Geodesic Discriminant Analysis for manifold-valued data

Maxime Louis Benjamin Charlier Stanley Durrleman
Institut du Cerveau et de la Moelle épinière, Inserm, CNRS, Sorbonne Université, Paris, France
Inria, Aramis project-team, Paris, France
{maxime.louis, benjamin.charlier, stanley.durrleman}@icm-institute.org

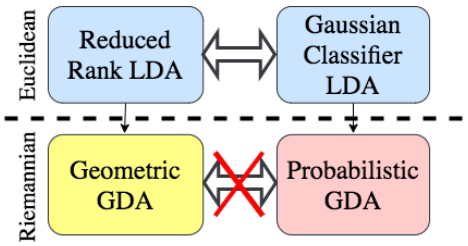
Abstract

In many statistical settings, it is assumed that high-dimensional data actually lies on a low-dimensional manifold. In this perspective, there is a need to generalize statistical methods to nonlinear spaces. To that end, we propose generalizations of the Linear Discriminant Analysis (LDA) to manifolds. First, we generalize the reduced rank LDA solution by constructing a geodesic subspace which optimizes a criterion equivalent to Fisher's discriminant in the linear case. Second, we generalize the LDA formulated as a restricted Gaussian classifier. The generalizations of those two methods, which are equivalent in the linear case, are in general different in the manifold case. We illustrate the first generalization on the sphere $\mathbb{S}^2$. Then, we propose applications using the Large Deformation Diffeomorphic Metric Mapping (LDDMM) framework, in which we rephrase the second generalization. We perform dimension reduction and classification on the kimia-216 dataset and on a set of 3D brain structures segmented from Alzheimer's disease and control subjects, recovering state-of-the-art performances.

1. Introduction

Large quantities of high-dimensional structured data are now routinely acquired such as various types of images, videos or 2D and 3D shape data. The raw description of this kind of data does not in general reflect its intrinsic structure: it hides the generally low number of degrees of freedom that produced the observations, it is often very high-dimensional and the use of usual distances on raw data is not appropriate. To obtain a better description of the data, a standard approach is to construct or learn a low-dimensional manifold which best approximates a set of observations under a predefined criterion. If an invariance property is expected in the data, such as an invariance by rotation and scaling, it is possible to project the set of observations in the corresponding quotient space [14] or at least to build a quotiented description of the data to more classic machine learning methods as it is commonly done in scattering [19] and convolutional networks. Otherwise, the manifold structure can be learned from the data itself as it is proposed in manifold learning approaches, which project the data onto $\mathbb{R}^n$ for some small $n \in \mathbb{N}$ while trying to preserve some local or global structure observed in the high-dimensional data (see [17, 24]).

All of these approaches produce a large amount of manifold-valued data for which it is necessary to adapt usual linear machine learning approaches. The Linear Discriminant Analysis (LDA) method is a popular classification algorithm assuming a linear structure in the data. It can be formulated in two different ways. First, as a dimension reduction problem which seeks to maximize the between-class variance with respect to the within-class variance. Second, LDA can be formulated as a classification problem supposing each class is distributed as a Gaussian random variable with common covariance matrix. In this paper, we propose generalizations of those two formulations of LDA to manifold-valued data. So far, most classifications of manifold-valued data was done after having projected the data onto a common tangent space (see [6, 16]), or using a coordinate chart on the manifold. Both of these approaches only see a simplified version of the geometry of the manifold. The proposed generalizations of LDA address this by taking into account the intrinsic geometry of the data to perform dimension reduction and classification.



The first generalization, derived in Section 2, that we call geometric Geodesic Discriminant Analysis (geometric GDA), is obtained by rewriting the Fisher Discriminant Ratio (FDR) –which measures the ratio between the between-class variance and the within-class variance– using geodesic distances on the manifold. We propose to build a geodesic subspace on the manifold on which this criterion is maximized.

Because the optimization of the criterion formulated for the geometric GDA is not always tractable, we proceed with a second generalization of LDA. Derived in Section 3, it extends the restricted Gaussian Classifier formulation of LDA. To extend this formulation to manifolds, we model the classes distributions as Riemannian exponentials of Gaussian distributions defined on a tangent space at a specific common point. We then propose to optimize the point on which this construction is centered at and to use convenient descriptions of the between-class covariance and the within-class covariance, basing our work on [21]. We call this method probabilistic GDA. We will show how to make this approach computationally efficient for a wide variety of manifolds.

It has been shown in [10] that, in the linear case, this approach is equivalent to the reduced rank LDA: it produces the same dimension reduction and classification rule. In the nonlinear case, probabilistic GDA and geometric GDA will not be equivalent.

A particular case of manifold structure can be obtained under the action of a group of diffeomorphisms on a set of shapes. Our formulation of the Large Deformation Diffeomorphic Metric Mapping (LDDMM) provides a way to parametrize a finite-dimensional manifold of diffeomorphisms. This family of diffeomorphisms then allows the comparison of shapes on which they act. We introduce this framework in Section 4. In Section 5 we provide results of the algorithm on 2D shapes extracted from the kimia-216 dataset as well as on 3D Brain structures segmented from the ADNI dataset. The probabilistic GDA is however generic and efficient enough to be applicable to a much broader family of manifolds.

Among related work, Exact Principal Geodesic Analysis (Exact PGA), initially formulated in [8], proposes to construct a geodesic subspace on which the explained variance, as measured using geodesic distances is maximized. Several variations have been proposed, among which Bayesian PGA [26] or Horizontal Component Analysis [22]. Our work differs from these methods since we propose a supervised learning algorithm which optimizes class separation, not explained variance, to increase classification performances.

We summarize our contributions:

1. We propose geometric GDA, a generalization of the reduced rank definition of LDA to manifold-valued data.
2. We propose probabilistic GDA, a generalization of the restricted Gaussian classifier definition of LDA to manifold-valued data.
3. We illustrate the geometric GDA method on $\mathbb{S}^2$ with synthetic data and the probabilistic GDA model on the kimia-216 dataset and on a dataset of hippocampi extracted from magnetic resonance images (MRI).

2. Geometric Geodesic Discriminant Analysis

In this section, we introduce geometric GDA, a generalization of LDA to manifold-valued data using Fisher approach to LDA [7]. In this paper, we consider a set of labelled observations $(y_i)_{i=1,...,N} \in \mathcal{M}$ from $C > 0$ different classes, where $\mathcal{M}$ is a smooth Riemannian manifold that we assume geodesically complete. For $p, q \in \mathcal{M}$, we note $d(p, q)$ the geodesic distance between p and q, and $s \in \mathbb{N}$ the dimension of the manifold.

If the manifold is a vector space, reduced rank LDA [7] consists in projecting the observations onto a linear subspace on which the between-class variance is maximized with respect to the within-class variance. Fisher proposed to find unit vectors a via maximization of the Fisher Discriminant Ratio (FDR):

$$0 < \frac{a^\top B a}{a^\top W a} \tag{1}$$

where $^\top$ denotes transposition, B is the between-class covariance matrix –the covariance matrix of the class centroids– and W is the within-class covariance matrix. To provide an expression generalizable to manifolds, we rewrite this FDR:

$$\frac{\frac{1}{C-1}\sum_{c=1}^{C}(a^\top \mu - a^\top \mu_c)^2}{\frac{1}{N-C}\sum_{c=1}^{C}\sum_{i\in I_c}(a^\top \mu_c - a^\top x_i)^2} \tag{2}$$

where μ is the mean of the observations, for each $c \in \{1, ..., C\}$, μ_c is the empirical mean of the class c and I_c is the set of indices of the observations of the class c. For any observation x, $a^\top x$ may be interpreted as the projection of the observation onto the space spanned by a.

If the manifold is non flat, instead of constructing a linear subspace, we will build a geodesic subspace

on the manifold, as proposed for PGA in [8]. For any $m \in \mathcal{M}$, for any subspace $V \subset T_m\mathcal{M}$, we define the geodesic subspace $\mathrm{Exp}_m V = \{\mathrm{Exp}_m(v) | v \in V\}$ where $\mathrm{Exp}_m : T_m\mathcal{M} \to \mathcal{M}$ is the Riemannian exponential at m. We define a projection operator on S by $\pi_S(x) = \mathrm{argmin}_{y \in S} \mathrm{d}(x, y)^2$. This projection, defined by minimization, might be ill-defined unless we restrict ourselves to a neighborhood of m. Assuming it is well-defined, equation (2) can now be generalized to manifolds by using geodesic distances measured after projection on S:

$$\frac{\frac{1}{C-1}\sum_{c=1}^{C} \mathrm{d}(\pi_{\mathrm{Exp}_m(V)}(\mu), \pi_{\mathrm{Exp}_m(V)}(\mu_c))^2}{\frac{1}{N-C}\sum_{c=1}^{C}\sum_{i \in I_c} \mathrm{d}(\pi_{\mathrm{Exp}_m(V)}(\mu_c), \pi_{\mathrm{Exp}_m(V)}(x_i))^2}. \quad (3)$$

where μ (resp. μ_c) are the Fréchet means [13] of the observations (resp. of the observations of class c). Reduced rank LDA on the manifold becomes the problem of maximizing this with respect to $m \in \mathcal{M}$ and V linear subspace of $T_m\mathcal{M}$. V can be constructed in a forward fashion by a basis $\{v_1, ..., v_k\}$ where k is a chosen number of component. Note that an alternative generalization could propose to recompute the Fréchet means after the projection, which yields a criterion different than equation (3) since the Fréchet mean of the projection is in general not the projection of the Fréchet mean. This alternative generalization of equation (1) would be more expensive to compute, since the computation of the different Fréchet means of the projections would be required at every step of the optimization procedure.

2.1. Inference

In practice, it is hard to find a robust procedure which optimizes both m and V at the same time. We propose a greedy procedure: we first optimize jointly m and a first geodesic component $v_1 \in T_m\mathcal{M}$, and then add new components v_k one at a time. In the linear case, this procedure yields the exact same optimum. A theoretical discussion about the validity of this procedure in the nonlinear case will be part of further work.

Note that if closed-form expressions are available for Riemannian logarithms and exponentials, the proposed geometric GDA can be computed efficiently. This is the case for Kendall shape space, the sphere or the manifold of symmetric positive-definite matrices with affine-invariant metric for instance. We will provide results in the case of the sphere $\mathbb{S}^2$ in Section 5.1.

Note also that the optimization problem (3) might be ill-defined if some degeneration is observed in the data, for instance if all the data points lie on a single geodesic, as in the linear case when the between-class covariance matrix does not have full rank. The study of the conditions for this estimation procedure to be well-defined will not be conducted in this paper.

2.2. Dimension reduction and classification

After estimation of m and V, one can project the observations onto V by taking the coordinates of $\pi_{\mathrm{Exp}_m(V)}(y)$ for each observation y. This gives a low-dimensional representation of the data, on a space in which classes difference predominate. This representation has the same range of applications as dimension reduction with linear LDA.

Classification can then be done in one of two ways. First, directly in the low-dimensional space $\mathrm{Exp}_m(V)$ by comparison of test observations geodesic distances to the different classes centroids on $\mathrm{Exp}_m(V)$, in a fashion very similar to the classic LDA. Second, it can be done after projection of the data onto $V \subset T_m\mathcal{M}$ using any usual classifier, whose performances will in general be improved if the FDR (3) has been correctly optimized.

Unfortunately, when no closed-form expressions are available for Riemannian logarithms or exponentials, geometric GDA is intractable. To remedy this, we propose a generalization of the alternative formulation of LDA.

3. Probabilistic Geodesic Discriminant Analysis

In the linear case, the restricted Gaussian classifier formulation of LDA assumes each class is distributed along a normal distribution, with common covariance Σ. In this linear setting, the probability of an observation y, if it is of class c, is:

$$y | c = c \sim \mathcal{N}(y | \mu_c, \Sigma). \quad (4)$$

In [10], the authors show that maximimizing the likelihood of this model with a rank constraint on the means μ_c ($\mathrm{rank}(\mu_c)_{c=1,...,C} < K$) is equivalent to projecting the observations onto the K first discriminant components found by maximization of the Fisher Discriminant Ratio (1), even when the within-class covariance matrix Σ is unknown.

There is no natural way to generalize equation (4) to manifold-valued data. In particular, it is hard to make sense of the homoscedasticity hypothesis in LDA since it involves comparing covariance matrices defined at different tangent spaces on the manifold. One possible generalization of equation (4) is to consider:

$$y | c = c \sim \mathrm{Exp}_m(d_c + \alpha) \quad (5)$$

where m is a point on the manifold and $\alpha \sim \mathcal{N}(0, \Sigma)$ is a normal distribution on the tangent space $T_m\mathcal{M}$. If the manifold is flat, this model is equivalent to (4), and the rank constraint can in theory be enforced on the vectors $d_c \in T_m\mathcal{M}$. The homoscedasticity is replaced with the assumption that, as seen from the tangent space to m, the logarithms of the observations of the different classes are distributed along normal distributions with the same covariance matrix Σ. This approach is still hardly tractable in practice. First, learning the model will require estimating the full within-class covariance matrix Σ. Second, the rank constraint is difficult to implement in practice. We therefore extend the model defined in [21], which is similar to LDA, to:

$$y_i|c \sim \mathcal{N}(\mathrm{Exp}_m(F\alpha_c + G\beta_i), \sigma) \tag{6}$$

where:

- $\mathcal{N}$ is a normal distribution on $\mathcal{M}$ with density $p(y, \mu, \sigma) = \frac{1}{D(\mu,\sigma)} e^{-\frac{1}{2\sigma} d^2(y,\mu)}$, as defined in [8]. It can also be taken to be a normal distribution on a larger space of observations, to ease computations, as used in the applications below,
- F is a s times $C-1$ matrix which can be seen as the between-class covariance matrix,
- G is a s times N_G matrix where $N_G \in \mathbb{N}$ is the selected number of intra-class components to estimate: it can be seen as the principal components of the within-class variations, as seen from $T_m\mathcal{M}$,
- For each class c, α_c in $\mathbb{R}^{C-1}$ contains the coordinates of the class c in the $C-1$-dimensional space represented in F,
- β_i in $\mathbb{R}^{N_G}$ is a hidden variable which contains the coordinates of the i-th observation within its class, in the N_G-dimensional space represented in G.

We put normal priors on α and β, and an automatic relevance determination prior on G as in [18]:

$$P(G; \gamma) = \prod_{i=1}^{N_G} \left(\frac{\gamma_i}{2\pi}\right)^{\frac{s}{2}} \exp\left(-\frac{\gamma_i}{2}\|G_i\|_2^2\right) \tag{7}$$

where $(\gamma_i)_{i=1,\dots,N_G}$ is a set of parameters which are estimated during the learning procedure and $(G_i)_{i=1,\dots,N_G}$ are the columns of G. This prior allows the automatic selection of a relevant number of dimensions in the within-class covariance structure.

Compared to the tangent LDA, which consists in performing an LDA after having projected the observations onto the tangent space to the Fréchet mean, the proposed method updates the within and between-class components with a constant feedback from the real geometry of the data. Besides, we allow the joint optimization of the point m and do not constrain it to be the Fréchet mean of the data, which may not be optimal in the perspective of class separation (in [11] the authors show it is not optimal in the case of exact PGA).

3.1. Inference

As in [26], the mode of the posterior distribution of the variables α and the optimal values of the parameters can be obtained as a maximum a posteriori using a gradient descent. In more details, we maximize $P(y_j|\theta, \beta)P(\theta)P(\beta)$ with respect to the parameters $\theta = (F, G, \alpha, m, \sigma, \gamma)$ and β. The computation of the gradient requires the differentiation of a function of a geodesic endpoint with respect to its initial conditions. It can be done by backward integration using the method described in [23].

This approach is tractable in a wide variety of situations:

- Even if there is no closed-form expression for Riemannian exponential, geodesics can still be computed through integration of the Hamiltonian system of equations, using only the inverse of the metric and its gradient, as shown in [5]. In that case, automatic differentiation is a competitive way to compute the gradients, as shown in [15].
- The normal distribution in equation (6) can be replaced with a normal distribution on a larger space which contains the observations e.g.a pixel-wise normal distribution for images, or a noise in $\mathbb{R}^3$ for $\mathbb{S}^2$. This saves the computation of the normalization constant of the Riemannian normal distribution and of geodesic distances. Since this distribution is used only to measure residuals, we believe it has a limited effect on the model.
- The estimation procedure can be parallelized among the different subjects, rendering it efficient even with large data sets.

3.2. Dimension reduction and classification

After estimation of the parameters of the model, it is possible to project an observation y onto the geodesic subspace defined by F by optimization of:

$$\delta \to d(\mathrm{Exp}_m(F\delta), y)^2 \tag{8}$$

with respect to $\delta \in \mathbb{R}^{N_c-1}$, which indicates the position of the observation y in the geodesic subspace $\mathrm{Exp}_m(F)$. Doing so yields a low-dimensional description of each

data point. Classification can be performed after dimension reduction of the dataset. We will show results of this classification procedure in Section 5.

Classification can also be done using the probabilistic GDA model, by maximizing the likelihood of an observation with respect to the classes. For each unobserved y, using Bayes rule:

$$p(c = c_k|y) = \int_\beta p(y|c = c_k, \beta_k = \beta)p(\beta)p(c = c_k). \tag{9}$$

The integral over the hidden variable β corresponds to looking at the observation J through all its possible representations as an object of class c_k, where the representations have been learned through the matrix G. This integral is expensive to compute or approximate in most cases and we decide to settle for the mode:

$$p(c = c_k|y) \propto p(y|c = c_k, \beta^*)p(c = c_k) \tag{10}$$

where $\beta^* = \text{argmax}_\beta p(y|c = c_k, \beta)$, which can be estimated via gradient descent.

The ability to compute the integral (9) would allow to evaluate the new observation as an element in the space quotiented by the different representations of the elements of the class c_k. Additionally, as described in [21], it would also allow to do one-shot learning i.e. being able to decide if a new observation is in the set of known classes or if it is more likely to belong to a yet unobserved class.

4. Probabilistic GDA for shape analysis.

The Probabilistic GDA introduced above can be applied in a variety of situations, we will focus on examples of applications in the case of shapes modelled using the LDDMM framework [20,25]. We first introduce this framework, before rewriting the model (6) in this particular case.

4.1. Embedding shapes and images on a manifold

The LDDMM framework provides a way to compare shapes via the action of diffeomorphisms of the ambient space. Such diffeomorphisms are obtained by integration of the flow of a square integrable time-varying vector field. The parametrization of the diffeomorphisms then amounts to the parametrization of time-varying vector fields. In our approach, we use as in [6] a sparse description of vector fields:

$$X(x) = \sum_{i=1}^{p} k(x, q_i)p_i \tag{11}$$

where $p \in \mathbb{N}$ is fixed, $(q_i)_{i=1,\dots,p}$ is a set of control points, $(p_i)_{i=1,\dots,p}$ is a set of momenta and k is a Gaussian kernel of fixed width ρ. The space of such vector fields is a Reproducible Kernel Hilbert Space (RKHS) K with

$$\langle X, X' \rangle_K = \sum_{i,j=1}^{p} k(q_i, q'_j)p_i^\top p'_j. \tag{12}$$

Given an initial vector field X of this form, one can show [6] that there is a unique time-varying vector field $X(t, \cdot)$ such that $X(0, \cdot) = X$ which minimizes $\int_0^1 \|X(t, \cdot)\|_K^2$. We call this the geodesic flow of the initial vector field. Considering only such geodesics, we get a parametrization of diffeomorphisms solely determined by the initial set of control points and momenta. Following this construction, the obtained set of diffeomorphisms form a finite-dimensional Riemannian manifold. On this manifold, the exponential map corresponds to computing the geodesic flow until time 1 of the time-varying vector field with a given set of initial momenta. We denote $\Phi_{q,p} \cdot M$ the action of the diffeomorphism $\Phi_{q,p}$ parametrized by the initial control points and momenta q, p on the shape M. If M is a mesh embedded in $\mathbb{R}^n$, then $\Phi_{q,p}$ acts on the points of the meshes directly. If M is an image, $\Phi_{q,p} \cdot M = M \circ \Phi_{q,p}^{-1}$ where M is seen as an element of $L_2(\mathbb{R}^n; \mathbb{R})$ for some integer n.

4.2. A generative model

Let us assume that we have a collection of shapes $(y_k)_{i=1,\dots,N}$ where $N \in \mathbb{N}$. We note n the dimension of the ambient space and p the number of control points, so that the considered manifold of diffeomorphisms is of dimension $p \times n$. As described in equation (6), we assume that each shape y_k was generated with probability:

$$\frac{1}{(2\pi)^{\frac{\Lambda}{2}}\sigma^\Lambda} \exp\Big(-\frac{1}{2\sigma^2}\|\Phi_{q,F\alpha^{c_k}+G\beta_k} \cdot M - y_k\|_\Lambda^2\Big) \tag{13}$$

where:

- $\Lambda \in \mathbb{N}$ is the dimension of the observations *e.g.* number of voxels for the images, number of faces for varifolds. We embed those observed shapes in a Λ-dimensional space on which we define a norm $\|_\Lambda$ (L^2 for images, varifold norm for meshes as in [9]),
- M is a template shape,
- $\Phi_{q,F\alpha^{c_k}+G\beta_k}$ is the diffeomorphism obtained with the initial momenta $p_k = F\alpha^{c_k} + G\beta_k$ and control points q.

Note that we replaced the normal distribution on the manifold by a normal distribution on the set of shapes, that can be defined for images, varifolds or currents as

shown in [9]. There are two reasons for this. First, the orbit of M under the action of the group of diffeomorphisms does not in general contain the observations, the idea being to describe shape variability with strong smoothing constraints on the shape structures. Second, even if we could use geodesic distances on the manifold of diffeomorphisms, the computation of this geodesic distance would be too expensive in general to make the inference of the model tractable.

For the inference, we estimate the mode of the logarithm of the posterior distribution, which writes, using Bayes rules and assuming that F, G, α and β are independent:

$$\begin{aligned} l(\theta) = \log(P(F,G,\alpha,\beta|y_k;\gamma,I,\sigma)) = -\Lambda\log(\sigma) \\ - \frac{1}{2\sigma^2}\|\Phi_{q,F\alpha^{c_k}+G\beta_k}\cdot M - y_k\|^2_\Lambda - \frac{1}{2}\|\beta\|^2_2 \\ \sum_{i=1}^{N_G}\frac{\Lambda pn}{2}\log(\frac{\gamma_i}{2\pi}) - \sum_{i=1}^{N_G}\frac{\gamma_i}{2}\|G_i\|^2_2 - \frac{1}{2}\|\alpha\|^2_2 \end{aligned} \tag{14}$$

Derivating (14) yields the closed-form updates for σ and γ:

$$\gamma_i = \frac{\Lambda pn}{\|G_i\|^2_2}. \tag{15}$$

$$\sigma^2 = \frac{\sum_{k=1}^{N}\|(\Phi_{F\alpha^{c_k}+G\beta_k})\cdot M - y_k\|^2_\Lambda}{\Lambda N} \tag{16}$$

The computation of the gradients with respect to the momenta p, the control points q and the template M can be done by backward integration of system of adjoint equations as detailed in [6,23,26] and propagated to α, β, F and G using the chain rule. The optimized functional is once again not convex in general. Algorithm 1 gives a pseudo-code for the estimation procedure. A complete code of the model is made available [1].

Algorithm 1 Probabilistic GDA inference on shapes

$F, G, \alpha, \beta, M, q \leftarrow$ Initialization from Tangent LDA

$\gamma, \sigma \leftarrow$ (15)(16): for initialization.

while no convergence **do**

Compute $l(\theta)$

Compute $\nabla_M l(\theta)$, $\nabla_p l(\theta)$, $\nabla_q l(\theta)$.

Propagate to $\nabla_F l(\theta)$, $\nabla_G l(\theta)$, $\nabla_\alpha l(\theta)$, $\nabla_\beta l(\theta)$

Update $(F, G, \alpha, \beta, M, q)$ by line search.

$\gamma, \sigma \leftarrow$ (15)(16): closed-form update.

return $F, G, \alpha, \beta, M, q, \sigma, \gamma$

[1] A code for the model is available at www.deformetrica.org

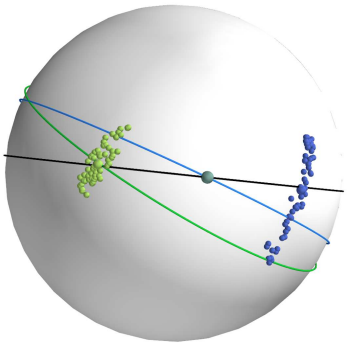


Figure 1: The points are labelled data. The black geodesic is the first component of the PGA method. The blue geodesic is obtained by geometric GDA with m set to the Fréchet mean. The green geodesic is obtained by geometric GDA with optimization of m.

5. Applications and Results

5.1. Geometric GDA on $\mathbb{S}^2$

We performed the optimization of the criterion given in equation (3) in the case of the sphere $\mathbb{S}^2$ with the metric induced from $\mathbb{R}^3$, on a synthetic set of points of two classes. Note that, whether we optimize the position of the point m on which the geodesic subspace is built or not, the optimization problem is in general not convex. We therefore perform multiple gradient descents with randomly chosen initial conditions, and select the final estimated values which give the optimum of the Fisher Discriminant Ratio. We compare three methods: an LDA performed in the tangent space to the Fréchet mean, a geometric GDA performed with a geodesic subspace set to the Fréchet mean (GDA) and a geometric GDA performed with the joint estimation of the geodesic subspace and of the point on which it is built (full GDA).

Figure 1 shows the estimated geodesics in the different cases of GDA, as well as the result of an exact PGA built by optimization of the explained variance on a geodesic subspace at the Fréchet mean. In each case, we measure the FDR after projection onto the first component found after optimization. Note that the FDR measured after projection assuming a linear structure differs from the FDR defined in equation (3). Indeed, the projection of the classes centroids is in general different from the class centroids of the projections, unlike in the linear case. We provide the values of the FDRs measured after projection and the FDRs measured in equation (3) in Table 1.

The geometric GDA outperforms an LDA performed in the tangent space to the Fréchet mean in terms of class separation, indicating that we may obtain better

Method	Tangent LDA	GDA	full GDA
FDR of projection	495	514	647
FDR equation (3)	x	505	636

Table 1: Fisher Discriminant Ratios. Higher FDRs indicate a better class separation.

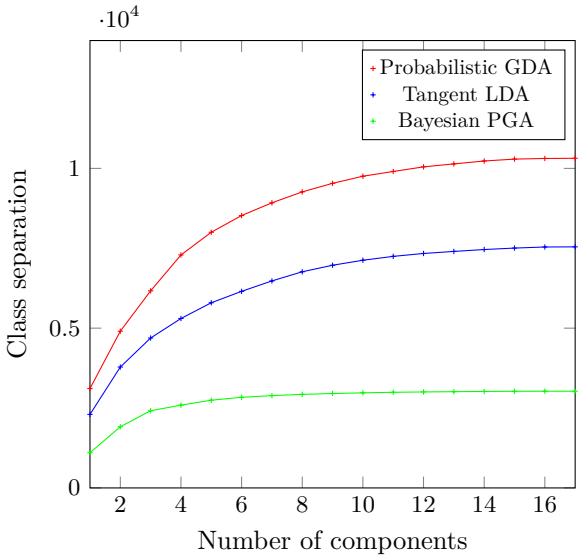


Figure 2: Class separation for each number of selected components, on the kimia-216 dataset.

classification results in some situations. In addition, as mentioned in Section 2, the optimization of m in equation (3) allows a significant improvement.

5.2. Kimia-216

We used the setting described in Section 4 on shapes from the kimia-216 dataset. The kimia-216 dataset consists of 18 classes each containing 12 observations. We extracted the contour of the shapes on the images and modelled them as varifolds with a Gaussian kernel of width set at 13 (expressed in pixels of the original images). The number of points of the obtained shapes is not controlled and vary between 300 and 800. For each class, we randomly selected 9 observations that we rigidly aligned one to another. We proceeded to the estimation of the model described in equation (6), simultaneously estimating the matrices F and G, the vectors α for each class, the mode of β for each observation, as well as the template shape I and the set of control points. We set the kernel width ρ of the diffeomorphisms to 10 (in terms of pixels of the original images).

After estimation of the parameters of the model, we projected each training observation using the method described in Section 3.2. To measure the quality of the projection, we compute, in the low-dimensional space, the between-class covariance matrix B and the within class-covariance matrix W and compute the eigenvalues of $W^{-1}B$. Those eigenvalues measure the separation of the classes, and are equivalent to the FDR in the linear case. We compute those eigenvalues for the probabilistic GDA, the tangent LDA and the bayesian PGA with 17 components. Note that the bayesian PGA is a special case of the model 6 when there is a single class. The results are provided in Figure 2. The probabilistic GDA outperforms both the tangent LDA and the Bayesian PGA in the separation it provides after projection of the data.

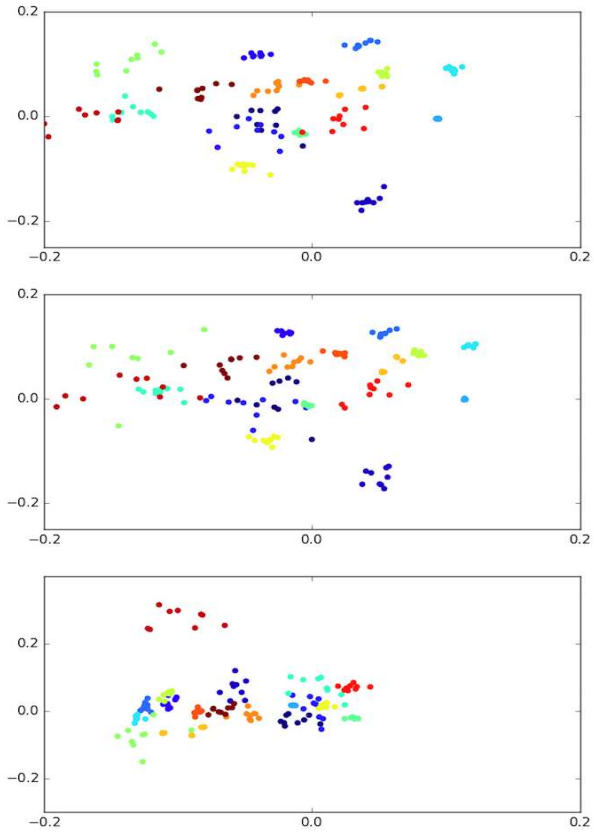


Figure 3: First two components of probabilistic GDA (top), tangent LDA (middle), Bayesian PGA (bottom), (arbitrary units).

Then, we provide a plot of the two first components of each observations found using the probabilistic GDA, for visualization purposes, to be compared with the same components for tangent LDA and bayesian PGA, on Figure 3.

Finally, we investigated classification performances on the kimia-216 database, using the classification procedure described in Section 3.2. In details, for each test

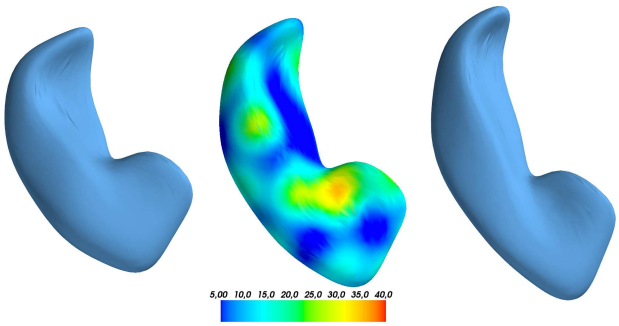


Figure 4: Three observations on the first geodesic discriminant component estimated from the ADNI dataset. From left to right, we follow the geodesic from normal controls to MCIc subjects. The middle observation is the estimated M template, colored with the initial velocity field norm.

observation and each candidate class k, we evaluated the mode of the integral (10) by rigidly aligning the test observation to an element of the class k and performing a gradient descent on β with α set to α_k, the position of the class k in the space spanned by F. We take the class which gives the smallest residual after the descent. A 3-fold result of this classification procedure gives an average accuracy of 89%. Note that such low classification results compared to usual benchmarks [16] can be expected since this dataset is not well adapted to deformable models: the differences between the shapes occur both on small and large scales.

5.3. Brain structures in the course of Alzheimer's disease

From MRI images in the ADNI dataset, we segmented hippocampi from 125 normal controls and MCIc subjects (subjects who have or will convert to Alzheimer's disease) using Freesurfer [3]. We then ran the probabilistic GDA model four times on randomly extracted training set and test set in the data. Each run provided an estimation of a single discriminant geodesic component. Figure 4 shows an example of discriminant geodesic component.

After estimation on the train set, we projected both test and train set onto the first geodesic component by optimization of (8). We then trained a logistic regression classifier on the projected, 1-dimensional, data. This is common practice after LDA dimension reduction, to learn the appropriate threshold for the classification and to correct for the strict homoscedasticity hypothesis of the model.

The AUC and accuracy scores are available in Table 2 and compared to a classification based on the hippocampi volumes on the exact same folds, as well as to other reference methods performing cross-sectional hippocampus-based classification of normal controls versus MCIc subjects. Our method provides state-of-the-art accuracy and AUC results. Note that the problem of classifying MCIc versus normal controls is in general much better solved using whole T1 MRIs, which could be future work using the same proposed probabilistic GDA applied to full 3D images.

	AUC (std)	Accuracy (std)
Tangent LDA	0.77 (0.06)	0.76 (0.07)
Probabilistic GDA	0.78 (0.07)	0.77 (0.08)
Volumes	0.68 (0.003)	0.56 (0.003)
Chupin et al. [1]	x	0.71
Cuignet et al. [2]	x	0.73

Table 2: AUC and accuracy scores at MCIc vs normal controls classification using only the hippocampus.

Our method in this case could be compared with [12] in which the authors do an analysis of hippocampi differences between Alzheimer's and normal controls modelling the shapes using the elastic shape framework [12]. However, their analysis of the differences is done after having performed a PCA on the tangent space to the Fréchet mean, and their approach requires a parametrization of the surfaces.

6. Conclusion

We propose generalizations of the different formulations of LDA. The geometric GDA constructs a geodesic subspace which maximizes the FDR as seen in this curved space, but is hard to compute in general. The probabilistic GDA, generalization of the gaussian classifier formulation of LDA, is much more efficient to compute. We illustrated the methods with dimension reduction and classification tasks, with an example on a set of 3D shapes segmented from subjects with Alzheimer's disease where we reach state-of-the-art classification results.

Applications to data sets of different types would allow to best show the applicability of the method. Future work also includes improving the estimation procedure for the probabilistic model, to take full advantage of the hidden variable β, using for instance use a stochastic version of the EM algorithm [4]. In addition, several theoretical discussions could be conducted: to see when the π operation is well-defined, to study the consistency of the estimation and the identifiability of the model or to formulate criteria to identify and handle degenerate cases.

References

[1] M. Chupin, E. Gérardin, R. Cuingnet, C. Boutet, L. Lemieux, S. Lehéricy, H. Benali, L. Garnero, and O. Colliot. Fully automatic hippocampus segmentation and classification in alzheimer's disease and mild cognitive impairment applied on data from adni. *Hippocampus*, 19(6):579–587, 2009. 8

[2] R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, S. Lehéricy, M.-O. Habert, M. Chupin, H. Benali, O. Colliot, A. D. N. Initiative, et al. Automatic classification of patients with alzheimer's disease from structural mri: a comparison of ten methods using the adni database. *NeuroImage*, 56(2):766–781, 2011. 8

[3] A. M. Dale, B. Fischl, and M. I. Sereno. Cortical surface-based analysis: I. segmentation and surface reconstruction. *Neuroimage*, 9(2):179–194, 1999. 8

[4] B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the em algorithm. *Annals of statistics*, pages 94–128, 1999. 8

[5] M. P. Do Carmo. *Riemannian geometry*. Birkhauser, 1992. 4

[6] S. Durrleman, M. Prastawa, N. Charon, J. R. Korenberg, S. Joshi, G. Gerig, and A. Trouvé. Morphometry of anatomical shape complexes with dense deformations and sparse parameters. *NeuroImage*, 2014. 1, 5, 6

[7] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals Eugen.*, 7:179–188, 1936. 2

[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. 2, 3, 4

[9] P. Gori, O. Colliot, Y. Worbe, L. Marrakchi-Kacem, S. Lecomte, C. Poupon, A. Hartmann, N. Ayache, and S. Durrleman. *Bayesian Atlas Estimation for the Variability Analysis of Shape Complexes*, pages 267–274. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. 5, 6

[10] T. Hastie and R. Tibshirani. Discriminant analysis by gaussian mixtures. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 155–176, 1996. 2, 3

[11] T. Hotz, S. Huckemann, A. Munk, D. Gaffrey, and B. Sloboda. Shape spaces for prealigned star–haped objects-studying the growth of plants by principal components analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(1):127–143, 2010. 4

[12] S. H. Joshi, Q. Xie, S. Kurtek, A. Srivastava, and H. Laga. Surface shape morphometry for hippocampal modeling in alzheimer's disease. In *Digital Image Computing: Techniques and Applications (DICTA), 2016 International Conference on*, pages 1–8. IEEE, 2016. 8

[13] H. Karcher. Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics*, 30(5):509–541, 1977. 3

[14] D. G. Kendall. A survey of the statistical theory of shape. *Statist. Sci.*, 4(2):87–99, 05 1989. 1

[15] L. Kühnel and S. Sommer. Computational anatomy in theano. In *Graphs in Biomedical Image Analysis, Computational Anatomy and Imaging Genetics*, pages 164–176. Springer, 2017. 4

[16] S. Lee, N. Charon, B. Charlier, K. Popuri, E. Lebed, M. V. Sarunic, A. Trouvé, and M. F. Beg. Atlas-based shape analysis and classification of retinal optical coherence tomography images using the functional shape (fshape) framework. *Medical image analysis*, 35:570–581, 2017. 1, 8

[17] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008. 1

[18] D. J. C. MacKay. Probable networks and plausible predictions - - a review of practical bayesian methods for supervised neural networks, 1995. 4

[19] S. Mallat. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012. 1

[20] M. I. Miller, A. Trouvé, and L. Younes. Geodesic shooting for computational anatomy. *Journal of Mathematical Imaging and Vision*, 24(2):209–228, Mar 2006. 5

[21] S. J. D. Prince. Probabilistic Linear Discriminant Analysis for Infermences About Identity. In *Proc. International Conference on Computer Vision*, 2007. 2, 4, 5

[22] S. Sommer. Horizontal dimensionality reduction and iterated frame bundle development. In *Geometric Science of Information*, pages 76–83. Springer, 2013. 2

[23] S. Sommer, F. Lauze, S. Hauberg, and M. Nielsen. *Manifold Valued Statistics, Exact Principal Geodesic Analysis and the Effect of Linear Approximations*, pages 43–56. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. 4, 6

[24] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. 1

[25] L. Younes. *Shapes and diffeomorphisms*. Heidelberg: Springer, 2010. "A direct application of what is presented in the book is a branch of the computerized analysis of medical images called computational anatomy"–Back cover. 5

[26] M. Zhang and P. T. Fletcher. Bayesian principal geodesic analysis in diffeomorphic image registration. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. SpringerLink, 2014. 2, 4, 6