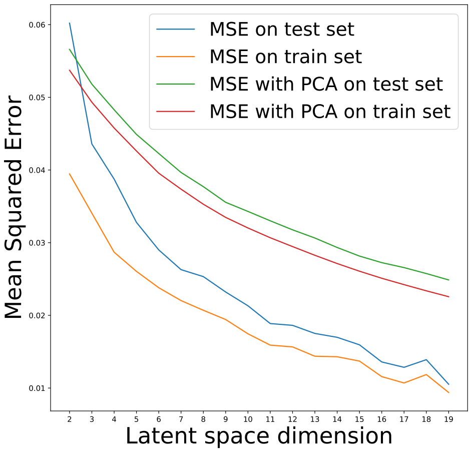

Figure 3: MSE with standard PCA and model (2) for varying dimension d.

We generate thirty different data sets with 200 subjects and 7 observations per subject according to this distribution, 25% of the subjects are kept in a test set. We add a white noise on the images with standard deviation of 0.025, 0.05 or 0.075. We estimate the model with $d = 3$ on the train set. Figure 6 show the obtained geodesic progression and two different patterns $\Psi_w(v_0 t + e_i)_{i=2,\ldots,3}$. Figure 4 (left) shows the estimation of the noise variance, averaged over the ten folds for each noise level.

We then fit the test observations onto the obtained model by gradient descent on the individual parameters u. We compare the MSE of reconstruction between train and test set, to check if it generalizes well. We also compute the MSE between the reconstructed images and the original images without the added noise to see that the generative network has a denoising effect. The results are given on Figure 4 (right). Interestingly, for all the noise levels, the model, which was trained on the noisy images, recovers equally well the original images without noise, suggesting a resilience of the model with respect to noise component outside of the data manifold. These experiments indicate the ability of the model to capture the synthetic submanifold and its Riemannian structure.

4.2.2 Cognitive scores

We use the cognitive scores grading the subjects memory, praxis, language and concentration extracted from the ADNI database as in [23] and run the model to estimate the geometry adapted to the progression of the cognitive scores. We emphasize that in [23], and in its derivations for other types of data [4, 16], the authors consider user-defined metrics specifically manufactured for the data set. Here, we propose a generic way to learn both the manifold and metric regardless of the kind of data considered. We do recover a geometry which is similar to the one postulated in [23] with logistic-like geodesics, as shown on Figure 7, although this geometry was not prescribed a priori.

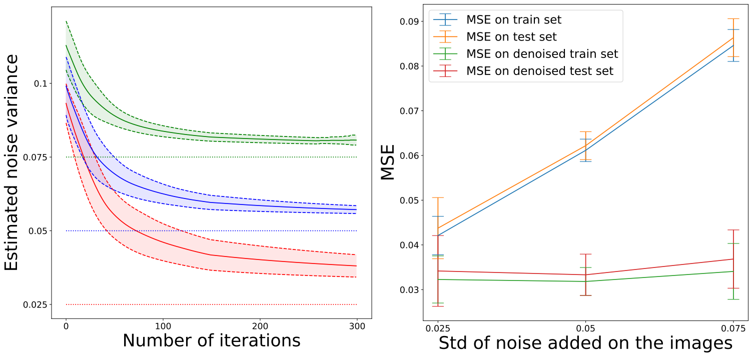

Figure 4: **Left**: Noise variance and standard deviation vs number of iterations for each noise level 0.025, 0.05 and 0.075, averaged over ten folds. Dashed line are the simulation noises. **Right**: MSE of the model with train and test, noised or denoised, images.

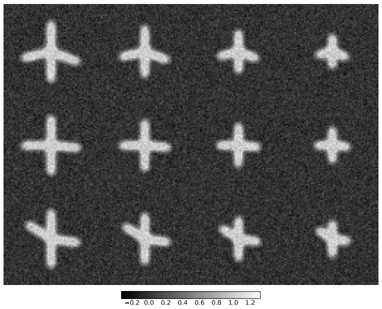

Figure 5: Each row is a synthetic subject.

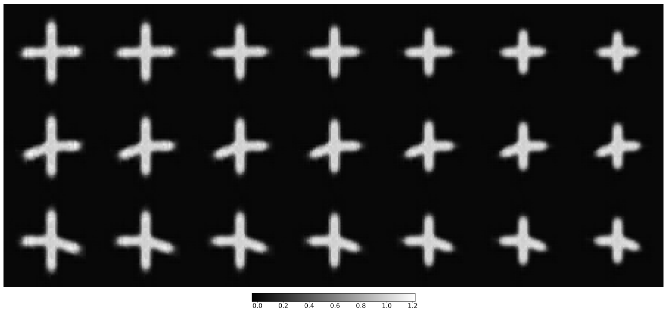

Figure 6: Top row is $t \mapsto \Psi_w(e_0 t)$. Following rows are $t \mapsto \Psi_w(e_0 t + e_i)$ for $i \in \{2, 3\}$.

In addition, our generalization allows more flexibility with different behaviors for each score and non-parallel geodesics. The estimation of a non-diagonal matrix allows a more complicated form of the parallel curves $t \mapsto \Psi_w(e_1 t + e_i)$ and different shapes of these parallel curves for the different bio-markers as visible for memory and language on Figure 7.

4.2.3 Disease model progression

We use Magnetic Resonance Images from the ADNI database. We select the images from subjects which ultimately develop the Alzheimer's disease and are at least observed twice. A common 128×128 slice is extracted from all rigidly aligned images, chosen so as to contain the Hippocampi and the Amygdala, which are particularly affected during Alzheimer's progression. We perform a 10-fold cross-validation on the subjects. After estimation, the test set of subjects is fitted onto the model by gradient descent on the individual parameters u_i.

A mean trajectory and the patterns of progressions $\Psi_w(v_0 t + e_i)_{i=2,\dots,d}$ are given on Figure 8. Those different patterns correspond to parallel trajectories to the reference geodesic on the learned submanifold as shown on Figure 1. The mean trajectory recovers grey matter loss during the disease progression, and especially grey matter loss near the Hippocampi and ventricles. The different patterns of progression on Figure 8 illustrate slight differences in the global pattern, which correct for the different individual anatomies as well as for different progression patterns.

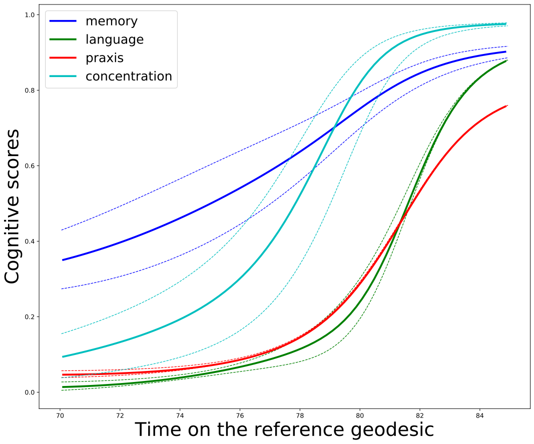

Figure 7: Learned main progression of the cognitive scores. Dashed curves are $t \mapsto \Psi_w(e_0 t \pm e_1)$: the first estimated parallel direction.

Reconstructing observations. To illustrate the reconstruction of the images by the generative model, we provide on Figure 9 the original images for a subject as well as its reconstructions. The reconstructed images are smoother and blurrier than the original images, since the model discarded what it sees as noise in the manifold estimation and since the cost is an ℓ^2 distance. To check whether the model generalizes to unseen observations, we fit, for each fold, the test set of observations to the model by gradient descent on the latent variables u_i. The MSE on the train set is of 0.087 ± 0.0008 while it is of 0.088 ± 0.003 on the test set, which shows that the model did not overfit.

Changing the latent space dimension. To analyze the effect of the dimension d of the latent space, we repeat the estimation of the model with d varying between 2 and 22. Figure 10 shows the results. The model MSE decreases steadily until $d \sim 18$, where it stagnates. This may indicate that the MRI data lies on a 18-dimensional submanifold of the whole space of observations.

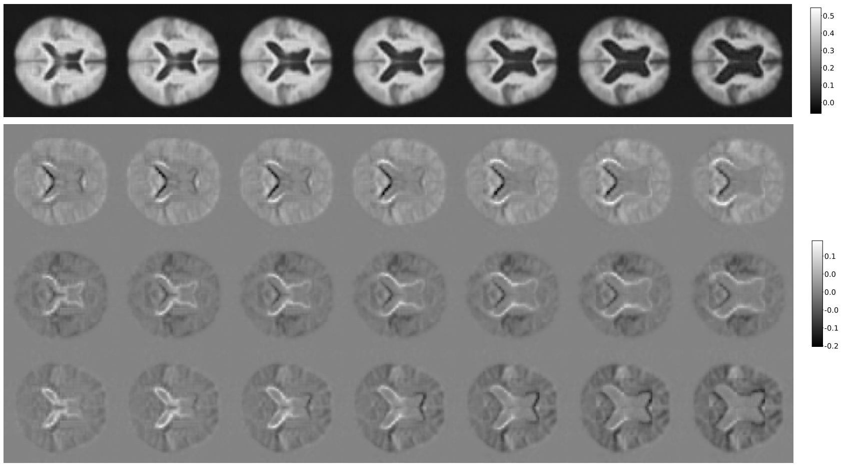

Figure 8: Top row is the reference geodesic $t \mapsto \Psi_w(e_0 t)$. Following rows are $t \mapsto \Psi_w(e_0 t + e_i) - \Psi_w(e_0 t)$ for $i \in \{2, \ldots, d\}$. Grey matter loss is visible in the main progression, with some variations in the parallel curves.

Analyzing the model output. We compare the distributions of the latent variables between sub-populations having different ages of diagnosis, alleles of the APOE gene -which indicates strong predisposition to Alzheimer's disease- or gender. For all the folds, we found a negative correlation between the pace of progression of the disease and the age of diagnosis, as observed in the literature [3], a pace of progression higher when at least one allele of the APOE gene is present, as described in [22], and a higher pace of progression for female subjects than for male subjects [7]. This shows that the Riemannian modelling proposed in the paper does yield informative results on real data.

5 Discussion

We show how to generalize two linear mixed-effect models to learn a Riemannian manifold of the space of observations and a Riemannian metric on this manifold which optimize the likelihood of the model. We show the importance of learning a Riemannian metric that is adapted to the model on top of learning a submanifold of the space of observations. This generalization procedure may be applied to a variety of generative linear models. The generalization of the mixed-effect model for longitudinal data is promising in the perspective of unsupervised disease progression model, since it identifies a geometry of progression for the images and a metric which may be of relevance for tasks such as early diagnosis. Further analysis of the obtained geometry such as classification with the learned metric are natural continuations of this work. Because of the huge parametric family of considered manifolds and metrics parametrized by the neural network, the identifiability of the parameters of the proposed models remains an open question, shared by all current deep learning methods.

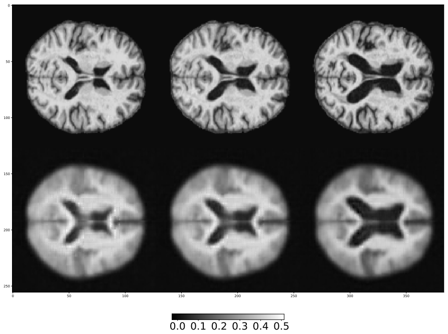

Figure 9: Top row: original trajectory. Bottom row: reconstruction by the model.

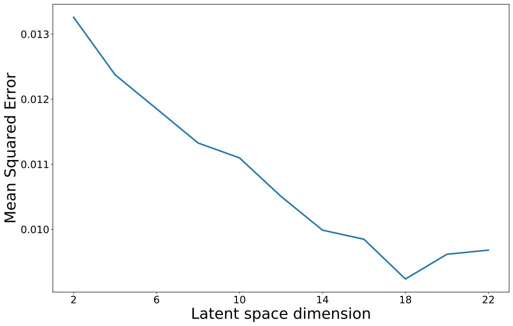

Figure 10: MSE with varying latent space dimension.

References

[1] G. Arvanitidis, L. K. Hansen, and S. Hauberg. A locally adaptive normal distribution. In *Advances in Neural Information Processing Systems*, pages 4251–4259, 2016.

[2] G. Arvanitidis, L. K. Hansen, and S. Hauberg. Latent space oddity: on the curvature of deep generative models. *arXiv preprint arXiv:1710.11379*, 2017.

[3] E. Bigio, L. Hynan, E. Sontag, S. Satumtira, and C. White. Synapse loss is greater in presenile than senile onset alzheimer disease: implications for the cognitive reserve hypothesis. *Neuropathology and applied neurobiology*, 28(3):218–227, 2002.

[4] A. Bône, O. Colliot, and S. Durrleman. Learning distributions of shape trajectories from longitudinal datasets: a hierarchical model on a manifold of diffeomorphisms. *arXiv preprint arXiv:1803.10119*, 2018.

[5] N. Chen, A. Klushyn, R. Kurle, X. Jiang, J. Bayer, and P. van der Smagt. Metrics for deep generative models. *arXiv preprint arXiv:1711.01204*, 2017.

[6] B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the em algorithm. *Annals of statistics*, pages 94–128, 1999.

[7] L. A. Farrer, L. A. Cupples, J. L. Haines, B. Hyman, W. A. Kukull, R. Mayeux, R. H. Myers, M. A. Pericak-Vance, N. Risch, and C. M. Van Duijn. Effects of age, sex, and ethnicity on the association between apolipoprotein e genotype and alzheimer disease: a meta-analysis. *Jama*, 278(16):1349–1356, 1997.

[8] P. T. Fletcher, C. Lu, S. M. Pizer, and S. Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE transactions on medical imaging*, 23(8):995–1005, 2004.

[9] T. Fletcher. Geodesic regression on riemannian manifolds. In *Proceedings of the Third International Workshop on Mathematical Foundations of Computational Anatomy-Geometrical and Statistical Methods for Modelling Biological Shape Variability*, pages 75–86, 2011.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[11] M. Hauser and A. Ray. Principles of riemannian geometry in neural networks. In *Advances in Neural Information Processing Systems*, pages 2804–2813, 2017.

[12] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[13] H. Karcher. Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics*, 30(5):509–541, 1977.

[14] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[16] I. Koval, J.-B. Schiratti, A. Routier, M. Bacci, O. Colliot, S. Allassonnière, S. Durrleman, A. D. N. Initiative, et al. Statistical learning of spatiotemporal patterns from longitudinal manifold-valued networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 451–459. Springer, 2017.

[17] J. B. Kruskal and M. Wish. *Multidimensional scaling*, volume 11. Sage, 1978.

[18] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[19] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[20] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

[21] S. J. Prince and J. H. Elder. Probabilistic linear discriminant analysis for inferences about identity. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[22] S. Sadigh-Eteghad, M. Talebi, and M. Farhoudi. Association of apolipoprotein e epsilon 4 allele with sporadic late onset alzheimer's disease. *A meta-analysis. Neurosciences (Riyadh)*, 17(4):321–326, 2012.

[23] J.-B. Schiratti, S. Allassonniere, O. Colliot, and S. Durrleman. Learning spatiotemporal trajectories from manifold-valued longitudinal data. In *Advances in Neural Information Processing Systems*, pages 2404–2412, 2015.

[24] H. Shao, A. Kumar, and P. T. Fletcher. The riemannian geometry of deep generative models. *arXiv preprint arXiv:1711.08014*, 2017.

[25] A. Srivastava, S. H. Joshi, W. Mio, and X. Liu. Statistical shape analysis: Clustering, learning, and testing. *IEEE Transactions on pattern analysis and machine intelligence*, 27(4):590–602, 2005.

[26] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[27] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[28] M. Zhang and T. Fletcher. Probabilistic principal geodesic analysis. In *Advances in Neural Information Processing Systems*, pages 1178–1186, 2013.