UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA MATEMÁTICA

NONLINEAR PRINCIPAL COMPONENTS ANALYSIS FOR MEASURES AND IMAGES

TESIS PARA OPTAR AL GRADO DE DOCTOR EN CIENCIAS DE LA INGENIERIA, MENCION MODELACION MATEMATICA

ALFREDO IGNACIO LOPEZ ALFAGEME

PROFESOR GUÍA:
RAUL GOUET BAÑARES

MIEMBROS DE LA COMISIÓN:
ROLANDO BISCAY LIRIO
JAIME ORTEGA PALMA
JAIME SAN MARTIN ARISTEGUI

SANTIAGO DE CHILE
2013

RESUMEN DE LA TESIS PARA OPTAR AL GRADO DE DOCTOR EN CIENCIAS DE LA INGENIERÍA, MENCIÓN MODELACIÓN MATEMÁTICA
POR: ALFREDO IGNACIO LOPEZ ALFAGEME
FECHA: SEPTIEMBRE 2013
PROF. GUÍA: DR. RAUL GOUET

ANÁLISIS NO-LINEAL DE COMPONENTES PRINCIPALES APLICADO A MEDIDAS E IMÁGENES

En esta tesis definimos dos adaptaciones no-lineales del análisis de componentes principales, para el estudio de la variabilidad de datos conformados por medidas de probabilidad y por imágenes.

En el Capitulo 2 introducimos el método de análisis de componentes principales geodésico (ACPG) en el espacio de medidas de probabilidad en la línea real, con segundo momento finito, dotado de la métrica de Wasserstein. Apoyándonos en la estructura pseudo-riemanniana del espacio de Wasserstein, definimos el ACPG basado en adaptaciones del ACP a variedades, propuestas en la literatura. En este contexto, el ACPG se define por medio de un problema de minimización sobre el espacio conformado por los subconjuntos geodésicos del espacio de Wasserstein. Usando argumentos de compacidad y de gama-convergencia, establecemos la consistencia del método, demostrando que el ACPG converge a su contraparte poblacional, cuando el tamaño de la muestra crece a infinito. Discutimos las ventajas de este método, respecto a un ACP funcional estándar de medidas de probabilidad en el espacio de Hilbert de funciones a cuadrado integrable. Con el fin de mostrar los beneficios de este procedimiento para el análisis de datos, exhibimos algunos ejemplos ilustrativos en un modelo estadístico simple.

En el Capitulo 3 describimos el método de análisis de componentes principales geométrico (ACP geométrico) para analizar los modos principales de variación geométrica de un conjunto de imágenes. En este contexto proponemos modelar la variabilidad geométrica de las imágenes, respecto a un patrón medio de referencia, por medio de un operador de deformación parametrizado por un espacio de Hilbert. El ACP geométrico consta de dos etapas: (1) registro de imágenes usando un operador de deformación y (2) ACP estándar en los parámetros asociados a las deformaciones. La consistencia del procedimiento es analizada en el contexto de un modelo estadístico de patrón deformable, con una doble asíntota, donde el número de observaciones tiende a infinito y el ruido aditivo converge a cero. Para destacar los beneficios de este procedimiento, describimos un algoritmo y su aplicación a algunos experimentos numéricos con imágenes reales.

Abstract

In this thesis we define two non-linear adaptations of principal components analysis for studying the variability of dataset of probability measures and of images.

In Chapter 2 we introduce the method of Geodesic Principal Component Analysis (GPCA) on the space of probability measures on the line, with finite second moments, endowed with the Wasserstein metric. Relying on the pseudo-Riemannian structure of the Wasserstein space, we define the GPCA based on existing PCA adaptations to manifolds. In this context, the GPCA is defined by means of a minimization problem over the space of geodesic subsets of the Wasserstein space. By using arguments of compactness and gamma-convergence, we establish the consistency of the method by showing that the empirical GPCA converges to its population counterpart, as the sample size tends to infinity. We discuss the advantages of this approach over, a standard functional PCA of probability densities in the Hilbert space, of square-integrable functions, and give illustrative examples on simple statistical models to show the benefits of this approach for data analysis.

In Chapter 3 we describe the method of Geometric Principal Component Analysis (geometric PCA) for analyzing the principal modes of geometric variability of images. In this context, we propose to model the geometric variability of images around a reference mean pattern by using a deformation operator parametrized over a Hilbert space. The geometric PCA is a two-step procedure: (1) image-registration using a deformation operator and (2) standard PCA on the resulting deformation-parameters. The consistency of this procedure is also analyzed in the context of statistical deformable models, with a double asymptotic setting, where the number of observations tends to infinity and the additive noise converges to zero. We describe a simple algorithm for estimating the geometric variability of a set of images and some numerical experiments on real data are proposed, to highlight the benefits of this approach.

Agradecimientos

Quiero agradecer a las personas e instituciones que me han apoyado en la realización de esta tesis doctoral. Agradezco al programa de doctorado del DIM por la oportunidad brindada y por el apoyo a través de la beca MECESUP y al Centro de Modelamiento Matemático, por el financiamiento otorgado durante el periodo de finalización de tesis. También agradezco a la embajada de Francia en Chile y a CONICYT PFCHA-2012, por el financiamiento para hacer dos estadías doctorales en Toulouse, Francia.

Doy gracias a mi supervisor de tesis, Dr. Raúl Gouet, quien me apoyó desde el proceso de postulación y me ha guiado en este trabajo con gran dedicación y profesionalismo. A Raúl lo conozco desde que dirigió mi memoria de ingeniero y una parte importante de mi formación como investigador se la debo a él.

También agradezco especialmente al Dr. Jérémie Bigot, con quien desarrollamos muchas de las ideas expuestas en esta tesis, a partir de su visita al CMM el año 2010. Destaco la excelente hospitalidad y apoyo financiero que Jérémie me brindó durante mis dos estadías doctorales en Toulouse, Francia.

Mi agradecimiento también es para los miembros de la comisión, Rolando Biscay, Jaime Ortega y Jaime San Martín.

Agradezco al laboratorio MOTIV-CMM y en especial a sus miembros Dr. Jaime Ortega y Dr. Takeshi Asahi, por las numerosas invitaciones a participar en proyectos aplicados al procesamiento de imágenes. Destaco los proyectos FONDEF D04-I-1237 y FONDECYT 1111012. Finalmente doy gracias a Dr. Julian Mattes, con quien trabajé en interesantes proyectos relacionados con procesamiento de imágenes, durante el periodo inicial de mis estudios doctorales.

Contents

1 Introduction **1**
1.1 PCA on Hilbert spaces 4
1.1.1 Standard PCA on probability measures 6
1.1.2 Standard PCA on images 6
1.2 Analogs of PCA for data belonging to a Riemannian manifold 8
1.2.1 Riemannian manifolds 8
1.2.2 Fréchet mean on a Riemannian manifold 9
1.2.3 PGA along geodesics on a Riemannian manifold 9
1.2.4 PGA via linearization in the tangent space 10
1.3 Summary of contributions and results of this thesis 11

2 Geodesic principal components analysis in the Wasserstein space **12**
2.1 Introduction 12
2.1.1 Main goal of this chapter 12
2.1.2 Main contributions and organization of the chapter 13
2.2 Preliminaries and definitions 13
2.2.1 A running example: the homothetic model 13
2.2.2 Optimal transportation theory 14
2.2.3 The pseudo-Riemannian structure of the Wasserstein space 15
2.2.4 Geodesics in the Wasserstein space 17
2.3 Population Fréchet mean and principal geodesics 19
2.3.1 Fréchet Mean 19
2.3.2 Principal geodesics 20
2.3.3 Empirical Fréchet mean and principal geodesics 22
2.3.4 Formulation of GPCA as an optimization problem in $L^2_\mu(\Omega)$ 23
2.3.5 PCA on logarithms 25
2.4 Some numerical examples of GPCA in W_2 26
2.4.1 The case of sufficiently concentrated data 26
2.4.2 The case of spread-out data 28
2.5 Consistency of the empirical GPCA 31
2.6 Conclusions and discussion of this chapter 33
2.7 Extensions and related problems 34
2.7.1 Consistency of GPCA in a statistical deformable model 35
2.7.2 GPCA in the case of non-compactly supported measures 36
2.7.3 GPCA based on kernel density estimation 40

3 Geometric principal components analysis of images **41**
3.1 Introduction . . . 41
3.1.1 An overview of PCA-like methods for analyzing geometric variability 41
3.1.2 Main contributions and organization of this chapter . . . 43
3.2 Geometric PCA . . . 43
3.2.1 Grenander's pattern theory of deformable templates . . . 43
3.2.2 Registration . . . 45
3.2.3 Numerical implementation and application of geometric PCA to handwritten digits data . . . 47
3.3 Consistency of geometric PCA in statistical deformable models . . . 52
3.3.1 Case of template registration . . . 55
3.3.2 Case of groupwise registration . . . 56
3.3.3 Translation operator . . . 57
3.3.4 Non-rigid diffeomorphic operator . . . 58
3.4 Conclusions and discussion of this chapter . . . 59
3.5 Extensions and related problems . . . 60
3.5.1 Tikhonov regularization . . . 60
3.5.2 Discrete model . . . 61

Conclusions **62**

References **63**

A Appendix of Chapter 2 **69**
A.1 Quantile functions . . . 69
A.2 Geodesics in metric spaces . . . 70
A.3 K-convergence . . . 71
A.4 $\Gamma-$convergence . . . 72
A.5 Minkowski functional . . . 73

B Appendix of Chapter 3 **76**
B.1 Preliminary technical results . . . 76
B.2 Proofs of main results . . . 81
B.2.1 Proof of Theorem 3.1 . . . 81
B.2.2 Proof of Theorem 3.2 . . . 82
B.2.3 Proof of Theorem 3.3 . . . 83
B.3 Technical results for translation operators . . . 83
B.3.1 Proof of Lemma 3.1 . . . 84
B.3.2 Proof of Proposition 3.3 . . . 84
B.4 Technical results for non-rigid diffeomorphic operators . . . 85
B.4.1 Proof of Proposition 3.4 . . . 85
B.4.2 Proof of Proposition 3.5 . . . 87

Chapter 1

Introduction

A central problem in many applications is the characterization a of population's variability around an estimated mean pattern. If the data belongs to a Hilbert space, then a common approach for addressing such problem is Principal component analysis (PCA). However, in many situations the data cannot be properly modeled as element in a Hilbert space and thus PCA does not lead to meaningful results. The latter situation occurs, in particular, for two instances of data we investigate in this thesis, namely images and probability measures on the line. In Chapter 2 we define and analyze the *geodesic principal analysis in the Wasserstein space*, a framework for analyzing the variability of a set of probability measures on the line. Chapter 3 is devoted to the description and analysis of the *geometric PCA of images*, which is a method well adapted for describing the geometric variability of a set of images. In both chapters, our main theoretical concern is the probabilistic consistency of these methods with respect to a given population model. Though both chapters are self contained and can be read independently, there exists a common basis and several links between them. In the following paragraphs we describe the different building-blocks of this thesis work.

PCA in multivariate data analysis. Commonly, the data can be represented as elements in the vector space $\mathbb{R}^d$ and, in this case, Principal Component Analysis (PCA) [Jolliffe, 2002] is one of the most used technique for describing variability. PCA on $\mathbb{R}^d$ consist of finding an orthonormal basis such that, for any $k \leq d$, the projection of the data onto the subspace generated by the first k vectors of the basis has maximum variance. This is achieved by diagonalizing the covariance matrix of the data, sorting the eigenvalues in decreasing order and taking the associated eigenvectors. PCA is a widely used method with applications in several fields, it is commonly used as a dimensionality reduction method and combined with other multivariate techniques such as discriminant, cluster or correlation analysis.

Functional PCA. In many applications the data have the form of functions defined on a continuous domain. For instance, time series in econometrics, medical images or probability density functions estimated from frequency histograms. Such data cannot always be represented by elements in a finite dimensional vector space and therefore an infinite dimensional framework is necessary for its analysis. The extension of finite dimensional PCA for functional data is usually known as functional PCA (FPCA), see [Ramsay and Silverman, 2002, 2005] for an introduction and [Shang, 2013] for a recent survey on FPCA. In this setting, the data are considered as elements in the Hilbert space $L^2_\mu(\Omega)$, of square integrable functions defined on a convex set $\Omega \subset \mathbb{R}^d$, endowed with its usual inner product. Analogously to the finite-dimensional case, FPCA amounts to computing the spectral decomposition of the covariance operator of the data and the first k eigenfunctions associated with the k largest eigenvalues recover the variability of the data in an optimal sense.

PCA on a Hilbert space. The general mathematical framework for defining PCA is to consider the data belonging to a Hilbert space H. This setting includes the cases $H = \mathbb{R}^d$ and $H = L^2_\mu(\Omega)$ described above. The choice of the inner product in H plays an important role in modeling the data variability. It allows for instance, to consider smoothness constraints by taking a Sobolev inner product [Silverman, 1996]. In Section 1.1 we describe the PCA on a Hilbert space in detail.

Analogs of PCA for data belonging to a Riemannian manifold. Modern applications of image analysis and computer vision require representation of the data in non euclidean spaces, such as Riemannian manifolds [Shi et al., 2009; Sommer et al., 2010], leading to the development of extensions of PCA for data belonging to a Riemannian manifold [Fletcher et al., 2004; Huckemann et al., 2010]. The main idea of such methods, is to find geodesic submanifolds (generalization of linear subspaces), capturing the variability of the data in an optimal sense. In [Fletcher et al., 2004] a linear approximation is considered, by means of which the data on the manifold is mapped onto a suitable Hilbert space, where a standard PCA (or FPCA) is performed. In contrast, the methods analyzed in [Huckemann et al., 2010; Sommer et al., 2010] deal with the inherent manifold structure, thus more sophisticated theoretical and numerical tools are required. For more details, see Sections 1.2.3 and 1.2.4.

Geodesic PCA in the Wasserstein space. In some situations of applied interest, the data takes the form of probability measures on a common space; see [Kneip and Utikal, 2001] for a motivating application. In this thesis we develop and study a new PCA-type technique for analyzing the variability of such data. If each of these probability measures admits a square integrable density with respect to Lebesgue's measure, then one can consider the data as points on the space of square-integrable functions and carry out a standard FPCA, as done in [Kneip and Utikal, 2001]. However, such an analysis does not always leads to interpretable results (see Section 1.1.1), as the set of densities is not a linear subspace and the Euclidean distance is not appropriate to compare densities. In this thesis we suggest that the appropriate framework to consider is the space W_2 of probability measures, with finite second moment, endowed with the Wasserstein distance (see [Villani, 2003]). We believe that, in our context, the Wasserstein metric is particularly well adapted because its definition relies on ideas close to "deformation" or "transportation" stress between two probability measures. Besides this attractive interpretation, the Wasserstein metric space has convenient technical properties well adapted to our methodology, such as a pseudo-Riemannian structure [Ambrosio and Savaré, 2007]. Using this property, we introduce a new concept of PCA in W_2 by mimicking the PCA on manifolds, mentioned in the previous paragraph. We believe this framework represents a novel contribution to the field of statistical applications of optimal transportation theory. This subject will be thoroughly developed in Chapter 2.
It worth to mention that there are a number of classical references (see [Amari and Nagaoka, 2000; Murray and Rice, 1993]) that define and study the Riemanian manifold structure of parametric families of probability measures, when the parameter space is a finite dimensional vector space and the dependence on the parameter is one-to-one and smooth. In this context, the associated Riemaniann metric (and thus geodesics) is based on the Fisher information

matrix. We believe that it might be of interest (but out of our scope) to apply the PCA-like methods defined in the previous paragraph, based on the definition of geodesics derived from the Fisher information matrix. In contrast to the previous framework, our approach does not require that the family of probability measures considered be smoothly parametrized over a Hilbert space, however, applying our method for such families of probability measures might constitute an interesting future line of research.

PCA-like methods for analyzing geometric variability of images. An important problem in the field of image analysis, is the characterization of the shape variability of a given object within a population. For instance, it is possible to asses the state of a degenerative disease, track a growth process or classify healthy versus ill subjects, by analyzing shape discrepancy with respect to a reference. Images can be considered as element of the Hilbert space $L^2_\mu(\Omega)$, however PCA applied to a set of images does not lead to meaningful results (see Section 1.1.2) as it consists only on a point-wise analysis and no geometric information of the objects represented in the images is considered. A standard approach to analyze geometric variability of a set of images is to use registration. This well-known method consists in computing parametric transforms on the domain Ω, so that the images deformed by these transforms are geometrically aligned. Then, the main idea to estimate the geometric variability of such data is to apply classical PCA to the resulting transformation parameters after registration and not to the images themselves. This approach is at the core of several methods to estimate the geometrical variability of images and constitutes the basis of the general framework we call geometric PCA.

Geometric PCA of images. Grenander's pattern theory of deformable templates [Grenander and Miller, 1993, 2007] provides a mathematical framework for analyzing the geometric variability of images and, in particular, for defining and analyzing the geometric PCA. Following Grenander's ideas, one may consider that the observed images (or, more generally, shapes) are obtained through the deformation of a common reference image (resp. shape). In this setting, images are treated as points in $L^2(\Omega)$ and the geometric variations of the images are modeled by Lie groups (groups that are Riemannian manifolds) of transformations acting on the domain Ω [Grenander and Miller, 1998; Miller and Younes, 2001]. We propose to model the geometric variability through the use of what we call *deformation operators*. A deformation operator can be understood as a family of smooth and invertible mappings from Ω onto Ω, parametrized by elements in a Hilbert space. Then, we define the geometric PCA as a two step procedure for analyzing the geometric variability of a set of images. The first step consists of groupwise registration of the images by means of a deformation operator. The second step consists of PCA applied to the resulting parameters (which are elements of a Hilbert space) after registration. We remark here that, from the methodological point of view, geometric PCA can be interpreted as PCA on the "manifold" of deformations and hence, there exists a link between the geometric PCA and the PCA-like methods for manifold we mentioned before. In Section 3.1.1 we provide a brief overview of related method existing in the literature and specially in the context of Grenander's theory of deformable templates. The geometric PCA of images is defined and analyzed in Chapter 3.

Probabilistic consistency. Our main theoretical results deal with the probabilistic consistency of the methods we propose. We assume that the data are independent observations from a given stochastic model and we study the convergence of the method's output to its population counterpart when the sample size growth. This is a technically complicated problem, that constitutes one of the original contributions of this thesis. On the other hand, various author have studied the asymptotic properties of PCA in the multivariate and the functional settings [Anderson, 1963; Dauxois et al., 1982].

1.1 PCA on Hilbert spaces

Let us introduce some tools and notations related to the standard PCA in a Hilbert space, that will be used throughout the thesis. Let H be a separable Hilbert space, endowed with the inner product $\langle\cdot,\cdot\rangle$ and the associated norm $\|\cdot\|$. Let d be the induced distance in H and denote by $d(v,S)=\inf_{v'\in S}\|v-v'\|$ the distance from $v\in H$ to $S\subset H$. Let V be an H-valued random variable. If $\mathbb{E}\|V\|<+\infty$, then V has expectation $\mathbb{E}V\in H$, which happens to be the unique element satisfying $\langle \mathbb{E}V, v\rangle = \mathbb{E}\langle V, v\rangle$, for all $v\in H$. If $\mathbb{E}\|V\|^2<+\infty$, then the (population) covariance operator $K: H\to H$ corresponding to V is

$$Kv = \mathbb{E}\langle V-\mathbb{E}V, v\rangle (V-\mathbb{E}V), \text{ for } v\in H. \tag{1.1}$$

Moreover, the operator K is self-adjoint, positive semidefinite and trace-class. Hence, K is compact, with nonnegative (population) eigenvalues $(\lambda_k)_{k\in\mathbb{K}}$ and orthonormal (population) eigenvectors $(\phi_k)_{k\in\mathbb{K}}$ and such that

$$Kv = \sum_{k\in\mathbb{K}} \lambda_k \langle v, \phi_k\rangle \phi_k,$$

where $\mathbb{K}=\{1,\dots,dim(H)\}$, if $dim(H)<\infty$ or $\mathbb{K}=\mathbb{N}$ otherwise. If we assume that the eigenvalues are arranged in decreasing order $\lambda_1\geq\lambda_2\geq\ldots\geq 0$, the k-th mode of variation of the random variable V is defined by the curve

$$\gamma_k(t) = \mathbb{E}V + t\phi_k,\ t\in\mathbb{R}. \tag{1.2}$$

We define the PCA of V as the problem of diagonalizing the covariance operator K defined in (1.1). On the other hand, it is well known that the PCA can be formulated as the problem of finding a sequence of nested affine subspaces, minimizing the norms of the projection residuals of the data. It can be shown that $\mathbb{E}V$ can be characterized as

$$\mathbb{E}V = \arg\min_{v\in H} \mathbb{E}\|v-V\|^2 \tag{1.3}$$

and the eigenvectors $(\phi_k)_{k\in\mathbb{K}}$ of K can be characterized as

$$\begin{aligned} \phi_k &\in \arg\min_{v\in E_k^\perp, \|v\|=1} \mathbb{E}\left(d^2(V,\ EV+\operatorname{span}(v))\right) && (1.4)\\ &= \arg\min_{v\in E_k^\perp, \|v\|=1} \mathbb{E}\|V-\mathbb{E}V-\langle V-\mathbb{E}V, v\rangle v\|^2 && (1.5)\\ &= \arg\max_{v\in E_k^\perp, \|v\|=1} \mathbb{E}\langle V-\mathbb{E}V, v\rangle^2, && (1.6)\end{aligned}$$

where $E_1=\{0\}$, $E_k=\operatorname{span}(\phi_1,\dots,\phi_{k-1})$, $k\geq 2$, and $\operatorname{span}(\mathcal{U})$ is the subspace spanned by $\mathcal{U}\subset H$.

Now, let us consider $v_1, \ldots, v_n \in H$ and let $V^{(n)} \in H$ be the random element such that $\mathbb{P}(V = v_i) = 1/n,\ i = 1, \ldots, n$. We define the empirical PCA of $v_1, \ldots, v_n$ as the PCA of the random element $V^{(n)}$, i.e., the empirical PCA consists of diagonalizing the empirical covariance operator $K_n : H \to H$, given by

$$K_n v = \frac{1}{n} \sum_{i=1}^{n} \langle v_i - \bar{v}, v \rangle (v_i - \bar{v}), \text{ for } v \in H, \tag{1.7}$$

where, $\bar{v} = \frac{1}{n} \sum_{i=1}^{n} v_i$. As K is also self-adjoint, positive semidefinite and compact, it admits the decomposition

$$K_n v = \sum_{k \in \mathbb{K}} \lambda_k \langle v, \phi_k \rangle \phi_k,$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ are the (empirical) eigenvalues, and $(\phi_k)_{k \in \mathbb{K}}$ is the set of (empirical) orthonormal eigenvectors of K_n. The (empirical) principal k-th mode of linear variation of $v_1, \ldots, v_n$ is defined by the curve

$$\gamma_k(t) = \bar{v} + t\phi_k,\ t \in \mathbb{R}. \tag{1.8}$$

As in (1.3) to (1.6), $\bar{v}$ can be characterized as

$$\bar{v} = \arg\min_{v \in H} \frac{1}{n} \sum_{i=1}^{n} \|v - v_i\|^2 \tag{1.9}$$

and the eigenvectors $(\phi_k)_{k \in \mathbb{K}}$ of K_n can be characterized as

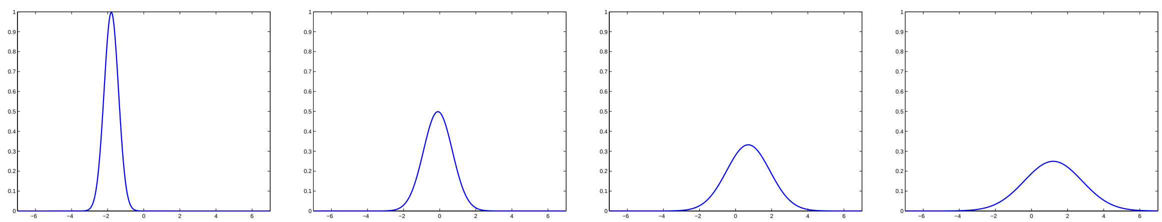
$$\begin{aligned}
\phi_k &\in \arg\min_{v \in E_k^{\perp}, \|v\|=1} \frac{1}{n} \sum_{i=1}^{n} d^2(v_i, \bar{v}_n + \text{span}(v)) && (1.10)\\
&= \arg\min_{v \in E_k^{\perp}, \|v\|=1} \frac{1}{n} \sum_{i=1}^{n} \|v_i - \bar{v}_n - \langle v_i - \bar{v}_n, v \rangle v\|^2 && (1.11)\\
&= \arg\max_{v \in E_k^{\perp}, \|v\|=1} \frac{1}{n} \sum_{i=1}^{n} \langle v_i - \bar{v}_n, v \rangle^2, && (1.12)
\end{aligned}$$

where $E_1 = \{0\}$ and $E_k = \text{span}(\phi_1, \ldots, \phi_{k-1}),\ k \geq 2$.

If H is finite-dimensional, diagonalizing K_n corresponds to the empirical PCA for vectors in a finite dimensional Euclidean space. If $H = L^2(\Omega) = \left\{ f : \Omega \to \mathbb{R},\ \|f\|_2^2 := \int_{\Omega} f^2(t) dt < \infty \right\}$ diagonalizing K_n is usually referred to as the method of functional PCA in nonparametric statistics (see e.g. [Dauxois et al., 1982; Ramsay and Silverman, 2005; Silverman, 1996]). Various authors (see e.g. [Dauxois et al., 1982], [Silverman, 1996] and references therein), under standard probabilistic assumptions on the data, have studied the consistency of the empirical PCA in Hilbert spaces and have stated sufficient conditions to ensure that the empirical eigenvalues and eigenvectors converge to the population eigenvalues and eigenvectors, as $n \to +\infty$.

1.1.1 Standard PCA on probability measures

The main goal of Chapter 2 is to define a notion of principal component analysis (PCA) of a family of probability measures $\nu_1, \ldots, \nu_n$, defined on the real line $\mathbb{R}$. In the case where the measures admit square-integrable densities $f_1, \ldots, f_n$, with respect to the Lebesgue measure, the standard approach is to use functional PCA (FPCA) (see e.g. [Dauxois et al., 1982; Ramsay and Silverman, 2005; Silverman, 1996]) on the Hilbert space $L^2(\mathbb{R})$ of square-integrable densities, endowed with its usual inner product. This method has already been applied in [Kneip and Utikal, 2001] for analyzing the main modes of variability of a set of densities.



(a) (b) (c) (d)

Figure 1: Graphs of Gaussian densities $f_1, \ldots, f_4$, with different means and variances.

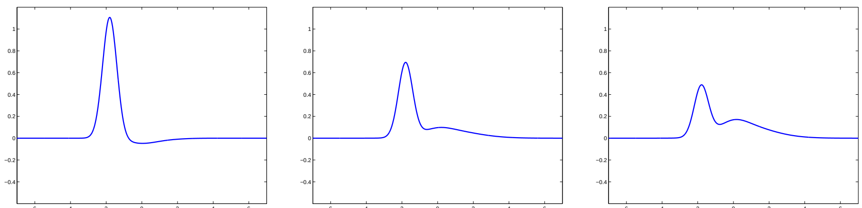
Let us illustrate the strategy described above on a set of Gaussian densities $f_1, \ldots, f_4$, shown in Figure 1, in the Hilbert space $H = L^2(\mathbb{R})$. We first compute the Euclidean mean $\bar{f} = \frac{1}{4}\sum_{i=1}^{4} f_i$, shown in Figure 2(d), and obtain a bi-modal density which is not a "satisfactory" average of the uni-modal densities $f_1, \ldots, f_4$. We also compute the first empirical mode of linear variation (1.8), defined by the curve (we omit the dependence on $k = 1$),

$$\gamma(t) = \bar{f} + t\phi,\ t \in \mathbb{R}, \tag{1.13}$$

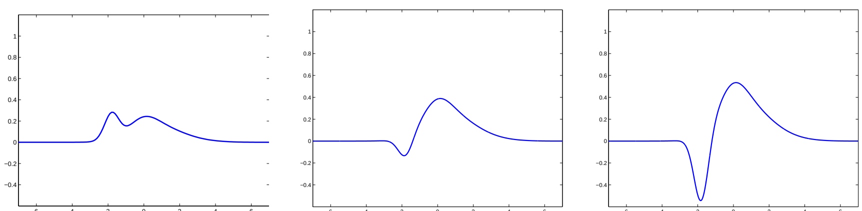
where $\phi \in L^2(\mathbb{R})$ is the eigenvector associated with the largest eigenvalue λ, of the empirical covariance operator $K_4 : L^2(\mathbb{R}) \to L^2(\mathbb{R})$, defined by $K_4(f) = \frac{1}{4}\sum_{i=1}^{4}\langle f_i - \bar{f}, f\rangle_2 (f_i - \bar{f})$. We observe that γ is not a "meaningful" descriptor of the data variability. Indeed, for $|t|$ sufficiently large, the function $\gamma(t)$ may take negative values and does not integrate to one, as illustrated in Figure 2(e),(f). Moreover, even for small values of $|t|$, $\gamma(t)$ does not represent the typical shape of the observed densities $f_1, \ldots, f_4$, as shown by Figure 2(c),(d). Therefore, functional PCA of densities in $L^2(\mathbb{R})$ is not always interpretable as it leads to principal modes of linear variation that may be not coherent with the shape variability that is observed in the data. This drawback of functional PCA comes from the fact that the Euclidean distance $\|f_1 - f_2\|_{L^2(\mathbb{R})}$ is not necessarily appropriate to compare two measures ν_1 and ν_2, admitting f_1 and f_2 as densities.

1.1.2 Standard PCA on images

In Chapter 3 we define a PCA-like method well suited for describing the variability of a set of images $y_1, \ldots, y_n$, which can be considered as square-integrable functions on a domain Ω, a convex subset of $\mathbb{R}^d$. Let us illustrate the method of standard PCA in the Hilbert space $H = L^2_\mu(\Omega)$, applied to a set of $n = 30$ images of handwritten digits $y_1, \ldots, y_{30}$; see Figure



(a) $\gamma(-2\lambda)$ (b) $\gamma(-\lambda)$ (c) $\gamma(-0.5\lambda)$



(d) $\gamma(0) = \bar{f}$ (e) $\gamma(\lambda)$ (f) $\gamma(2\lambda)$

Figure 2: An example of functional PCA of densities. First principal mode of linear variation $\gamma(t)$ in the Hilbert space $L^2(\mathbb{R})$ for $-2\lambda \leq t \leq 2\lambda$, see equation (1.13), of the densities displayed in Figure 1.

3. We compute the Euclidean mean $\bar{y} = \frac{1}{30}\sum_{i=1}^{30} y_i$ and, for $k = 1, 2$, we compute also the k-empirical mode of linear variation of the data (1.8), defined by the curve,

$$\gamma_k(t) = \bar{y} + t\phi_k, \; t \in \mathbb{R}, \tag{1.14}$$

where $\phi_k \in L^2_\mu(\Omega)$ is an eigenvector associated with the k-th largest eigenvalue $\lambda_k \geq 0$ of the empirical covariance operator $K_{30} : L^2_\mu(\Omega) \to L^2_\mu(\Omega)$, with $K_{30}y = \frac{1}{30}\sum_{i=1}^{30} \langle y_i - \bar{y}, y \rangle_2 (y_i - \bar{y})$.

Empirical PCA on $H = \mathrm{L}^2(\Omega)$, applied to this set of images, is a method to compute the principal directions of photometric variability of the y_i's around the Euclidean mean $\bar{y}$. However, these images also exhibit a large geometric variability; see Figure 3. In such a case, $\bar{y}$ is not a satisfying estimator of the typical shape of each individual image, see Figure 4 (a), and standard PCA does not meaningfully reflect the modes of variability of the data, see Figure 4 (b),(c). In particular, the second empirical mode of variation $\gamma_2(t)$ at $t = -\lambda_2$ is no longer a single digit but rather the superposition of two digits in different orientations.

The idea underlying PCA is that the Hilbert space $\mathrm{L}^2(\Omega)$, equipped with the standard inner product, is well suited to model natural images. However, the set of such objects (as those in Figure 3) typically cannot be considered as a linear sub-space of $\mathrm{L}^2(\Omega)$. Therefore, the Euclidean distance $\|y_1 - y_2\|_2$ is generally not well suited, since it is not adapted to the geometry of the set to which the images y_1 and y_2 truly belong. Actually one can see that the images in Figure 3 have mainly a geometric variability in space, which is much more important than the photometric variability.

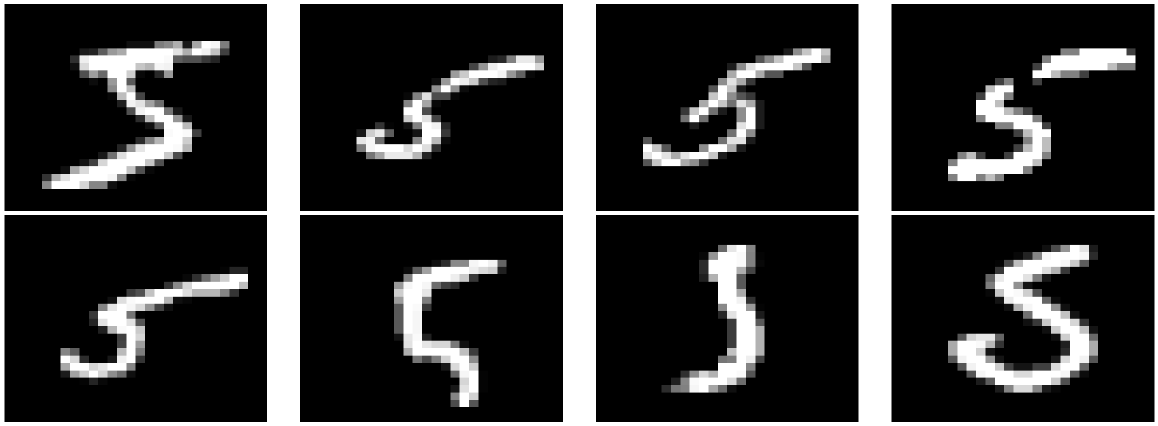
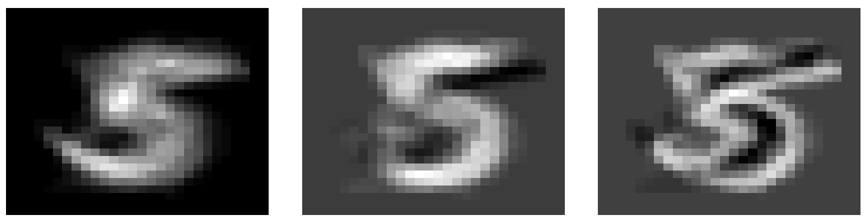


Figure 3: Digit 5. A sample of 8 images out of $n = 30$ taken from the Mnist database [LeCun et al., 1998]



(a) $\bar{y}$ (b) $\gamma_1(-\lambda_1)$ (c) $\gamma_2(-\lambda_2)$

Figure 4: Standard Euclidean mean, first and second empirical modes of variation $\gamma_1(t)$ and $\gamma_2(t)$ at $t = -\lambda_1$ and $t = -\lambda_2$ respectively.

1.2 Analogs of PCA for data belonging to a Riemannian manifold

There is currently a growing interest in the statistical literature on the development of non-linear analogs of PCA, for the analysis of data belonging to curved Riemannian manifolds; see e.g. [Fletcher et al., 2004; Huckemann et al., 2010; Sommer et al., 2010] and references therein. These methods, generally referred to as Principal Geodesic Analysis (PGA), extend the notion of classical PCA on Hilbert spaces. After providing a brief overview of Riemannian Manifolds, we describe some of the main ideas of PGA. In order to be coherent with the literature, we give such description in the empirical setting, that is, we describe the method of PGA applied to a data set $p_1, \ldots, p_n$ belonging to a complete Riemannian manifold $\mathcal{M}$. However, it could be formulated as a population PGA applied to a $\mathcal{M}$-valued random element.

1.2.1 Riemannian manifolds

A Riemannian manifold $\mathcal{M}$ is a topological space, locally homeomorphic at every $p \in \mathcal{M}$ to a Hilbert space $(T_p\mathcal{M}, \langle , \rangle_p)$, called the tangent space at p. Given a smooth curve $\gamma : I \to \mathcal{M}$ from an interval $I \subset \mathbb{R}$, it is possible to define the derivative or velocity at $t \in I$, which is an element of $T_{\gamma(t)}\mathcal{M}$. The length of γ is defined as the integral of the norm of the velocity over $t \in I$. The geodesic distance $d_{\mathcal{M}}(p_1, p_2)$ between $p_1, p_2 \in \mathcal{M}$ is the minimum length over all smooth curves with endpoints p_1 and p_2. Geodesics, which are generalizations of straight lines in Hilbert spaces to manifolds, are smooth curves that are locally length minimizer. The manifold is said to be complete if every geodesic can be extended indefinitely. Now, given a reference element $q \in \mathcal{M}$ and v a tangent vector in $T_q\mathcal{M}$, it is known that there exists a unique geodesic γ_v such that $\gamma_v(0) = q$, having v as initial velocity. If $\mathcal{M}$ is complete, then the Riemannian exponential map at q, $\exp_q : T_q\mathcal{M} \to \mathcal{M}$, is defined by $\exp_q(v) = \gamma_v(1)$. Therefore, the curve $t \to \exp_q(tv), t \in \mathbb{R}$, is a geodesic. Generally, the exponential map is a diffeomorphism only in a neighborhood of zero and its inverse is the Riemannian log map, that we denote by $\log_q$. The log map is not, in general, an isometry, however $d_{\mathcal{M}}(q, p) = \| \log_q(p)\|_q$ for all $p \in \mathcal{M}$, where $\|\|_q$ denotes the norm in $T_q\mathcal{M}$.

1.2.2 Fréchet mean on a Riemannian manifold

Consider $p_1, \dots, p_n \in \mathcal{M}$. In order to define a PGA one needs a notion of average on $\mathcal{M}$. It has been suggested [Fletcher et al., 2004] that the appropriate notion is the Frechet mean, defined as an element $q \in \mathcal{M}$ (not necessarily unique) minimizing the sum of squared distances to the data, namely

$$q \in \underset{p\in\mathcal{M}}{\arg\min} \, \frac{1}{n}\sum_{i=1}^{n} d^2_{\mathcal{M}}(p, p_i).$$

From (1.9) we can observe that the Fréchet mean generalizes the Euclidean mean in a Hilbert space. We refer to [Bhattacharya and Patrangenaru, 2003] for further details and properties of the Fréchet mean in Riemannian manifolds. After computing a Fréchet mean q, it is possible to define the main modes of variability of the data in two different ways, that are described below.

1.2.3 PGA along geodesics on a Riemannian manifold

Recall that the curve $t \to \exp_q(tv)$, $t \in \mathbb{R}$, with $v \in T_q\mathcal{M}$, is a geodesic, thus $\exp_q(\mathrm{span}(v))$ generalizes the concept of one dimensional affine sub-space through q. The notion of PGA along geodesics on $\mathcal{M}$ is inspired by problem (1.10), which characterizes standard PCA. In a first step, one computes

$$\phi_k^{geo} \in \underset{v\in E_k^{\perp},\, \|v\|=1}{\arg\min} \, \frac{1}{n}\sum_{i=1}^{n} d^2_{\mathcal{M}}\left(p_i, \exp_q(\mathrm{span}(v))\right), \tag{1.15}$$

where $E_1 = \{0\}$ and $E_k = \mathrm{span}(\phi_{1,n}^{geo}, \dots, \phi_{k-1,n}^{geo})$, $k \geq 2$. Then, in a second (and final) step, one projects the element $\phi_k^{geo} \in T_q\mathcal{M}$ onto $\mathcal{M}$, by computing the geodesic

$$\gamma_k^{geo}(t) = \exp_q(t\phi_k^{geo}), t \in \mathbb{R}.$$

We remark that, among other differences, PGA was originally proposed in [Fletcher et al., 2004] in terms of maximizing variance. In the case of PCA on Hilbert spaces, both formulations, minimizing residuals (1.11) and maximizing variance (1.12), are equivalent. In contrast, in the case of PGA, such formulations do not coincide, as discussed in [Huckemann et al., 2010; Sommer et al., 2010]. We also remark that PGA along geodesic, is a complicated procedure that requires sophisticated theoretical and numerical tools [Huckemann et al., 2010; Sommer et al., 2010].

1.2.4 PGA via linearization in the tangent space

As we said before, (1.15) is a difficult problem, thus in [Fletcher et al., 2004] an approximation through linearization in the tangent space is proposed. In this approach, the data $p_1, \dots, p_n$ is first projected on $T_q\mathcal{M}$ by means of the $\log_q$ map, thus obtaining $v_i = \log_q(p_i),\ i = 1, \dots, n$. Next, a standard PCA of $v_1, \dots, v_n$ is performed in the linear space $(T_q\mathcal{M}, \langle \cdot, \cdot \rangle, \| \cdot \|)$, which amounts to computing the eigenvectors $(\phi_k^{lin})_{k \in \mathbb{K}}$ associated with the eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots$, of the covariance operator

$$K_n v = \frac{1}{n} \sum_{i=1}^{n} \langle v_i - \bar{v}, v \rangle (v_i - \bar{v}),\ v \in T_q\mathcal{M},$$

where $\bar{v} = \frac{1}{n} \sum_{i=1}^{n} v_i$. Finally, ϕ_k^{lin} is projected back onto $\mathcal{M}$ by means of the $\exp_q$ map, to obtain the geodesic

$$\gamma_k^{lin}(t) = \exp_q(t\phi_k^{lin}), t \in \mathbb{R},$$

which represents a second notion of principal geodesic.

PGA via linearization can be understood as an approximation of PGA along geodesic, described in Section 1.2.3. Indeed, consider the approximation

$$d_{\mathcal{M}}(p, p_i) \approx \| \log_q(p) - \log_q(p_i) \|_q,\ i = 1, \dots, n, \tag{1.16}$$

under the assumption that $p, p_1, \dots, p_n$ are well concentrated around the Fréchet mean q and the curvature of $\mathcal{M}$ at q is small. Then, from (1.9) and (1.16), $\bar{v} \approx \log_q(q) = 0$ and, again by (1.16), the objective function in (1.15) can be approximated as

$$\frac{1}{n} \sum_{i=1}^{n} d^2_{\mathcal{M}} \left(p_i, \exp_q(\mathrm{span}(v)) \right) \approx \frac{1}{n} \sum_{i=1}^{n} d^2_{\mu}(v_i, \mathrm{span}(v)) \approx \frac{1}{n} \sum_{i=1}^{n} d^2_{\mu}(v_i, \bar{v} + \mathrm{span}(v)). \tag{1.17}$$

The right-hand side above, characterizes the standard PCA of $v_1, \dots, v_n$ (see (1.10)).

This yields another definition of principal geodesics of the data and generally one has that $\gamma_k^{lin} \neq \gamma_k^{geo}$, except if $\mathcal{M}$ is a Hilbert space. Therefore, PGA via linearization on the tangent space and PGA along geodesics may lead to different definitions of principal geodesic in a curved manifold. A detailed analysis of the differences between these methods can be found in [Sommer et al., 2010]. Observe that PGA via linearization is easy to compute as it amounts to diagonalizing the covariance operator K_n. On the other hand, due to the approximation (1.17), if the data is not well concentrated around the Fréchet mean q or $\mathcal{M}$ has high curvature at q, then PGA via linearization might not be a good descriptor of the variability of the data [Huckemann et al., 2010; Sommer et al., 2010].

1.3 Summary of contributions and results of this thesis

The main contribution of this thesis are the following

1. Definition and analysis of the *geodesic principal analysis in the Wasserstein space*, a framework for analyzing the variability of a set of probability measures on the line. Since the Wasserstein space W_2 is not a Riemannian manifold, one cannot directly use the methods described in Section 1.2 to define a notion of PCA. Nevertheless, the space W_2 has a formal Riemannian structure [Ambrosio and Savaré, 2007] that we use to define a new notion of Geodesic PCA (GPCA) similar to those described in Section 1.2. Our main results of existence of GPCA are Theorem 2.2 and Theorem 2.3, corresponding to GPCA in two variants, called global and nested.

2. An illustrative example of GPCA based on Gaussian densities. We show in this case that GPCA can be solved using a simple numerical procedure.

3. Proof of the strong consistency of the empirical GPGA performed on a random data set. In this context, our main result is Theorem 2.4.

4. Definition and analysis of the *geometric PCA of images*, which is a method well adapted for describing the geometric variability of a set of images. The definition of the geometric PCA relies on standard procedures in the field of image processing. We define the geometric PCA in a framework where the spatial deformation operators (to represent geometric variability) are invertible and can be parametrized by elements of a Hilbert space. For estimating geometric variability, we use a preliminary registration step. Then, we apply classical PCA on Hilbert spaces to the resulting parameters representing the deformations after registration. The existence of geometric PCA is ensured by Proposition 3.1, under certain regularity conditions.

5. Development of an optimization procedure and numerical implementation of the geometric PCA, applied to a set of handwritten images.

6. Proof of the consistency of the geometric PCA in statistical deformable models. Here, our main results are Theorems 3.1 and 3.3, corresponding to geometric PCA using two alternative registration procedures, called template and groupwise, respectively.

Chapter 2

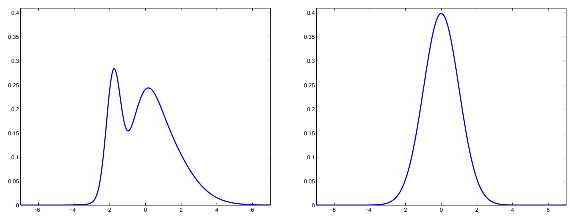
Geodesic principal components analysis in the Wasserstein space

2.1 Introduction

2.1.1 Main goal of this chapter

The main goal of this chapter is to define a notion of principal component analysis (PCA) of a family of probability measures $\nu_1, \ldots, \nu_n$, defined on the real line $\mathbb{R}$. In the case where the measures admit square-integrable densities $f_1, \ldots, f_n$, with respect to the Lebesgue measure, the standard approach is to use functional PCA (FPCA) (see e.g. [Dauxois et al., 1982; Ramsay and Silverman, 2005; Silverman, 1996]) on the Hilbert space $L^2(\mathbb{R})$ of square-integrable functions, endowed with its usual inner product. This method has already been applied in [Kneip and Utikal, 2001] for analyzing the main modes of variability of a set of densities.

In Section 1.1.1 we illustrate standard PCA on a set of $n = 4$ Gaussian densities $f_1, \ldots, f_4$, shown in Figure 1. From this analysis, we concluded that functional PCA of densities in $L^2(\mathbb{R})$ is not always interpretable as it leads to principal modes of linear variation that may be not coherent with the shape variability that is observed in the data. In this thesis, we suggest to rather consider that the measures $\nu_1, \ldots, \nu_n$ belong to the Wasserstein space W_2 of probability measures over $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, with finite second order moment, endowed with the Wasserstein distance d_{W_2}, associated with the quadratic cost; see [Villani, 2003] for information on the Wasserstein space. In this setting, it is not possible to define a notion of PCA in the usual sense as the Wasserstein space W_2 is not linear. Nevertheless, in this chapter we show how to define a proper notion of PCA in W_2, relying on the formal Riemannian structure of W_2, described in Section 2.2.3. A first idea in that direction is related to the mean of the data, which is an essential ingredient in any notion of PCA. We propose to use the Fréchet mean (also called barycenter) as introduced in [Agueh and Carlier, 2011], whose asymptotic properties have been studied in [Bigot and Klein, 2012]. It is significant to see that the barycenter of $f_1, \ldots, f_4$, has a shape which is coherent with the shapes of the densities; see Figure 5(b).



(a) (b)

Figure 5: (a) Euclidean mean of the densities shown in Figure 1.(b) Density of the barycenter $\bar{\nu}_n$ in the Wasserstein space W_2 of the probability measures $\nu_1, \ldots, \nu_4$ admitting $f_1, \ldots, f_4$ as densities.

2.1.2 Main contributions and organization of the chapter

Since the Wasserstein space W_2 is not a Riemannian manifold, one cannot directly use the methods described in Section 1.2. Nevertheless, the space W_2 has a formal Riemannian structure [Ambrosio et al., 2004; Ambrosio and Savaré, 2007] that we use to define a new notion of Geodesic PCA (GPCA) in W_2, similar to those described in Section 1.2. Another contribution of this chapter is a proof of consistency of the empirical GPGA performed on a random data set.

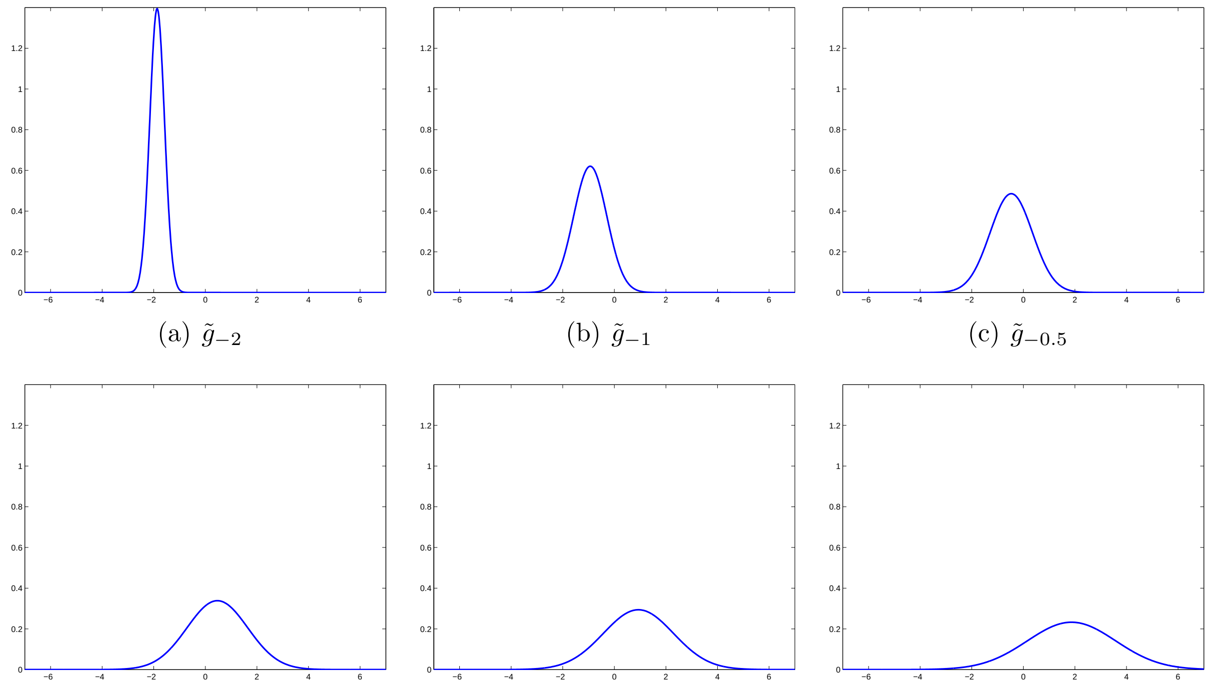
Before precisely defining GPCA, we display in Figure 6 the first principal mode of geodesic variation $\tilde{g}$ in W_2, of the data displayed in Figure 1; see equation (2.26). GPCA clearly gives a better description of the variability in the data compared to the results displayed in Figure 2, that correspond to the first principal mode of linear variation g in $L^2(\mathbb{R})$, as defined in (1.2).

The chapter is organized as follows. In Section 2.2 we give a precise definition of the Wasserstein space W_2 and we describe its formal Riemannian structure. In particular, we recall basic definitions such as tangent space, geodesics, exponential map and logarithmic map in the Wasserstein space framework, having their analogs in the Riemannian manifold setting. The main results are contained in sections 2.3 and 2.5. Section 2.3 is devoted to the construction and existence of GPCA in two variants, called global and nested, while Section 2.5 is dedicated to the consistency of GPCA, when the number n of random data points tends to infinity. Finally, we conclude the chapter in Section 2.6 with a discussion of the differences between GPCA and the methods described in Section 1.2. In order to be self-contained, we give in Appendix A some classical results on quantile functions, geodesic spaces, Kuratowski convergence and Γ-convergence. Throughout the chapter, we provide numerical experiments to illustrate the properties of GPCA in simple statistical models.

2.2 Preliminaries and definitions

2.2.1 A running example: the homothetic model

We begin this section by describing a simple construction of probabilities on the line, that will be used throughout the chapter as illustrative example. Let $\mu_0 \in W_2$ with density f_0



(a) $\tilde{g}_{-2}$ (b) $\tilde{g}_{-1}$ (c) $\tilde{g}_{-0.5}$

(d) $\tilde{g}_{0.5}$ (e) $\tilde{g}_{1}$ (f) $\tilde{g}_{2}$

Figure 6: An example of GPCA of densities. First principal mode of geodesic variation $\tilde{g}_t$ in the Wasserstein space W_2 for $-2 \leq t \leq 2$ of the densities displayed in Figure 1. See equation (2.26).

and cdf F_0. For $(a, b) \in (0, \infty) \times \mathbb{R}$, let us denote by $\nu^{(a,b)}$ the probability the density

$$f^{(a,b)}(x) := \frac{1}{a} f_0 \left(\frac{x-b}{a} \right), \; x \in \mathbb{R}. \tag{2.1}$$

and cdf

$$F^{(a,b)}(x) := F_0 \left(\frac{x-b}{a} \right), \; x \in \mathbb{R}. \tag{2.2}$$

Then, if $(a_1, b_1), \ldots, (a_n, b_n)$ belong to $(0, \infty) \times \mathbb{R}$, we consider as a running example the case where the data $\nu_1, \ldots, \nu_n$ satisfy the model

$$\nu_i = \nu^{(a_i, b_i)}, \; i = 1, \ldots, n, \tag{2.3}$$

to be referred as the *homothetic model* of measures in the rest of the chapter. The densities displayed in Figure 1 represent an example of realizations of this model with f_0 being the standard normal density and $n = 4$.

2.2.2 Optimal transportation theory

Let Ω denote either the real line $\mathbb{R}$ or an interval of $\mathbb{R}$. We denote by $W_2(\Omega)$ the set of probability measures over $(\Omega, \mathcal{B}(\Omega))$ with finite second order moment, where $\mathcal{B}(\Omega)$ is the sigma field of Borel subsets of Ω; if $\Omega = \mathbb{R}$, we write W_2 instead of $W_2(\mathbb{R})$. We also denote by

$W_2^{ac}(\Omega) \subset W_2(\Omega)$ the set of probability measures that are absolutely continuous with respect to the Lebesgue measure dx on Ω. For a measurable map $T : \Omega \to \Omega$, we recall that the push-forward measure $T\#\mu$ of $\mu \in W_2(\Omega)$ through the map T is defined as

$$(T\#\mu)(A) = \mu(T^{-1}(A)) \text{ for all } A \in \mathcal{B}(\Omega).$$

Definition 2.1. *Given $\mu, \nu \in W_2(\Omega)$, the Wasserstein distance between μ and ν is defined by*

$$d_{W_2}(\mu, \nu) := \left(\inf_{\pi \in \mathcal{P}(\mu,\nu)} \int_{\Omega \times \Omega} |x - y|^2 d\pi(x, y) \right)^{1/2},$$

where $\mathcal{P}(\mu, \nu)$ is the set of probability measures on $\Omega \times \Omega$ with marginals μ and ν.

It can be shown that $W_2(\Omega)$ endowed with the distance d_{W_2} is a metric space, the so-called Wasserstein space. For a detailed analysis of its properties, we refer to [Villani, 2003]. Let us now recall the following key theorem from optimal transportation theory, due to Brenier [Brenier, 1991].

Theorem 2.1. *Let $\mu \in W_2^{ac}(\Omega)$ and $\nu \in W_2(\Omega)$, then*

$$d^2_{W_2}(\mu, \nu) = \inf_{T \in \mathrm{MP}(\mu,\nu)} \int_\Omega |T(x) - x|^2 d\mu(x), \tag{2.4}$$

where $\mathrm{MP}(\mu, \nu) = \{T : \Omega \to \Omega \mid T$ is measurable and $\nu = T\#\mu\}$. Moreover, there exists $T^* \in \mathrm{MP}(\mu, \nu)$ such that $d^2_{W_2}(\mu, \nu) = \int_\Omega |T^*(x) - x|^2 d\mu$, characterized as the unique (up to a μ negligible set) element in $\mathrm{MP}(\mu, \nu)$ that can be represented as the gradient of a convex function.

Since we are in dimension one, a mapping $T : \Omega \to \Omega$ is the gradient of a convex function if and only if T is increasing. Therefore, if F and G are the cumulative distribution functions (cdf) of μ and ν, then the optimal mapping in Theorem 2.1 is given by $T^* = G^- \circ F$ and

$$d^2_{W_2}(\mu, \nu) = \int_\Omega (G^- \circ F(x) - x)^2 d\mu(x), \tag{2.5}$$

where $G^-(y) := \inf\{x \in \mathbb{R} : G(x) \geq y\}$ is the quantile function of ν. Note that G^- is increasing, hence almost everywhere differentiable. By the change of variable $x = F^-(y)$, we obtain the well-known characterization of the Wasserstein distance via the quantile functions of μ and ν:

$$d^2_{W_2}(\mu, \nu) = \int_0^1 (G^-(y) - F^-(y))^2 dy. \tag{2.6}$$

2.2.3 The pseudo-Riemannian structure of the Wasserstein space

The Wasserstein space over $(\Omega, \mathcal{B}(\Omega))$ has a formal Riemannian structure considered, for example, in [Ambrosio et al., 2004; Ambrosio and Savaré, 2007]. We first provide some basic definitions in the Wasserstein space framework, having their analogs in the Riemannian manifold setting. Following [Ambrosio et al., 2004; Ambrosio and Savaré, 2007], given $\mu \in W_2^{ac}(\Omega)$ we define the tangent space at μ as the Hilbert space $L^2_\mu(\Omega)$ of real-valued, μ-square-integrable functions on Ω, equipped with the standard inner product and its associated norm denoted by $\|\cdot\|_\mu$. Furthermore, we define the exponential and the logarithmic maps at μ as follows.

Definition 2.2. *Let $\mu \in W_2^{ac}(\Omega)$ be a reference measure. The exponential map $\exp_\mu : L^2_\mu(\Omega) \to W_2(\Omega)$ and the logarithmic map $\log_\mu : W_2(\Omega) \to L^2_\mu(\Omega)$ are defined respectively as*

$$\exp_\mu(v) = (\mathrm{id} + v)\#\mu \quad \text{and} \quad \log_\mu(\nu) = G^- \circ F - \mathrm{id}, \tag{2.7}$$

where F, G are the cdf of μ, ν respectively, id *is the identity function on Ω, and $(\mathrm{id} + v)\#\mu$ is the push-forward measure of μ through the map $T = \mathrm{id} + v$.*

Note that the exponential map is well defined in the sense that $\exp_\mu(v) \in W_2(\Omega)$ for any $v \in L^2_\mu(\Omega)$, since

$$\int_\Omega x^2 d\ \exp_\mu(v)(x) = \int_\Omega (x + v(x))^2 d\mu(x) \leq 2 \int_\Omega x^2 d\mu(x) + \int_\Omega v^2(x) d\mu(x) < +\infty.$$

Similarly, the logarithmic map is well defined in the sense that $\log_\mu(\nu) \in L^2_\mu(\Omega)$ since $\| \log_\mu(\nu)\|^2_\mu = d^2_{W_2}(\mu, \nu) < +\infty$, for all $\nu \in W_2(\Omega)$.

Example 2.1. *In order to illustrate the notions of exponential and logarithmic maps, we consider the homothetic model (2.3). Let $\mu \in W_2^{ac}$ be a reference measure with cdf F. Then*

$$\log_\mu(\nu^{(a,b)})(x) = [F^{(a,b)}]^- \circ F(x) - x, \ x \in \mathbb{R}.$$

Taking $\mu = \mu_0$, from (2.2), the expression above simplifies to

$$\log_{\mu_0}(\nu^{(a,b)})(x) = (a-1)x + b, \ x \in \mathbb{R}.$$

Therefore, letting $v(x) = (a-1)x + b, \ x \in \Omega$, then

$$\exp_{\mu_0}(v) = \nu^{(a,b)}. \tag{2.8}$$

In the setting of Riemannian manifolds, the exponential map at a given point is a local homeomorphism from a neighborhood of the origin in the tangent space onto the manifold. However, this is not the case for the exponential map $\exp_\mu$ defined above, as it is possible to find two arbitrarily small functions in $L^2_\mu(\Omega)$ with equal exponentials, see e.g. [Ambrosio et al., 2004; Ambrosio and Savaré, 2007]. On the other hand, we will show that $\exp_\mu$ is an isometry when restricted to a specific set of admissible functions defined below.

Definition 2.3. *Let $\mu \in W_2^{ac}(\Omega)$. The function $v \in L^2_\mu(\Omega)$ is said to be admissible if* $\mathrm{id} + v$ *is increasing μ-a.s. It is said to be strictly admissible if* $\mathrm{id} + v$ *is strictly increasing μ-a.s. The set of admissible (resp. strictly) functions is denoted by V_μ (resp. V_μ^s).*

We remark that V_μ is not a linear space.

Lemma 2.1. *Let (T_k) be a sequence in $L^2_\mu(\Omega)$ such that T_k is μ-a.s increasing , for all $k \geq 1$. (a) If (T_k) converges pointwise μ-a.s to $T \in L^2_\mu(\Omega)$ then T is μ-a.s increasing. (b) If (T_k) converges in norm to $T \in L^2_\mu(\Omega)$ then T is μ-a.s increasing.*

Proof. (a) It is possible to find a measurable set A, such that $\mu(A) = 1$, T_k is increasing in A, for all $k \geq 1$ and $T_k \to T$ pointwise in A. Then it directly follows that T is increasing in A.
(b) As (T_k) converges in norm to T and using a result from measure theory, we know there exists a subsequence (T_{k_j}) converging μ-a.s to T. Then the conclusion follows from (a). □

The following result is a key ingredient for the construction of GPCA in the Wasserstein space.

Proposition 2.1. *Let $\mu \in W_2^{ac}(\Omega)$, then V_μ is closed and convex in $L^2_\mu(\Omega)$. Moreover V_μ^s is convex and dense in V_μ.*

Proof. For convexity take $\lambda \in [0,1]$, $v_1, v_2 \in V_\mu$. Then, from Definition 2.3, $\mathrm{id} + v_1$ and $\mathrm{id} + v_2$ are increasing and so is $\mathrm{id} + \lambda v_1 + (1-\lambda)v_2 = \lambda(\mathrm{id} + v_1) + (1-\lambda)(\mathrm{id} + v_2)$. The same arguments apply for the convexity of V_μ^s. For closeness, take a sequence (v_k) in V_μ converging to $v \in L^2_\mu(\Omega)$. Then $\mathrm{id} + v_k \to \mathrm{id} + v$ and, by Lemma 2.1, $\mathrm{id} + v$ is increasing, that is $v \in V_\mu$. Finally, we prove that V_μ^s is dense in V_μ by taking a sequence (h_k) in V_μ^s such that $h_k \to 0$. Then, for $v \in V_\mu$, we have $v + h_k \in V_\mu^s$ and $v + h_k \to v$, so the conclusion follows. $\square$

The following proposition shows that the exponential map in $W_2(\Omega)$, restricted to the convex set of admissible functions V_μ, is an isometry. This is an important property that will allow us to define and to compute the GPCA in $W_2(\Omega)$.

Proposition 2.2. *Let $\mu \in W_2^{ac}(\Omega)$, then $\exp_\mu(V_\mu) = W_2(\Omega)$ and $\exp_\mu(V_\mu^s) = W_2^{ac}(\Omega)$. Also, the exponential map $\exp_\mu$ restricted to V_μ or V_μ^s is an isometric homeomorphism, with inverse given by $\log_\mu$.*

Proof. Let $\nu \in W_2(\Omega)$ then, from Theorem 2.1, $G^- \circ F$ is the unique μ-a.s. increasing map such that $(G^- \circ F)\#\mu = \nu$, where F and G are the cdf of μ and ν respectively. In other words, $v := \log_\mu(\nu) = G^- \circ F - \mathrm{id}$ is the unique element in V_μ such that $\exp_\mu(v) = \nu$. Now, take $\nu \in W_2^{ac}(\Omega)$ and observe that, since ν is absolutely continuous with respect to the Lebesgue measure, G^- is strictly increasing (see for instance [Embrechts and Hofert, 2013]). As F is strictly increasing μ-a.s, we have that $G^- \circ F$ is strictly increasing μ-a.s. and so, $v \in V_\mu^s$.

Let us now prove the isometry property. From (2.6) and a change of variable we obtain $d^2_{W_2}(\nu_1, \nu_2) = \| \log_\mu(\nu_1) - \log_\mu(\nu_2)\|^2_\mu$, for any $\nu_1, \nu_2 \in W_2^{ac}(\Omega)$, and so $\exp_\mu : V_\mu^s \to W_2^{ac}(\Omega)$ is an isometry. Finally, thanks to Proposition 2.1, V_μ^s is dense in V_μ, which implies that $\exp_\mu : V_\mu \to W_2(\Omega)$ is an isometry as well. $\square$

2.2.4 Geodesics in the Wasserstein space

A general overview of geodesics in a metric space is given in Appendix A.2. In this section, we consider the notion of geodesic in the metric space $W_2(\Omega)$ as given in Definition A.4. A direct consequence of Corollary A.1 and Proposition 2.2 is that geodesics in $W_2(\Omega)$ are exactly the image under $\exp_\mu$ of straight lines in V_μ, where $\mu \in W_2^{ac}(\Omega)$ is a reference measure. In particular, given two measures in $W_2(\Omega)$, there exists a unique shortest path connecting them. This property is stated in the following lemma.

Lemma 2.2. *Let $\mu \in W_2^{ac}(\Omega)$ and $\gamma : [0,1] \to W_2(\Omega)$ be a curve, where $v_0 := \log_\mu(\gamma(0))$ and $v_1 := \log_\mu(\gamma(1))$. Then γ is a geodesic if and only if $\gamma(t) = \exp_\mu((1-t)v_0 + tv_1)$, for all $t \in [0,1]$.*

Example 2.2. *To illustrate Lemma 2.2, let us consider again the homothetic model* (2.3) *and take $\mu = \mu_0$ as reference measure. If we let $\nu_0 := \mu_0$ and $\nu_1 = \nu^{(a,b)}$, with $a > 0$, $b \in \mathbb{R}$,* *then one has*

$$v_0(x) := \log_{\mu_0}(\nu_0) = 0 \text{ and } v_1(x) := \log_{\mu_0}(\nu_1) = (a-1)x + b, \; x \in \mathbb{R}.$$

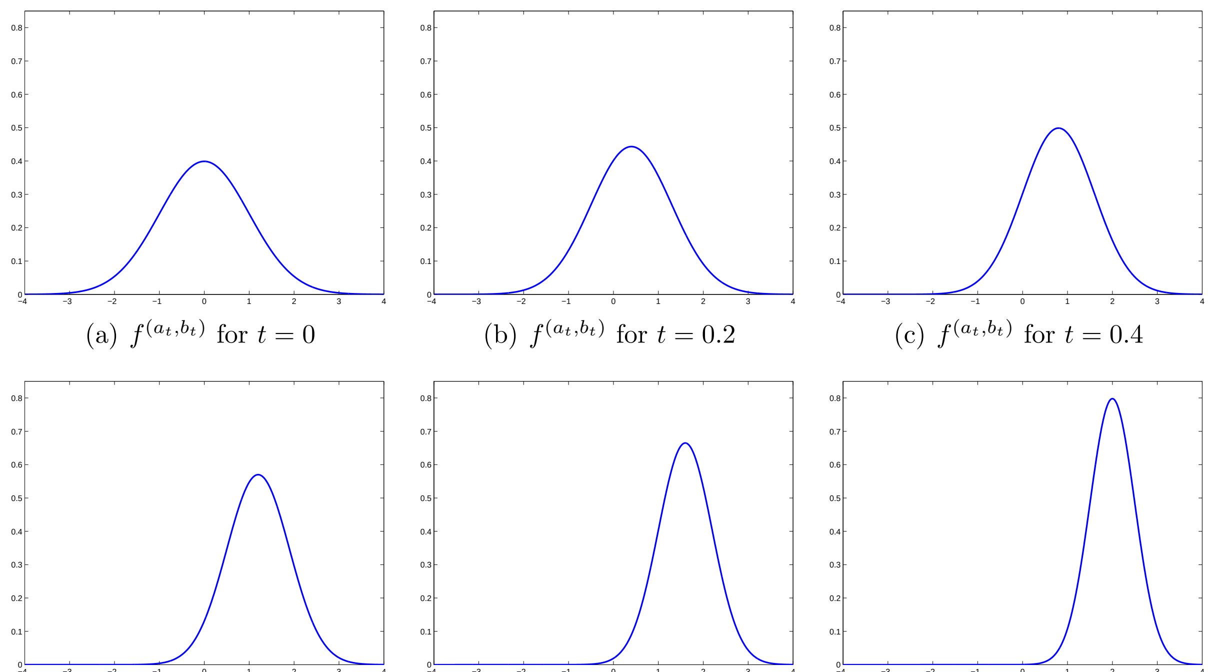
From (2.8) and Lemma 2.2, the curve $\gamma : [0,1] \to W_2(\Omega)$, defined by

$$\gamma(t) = \exp_{\mu_0}((1-t)v_0 + tv_1) = \exp_{\mu_0}(t(a-1)x + tb) = \nu^{(a_t,b_t)}, \; t \in [0,1],$$

is a geodesic such that $\gamma(0) = \mu_0 = \nu^{(1,0)}$ and $\gamma(1) = \nu^{(a,b)}$, where $a_t = 1 - t + ta$ and $b_t = tb$. Moreover, for each time $t \in [0,1]$, the measure $\gamma(t)$ admits the density

$$f^{(a_t,b_t)}(x) = \frac{1}{a_t} f_0\left(\frac{x - b_t}{a_t}\right), \; x \in \mathbb{R}. \tag{2.9}$$

In Figure 7, we display the densities $f^{(a_t,b_t)}$, for some values of $t \in [0,1]$, in the case where μ_0 is the standard Gaussian measure, $a = 0.5$ and $b = 2$.



(a) $f^{(a_t,b_t)}$ for $t = 0$ (b) $f^{(a_t,b_t)}$ for $t = 0.2$ (c) $f^{(a_t,b_t)}$ for $t = 0.4$

(d) $f^{(a_t,b_t)}$ for $t = 0.6$ (e) $f^{(a_t,b_t)}$ for $t = 0.8$ (f) $f^{(a_t,b_t)}$ for $t = 1$

Figure 7: Visualization of the densities $f^{(a_t,b_t)}$ associated with the geodesic curve $\gamma(t) = \nu^{(a_t,b_t)}$ in W_2, described in Example 2.2, with $a = 0.5$ and $b = 2$, in the case where $\mu = \mu_0$ is the standard Gaussian measure.

Note that, by Lemma 2.2, $W_2(\Omega)$ endowed with the Wasserstein distance d_{W_2} is a geodesic space. Moreover, we immediately have the following corollary.

Corollary 2.1. *Let $\mu \in W_2^{ac}(\Omega)$ be a reference measure. Then $G \subset W_2(\Omega)$ is geodesic (in the sense of Definition A.5) if and only if $\log_\mu(G)$ is convex.*

Finally, to define the GPCA in $W_2(\Omega)$ we will need the following definition of the dimension of a geodesic subset.

Definition 2.4. *Let G be a geodesic subset of $W_2(\Omega)$. We define the dimension of G, denoted* $\dim(G)$*, as the dimension of the smallest affine subspace of $L^2_\mu(\Omega)$ containing* $\log_\mu(G)$.

We observe that the previous definition does not depend on the choice of the measure μ. Indeed take $\mu' \in W_2(\Omega)$ and E an affine subspace of $L^2_\mu(\Omega)$ containing $\log_\mu(G)$. It is easy to see that $T_{\mu,\mu'} := \log_{\mu'} \circ \exp_\mu$ is an affine function from $L^2_\mu(\Omega)$ to $L^2_{\mu'}(\Omega)$, therefore $T_{\mu,\mu'}(E)$ is an affine subspace of $L^2_{\mu'}(\Omega)$ containing $\log_{\mu'}(G)$ and $\dim(E) = \dim(T_{\mu,\mu'}(E))$. Observe also that, if $\gamma : [0,1] \to W_2(\Omega)$ is a geodesic, then $\gamma([0,1])$ is a geodesic space of dimension 1.

2.3 Population Fréchet mean and principal geodesics

Throughout this section we consider a reference measure $\mu \in W_2^{ac}(\Omega)$ and a $W_2(\Omega)$-valued random element $\boldsymbol{\nu}$ which we assume to be square-integrable, in the sense of the following definition:

Definition 2.5. *A random measure $\boldsymbol{\nu} \in W_2(\Omega)$ is said to be square-integrable if*

$$\mathbb{E}(d^2_{W_2}(\boldsymbol{\nu}, \mu')) < +\infty,$$

for some (thus for all) $\mu' \in W_2(\Omega)$.

From now on all random measure $\boldsymbol{\nu}$ in $W_2(\Omega)$ are assumed to be square-integrable, according to the previous definition. Also, we consider $W_2(\Omega)$ equipped with the Borel sigma field relative to the Wasserstein metric.

2.3.1 Fréchet Mean

A natural notion of average in $W_2(\Omega)$ is the Fréchet mean, studied in [Bigot and Klein, 2012] in a general setting, for random measures with support in $\mathbb{R}^d$, $d \geq 1$. In what follows we define and give some properties of the population Fréchet mean $\mathcal{M}(\boldsymbol{\nu})$ of a random measure $\boldsymbol{\nu} \in W_2(\Omega)$. Note that results are stated for the one-dimensional case, that is, $d = 1$. The higher dimensional case is much more involved and we refer to [Agueh and Carlier, 2011; Bigot and Klein, 2012] for further details.

Let $\boldsymbol{u}$ be a random element in $L^2_\mu(\Omega)$, such that $\mathbb{E}(\|\boldsymbol{u}\|_\mu) < +\infty$. Then the expectation $\mathbb{E}(\boldsymbol{u})$ of $\boldsymbol{u}$ is defined by $\mathbb{E}(\boldsymbol{u})(x) = \mathbb{E}(\boldsymbol{u}(x))$, for all $x \in \mathbb{R}$. Observe that $\|\mathbb{E}(\boldsymbol{u})\|_\mu \leq \mathbb{E}(\|\boldsymbol{u}\|_\mu) < \infty$, hence $\mathbb{E}(\boldsymbol{u}) \in L^2_\mu(\Omega)$. Also, if $\mathbb{P}(\boldsymbol{u} \in V_\mu) = 1$, then clearly $\mathbb{E}(\boldsymbol{u}) \in V_\mu$ and if $\mathbb{P}(\boldsymbol{u} \in V_\mu^s) = 1$, then $\mathbb{E}(\boldsymbol{u}) \in V_\mu^s$. We refer to [Carrasco, 2005] and Chapter 8 of [Bonaccorsi and Priola, 2006] for a detailed exposition on probabilities in Banach and Hilbert spaces.

Proposition 2.3. *Let $\boldsymbol{\nu} \in W_2(\Omega)$ be a random measure. Then, one has the following.*

(i) There exists a unique element $\mathcal{M}(\boldsymbol{\nu}) \in \arg\min_{\pi \in W_2(\Omega)} \mathbb{E}\left(d^2_{W_2}(\boldsymbol{\nu}, \pi)\right)$ that we define as the (population) Fréchet mean of $\boldsymbol{\nu}$.

(ii) $\mathcal{M}(\boldsymbol{\nu}) = \exp_\mu(\mathbb{E}(\boldsymbol{v}))$, for any $\mu \in W_2^{ac}(\Omega)$, where $\boldsymbol{v} = \log_\mu(\boldsymbol{\nu})$.

(iii) $\mathcal{M}(\boldsymbol{\nu})$ has cdf $\bar{G}$ that satisfies $\bar{G}^- = \mathbb{E}(\boldsymbol{G}^-)$, where $\boldsymbol{G}$ is the cdf of $\boldsymbol{\nu}$.

(iv) If $\mathbb{P}(\boldsymbol{\nu} \in W_2^{ac}(\Omega)) = 1$, then $\mathcal{M}(\boldsymbol{\nu}) \in W_2^{ac}(\Omega)$.

Proof. *(i),(ii) Let $\mu \in W_2^{ac}(\Omega)$, and let $\boldsymbol{v} = \log_\mu(\boldsymbol{\nu})$. From Proposition 2.2 and as $\boldsymbol{\nu}$ is square-integrable, we have that $\mathbb{E}\left(\|\boldsymbol{v} - u\|_\mu^2\right) = \mathbb{E}\left(d_{W_2}^2(\boldsymbol{\nu}, \exp_\mu(u))\right) < \infty$, for all $u \in L_\mu^2(\Omega)$. Therefore*

$$\inf_{\pi \in W_2(\Omega)} \mathbb{E}\left(d_{W_2}^2(\boldsymbol{\nu}, \pi)\right) = \inf_{u \in V_\mu} \mathbb{E}\left(\|\boldsymbol{v} - u\|_\mu^2\right), \tag{2.10}$$

and u^* is a minimizer of the right-hand side (rhs) of (2.10) if and only if $\exp_\mu(u^*)$ is a minimizer of the left-hand side (lhs) of (2.10). On the other hand, observe that $\mathbb{E}(\boldsymbol{v})$ belongs to the convex set V_μ and that it is clearly the unique minimizer of $u \mapsto \mathbb{E}\left(\|\boldsymbol{v} - u\|_\mu^2\right)$ for $u \in V_\mu$. Hence $\mathbb{E}(\boldsymbol{v})$ is the unique minimizer of the rhs of (2.10) and thus, by Proposition 2.2, this implies that $\mathcal{M}(\boldsymbol{\nu}) = \exp_\mu(\mathbb{E}(\boldsymbol{v}))$ is the unique minimizer of the lhs of (2.10), for any $\mu \in W_2^{ac}(\Omega)$.

(iii) From Proposition A.1(vi), if $\mu \in W_2^{ac}(\Omega)$, with cdf F and if $T : \mathbb{R} \to \mathbb{R}$ is increasing and left-continuous, then the quantile of $T\#\mu$ is given by $G^- = T \circ F^-$. Therefore, by (ii) and Definition 2.2, we have

$$\bar{G}^- = (id + \mathbb{E}(\boldsymbol{v})) \circ F^- = \mathbb{E}(id + \log_\mu(\boldsymbol{\nu})) \circ F^- = \mathbb{E}(\boldsymbol{G}^- \circ F) \circ F^- = \mathbb{E}(\boldsymbol{G}^-),$$

where $\boldsymbol{v} = \log_\mu(\boldsymbol{\nu})$.

(iv) It is a direct consequence of (ii) and Proposition 2.2. □

One can remark that Proposition 2.3(ii) implies that $\exp_\mu(\mathbb{E}(\log_\mu(\boldsymbol{\nu})))$ does not depend on μ.

2.3.2 Principal geodesics

Let us first introduce some definitions and results that are needed to define the notion of principal geodesics of a random measure $\boldsymbol{\nu}$ in $W_2(\Omega)$.

Definition 2.6. *Let $G \subset W_2(\Omega)$ and $\nu \in W_2(\Omega)$. We define the distance between ν and G by*

$$d_{W_2}(\nu, G) = \inf_{\pi \in G} d_{W_2}(\nu, \pi),$$

Definition 2.7. *Let $\boldsymbol{\nu}$ be a random measure in $W_2(\Omega)$ and $G \subset W_2(\Omega)$. We associate to G the cost given by the expected value of the square residual of projecting $\boldsymbol{\nu}$ onto G, that is,*

$$\operatorname{cost}(G) := \mathbb{E}\left(d_{W_2}^2(\boldsymbol{\nu}, G)\right). \tag{2.11}$$

Note that $\operatorname{cost}(G)$ depends on the random measure $\boldsymbol{\nu}$ and that it is necessarily finite as $\boldsymbol{\nu}$ is assumed to be square-integrable. Observe also that cost is monotone, in the sense that

$$\operatorname{cost}(G) \geq \operatorname{cost}(F), \text{ if } G \subset F. \tag{2.12}$$

Definition 2.8. *Let CL be the metric space of nonempty and closed subsets of $W_2(\Omega)$, endowed with the Hausdorff distance h_{W_2} (see Definition A.7). Let also*

$$\mathrm{CG}_{\mu,k} = \{G \in \mathrm{CL} \mid \mu \in G,\ G \text{ is a geodesic set and } \dim(G) \leq k\},\ k \geq 1,$$

where $\dim(G)$ is defined in Definition 2.4.

In the rest of this Section 2.3.2, we concentrate on defining and showing the existence of principal geodesics under the assumption that Ω is compact. In this case it is well known that $(W_2(\Omega), d_{W_2})$ is also a compact metric space, see e.g. [Villani, 2003]. Note also that the compactness of $W_2(\Omega)$ implies that every random measure $\boldsymbol{\nu}$ in $W_2(\Omega)$ is square-integrable. Furthermore, compactness of $W_2(\Omega)$ also implies that $h_{W_2}(G_n, G) \to 0$, as $n \to \infty$ if and only if $G_n \to G$ in the sense of Kuratowski (see Definition A.6). From Proposition A.2 we obtain the following result.

Proposition 2.4. *For any $\nu \in W_2(\Omega)$, the function $G \in \mathrm{CL} \mapsto d_{W_2}(\nu, G)$ is continuous.*

Corollary 2.2. *The function* cost *of Definition 2.7 is continuous on* CL.

Proof. Let $\boldsymbol{\nu} \in W_2(\Omega)$ be a random measure and (G_n) be a sequence in CL converging to $G \in \mathrm{CL}$, in the sense of Hausdorff's distance. Take $\pi \in W_2(\Omega)$ and observe that

$$d^2_{W_2}(\boldsymbol{\nu}, G_n) \leq 2d^2_{W_2}(\boldsymbol{\nu}, \pi) + 2d^2_{W_2}(\pi, G_n).$$

Since $\boldsymbol{\nu}$ is square-integrable, one has that $\mathbb{E}(d^2_{W_2}(\boldsymbol{\nu}, \pi)) < \infty$ and, given that $W_2(\Omega)$ is compact, $d^2_{W_2}(\pi, G_n) \leq R$, for some constant $R > 0$. Hence, by Proposition 2.4 and the dominated convergence theorem, we conclude that $\mathrm{cost}(G_n) \to \mathrm{cost}(G)$. □

Proposition 2.5. *The sets* CL *and* $\mathrm{CG}_{\mu,k}$*, from Definition 2.8, are compact.*

Proof. Since the $\log_\mu$ map is a continuous bijection from the compact set $W_2(\Omega)$ onto V_μ (see Proposition 2.2), then V_μ is also compact. Hence, from Proposition A.3, $\mathrm{CL}(V_\mu) := \{C \subset V_\mu \mid C \text{ is nonempty and closed}\}$ and $\mathrm{CC}_{k,0}(V_\mu) := \{C \in \mathrm{CL}(V_\mu) \mid C \text{ is convex}, 0 \in C \text{ and } \dim(C) \leq k\}$ are compact metric spaces, for the topology induced by the Hausdorff distance h_μ of sets in $L^2_\mu(\Omega)$ (see Definition A.7). Now, consider the map

$$\begin{aligned} i_\mu : (\mathrm{CL}(V_\mu), h_\mu) &\to (\mathrm{CL}, h_{W_2}) \\ C &\to \exp_\mu(C) \end{aligned}$$

and observe that $i_\mu(\mathrm{CL}(V_\mu)) = \mathrm{CL}$, by Proposition 2.2, and $i_\mu(\mathrm{CG}_{k,0}(V_\mu)) = \mathrm{CG}_{\mu,k}$, by Corollary 2.1. On the other hand, as the exponential map is an isometry, it is easy to see that

$$h_\mu(C_1, C_2) = h_{W_2}(\exp_\mu(C_1), \exp_\mu(C_2)),\ C_1, C_2 \in \mathrm{CL}(V_\mu). \tag{2.13}$$

Hence i_μ is an isometric bijection and so, CL and $\mathrm{CG}_{\mu,k}$ are compact. □

Global principal geodesics

We now present a first definition of GPCA, namely the notion of global principal geodesic of a random measure $\boldsymbol{\nu}$, with respect to a reference measure $\mu \in W_2^{ac}(\Omega)$.

Definition 2.9. *Let $\mu \in W_2^{ac}(\Omega)$ be a reference measure, $\boldsymbol{\nu} \in W_2(\Omega)$ a random measure and $k \geq 1$ an integer. A global k-principal μ-geodesic of $\boldsymbol{\nu}$ is an element of*

$$\mathcal{G}_{\mu,k} := \arg\min\{\mathrm{cost}(G) \mid G \in \mathrm{CG}_{\mu,k}\}. \tag{2.14}$$

The following is the main result about the existence of global principal geodesics.

Theorem 2.2. *Let $\mu \in W_2^{ac}(\Omega)$ be a reference measure and $k \geq 1$ an integer. Then $\mathcal{G}_{\mu,k}$, from Definition 2.9, is nonempty.*

Proof. The result is a direct consequence of Corollary 2.2 and Proposition 2.5. □

Nested principal geodesics

Here we inductively define a second variant of GPCA that we call nested principal geodesics. This approach is inspired by the usual characterization of PCA in terms of a nested sequence of optimal linear subspaces.

Definition 2.10. *Let $\mu \in W_2^{ac}(\Omega)$ be a reference measure, $\boldsymbol{\nu} \in W_2(\Omega)$ a random measure and $k \geq 1$ an integer. A nested k-principal μ-geodesic of $\boldsymbol{\nu}$ is a k-tuple $(G_1, \dots, G_k)$ such that*

$$G_1 \in \arg\min\{\mathrm{cost}(G) \mid G \in \mathrm{CG}_{\mu,1}\}$$

and for $j = 2, \dots, k$,

$$G_j \in \arg\min\{\mathrm{cost}(G) \mid G \in \mathrm{CG}_{\mu,j}, G \supset G_{j-1}\}.$$

The set of nested k-principal μ-geodesics of $\boldsymbol{\nu}$ is denoted by $\mathcal{N}_{\mu,k}$.

Theorem 2.3. *Let $\mu \in W_2^{ac}(\Omega)$ be a reference measure and $k \geq 1$ an integer. Then $\mathcal{N}_{\mu,k}$, from Definition 2.10, is nonempty.*

Proof. Let $F, G_n \in \mathrm{CL}$ such that $F \subset G_n, n \geq 1$. If (G_n) converges to $G \in \mathrm{CL}$ then, thanks to (ii) in Definition A.6, $F \subset G$. Therefore, the set $\{G \in \mathrm{CL} : G \supset F\}$ is closed and so $\{G \in \mathrm{CG}_{\mu,k} : G \supset F\}$ is compact, thanks to Proposition 2.5. Thus, from Corollary 2.2,

$$\arg\min\{\mathrm{cost}(G) \mid G \in \mathrm{CG}_{\mu,k}, G \supset F\} \neq \emptyset.$$

Let us show that $\mathcal{N}_{\mu,k}$ is nonempty by induction on k. First observe that $\mathcal{N}_{\mu,1} = \mathcal{G}_{\mu,1}$; then, from Theorem 2.2, we have $\mathcal{N}_{\mu,1} \neq \emptyset$. Assume $\mathcal{N}_{\mu,k-1} \neq \emptyset$, let $(G_1, \dots, G_{k-1}) \in \mathcal{N}_{\mu,k-1}$ and define G_k as an element in $\arg\min\{\mathrm{cost}(G) \mid G \in \mathrm{CG}_{\mu,k}, G \supset G_{k-1}\}$, which was shown above to be nonempty. Finally $(G_1, \dots, G_{k-1}, G_k) \in \mathcal{N}_{\mu,k}$ by definition. □

Remark 2.1. *(1) For $k = 1$, the notions of global and nested principal geodesics coincide. However, this might be not the case for $k \geq 2$.*
(2) Our proofs of existence of principal geodesics, in Theorems 2.2 and 2.3, rely on the assumption that Ω is compact. However, this assumption is not a necessary condition as seen in Theorem 2.6, where we prove existence of principal geodesic in the case $\Omega = \mathbb{R}$ and $k = 1$. However, the non-compactness of Ω significantly complicates our proofs.

2.3.3 Empirical Fréchet mean and principal geodesics

We define the empirical Fréchet mean of measures $\nu_1, \dots, \nu_n \in W_2(\Omega)$ as follows.

Definition 2.11. *An empirical Fréchet mean of $\nu_1, \ldots, \nu_n \in W_2(\Omega)$ is defined as an element of*

$$\arg\min_{\nu \in W_2(\Omega)} \frac{1}{n} \sum_{i=1}^{n} d_{W_2}^2(\nu_i, \nu).$$

Definition 2.12. *Given $\nu_1, \ldots, \nu_n \in W_2(\Omega)$, we denote by $\boldsymbol{\nu}^{(n)}$ the random measure in $W_2(\Omega)$ such that $\mathbb{P}(\boldsymbol{\nu}^{(n)} = \nu_i) = 1/n$, for $i = 1, \ldots, n$.*

It is clear that $\boldsymbol{\nu}^{(n)}$, defined above, is square-integrable, according to Definition 2.5.

Proposition 2.6. *For any $\nu_1, \ldots, \nu_n \in W_2(\Omega)$ there exists a unique empirical Fréchet mean, denoted by $\bar{\nu}_n$. Furthermore,*

$$\bar{G}_n^- = \frac{1}{n} \sum_{i=1}^{n} G_i^-, \tag{2.15}$$

where $\bar{G}_n$ the cdf of $\bar{\nu}_n$ and $G_1, \ldots, G_n$ are the cdf of $\nu_1, \ldots, \nu_n$ respectively.

Proof. First observe that the empirical Fréchet mean can be characterized as the Fréchet mean (Proposition 2.3 (i)) of the random measure $\boldsymbol{\nu}^{(n)}$ (see Definition 2.12), that is, $\bar{\nu}_n = \mathcal{M}(\boldsymbol{\nu}^{(n)})$. Hence, the conclusions follow from Proposition 2.3 (i), (ii). □

Observe that (2.15) is the well known formula of quantile averaging in statistics, see e.g. [Zhang and Müller, 2011; Gallón et al., 2013]. A detailed characterization of the empirical Fréchet mean in the Wasserstein space, can be found in [Agueh and Carlier, 2011], for the general case of measures supported on $\mathbb{R}^d$, $d \geq 1$.

Definition 2.13. *Let $\mu \in W_2^{ac}(\Omega)$ be a reference measure, $\nu_1, \ldots, \nu_n \in W_2(\Omega)$ and $k \geq 1$ an integer. The empirical global and nested k-principal μ-geodesics of $\nu_1, \ldots, \nu_n$ are defined as in Definitions 2.9 and 2.10 respectively, applied to the random measure $\boldsymbol{\nu}^{(n)}$, given in Definition 2.12. The sets of empirical global and nested k-principal μ-geodesics are denoted by $\mathcal{G}_{\mu,k,n}$ and $\mathcal{N}_{\mu,k,n}$ respectively.*

Note that Theorems 2.2 and 2.3 yield the existence of empirical principal geodesics. Observe also that cost in (2.11) can be written in this empirical setting as

$$\mathrm{cost}_n(G) := \mathbb{E}\left(d_{W_2}^2(\boldsymbol{\nu}^{(n)}, G)\right) = \frac{1}{n} \sum_{i=1}^{n} d_{W_2}^2(\nu_i, G). \tag{2.16}$$

In this section we are assuming that the reference measure $\mu \in W_2^{ac}(\Omega)$ is arbitrary. However, by analogy with the PCA in Hilbert spaces, a natural choice for μ would be the Fréchet mean $\mathcal{M}(\boldsymbol{\nu})$, which belongs to $W_2^{ac}(\Omega)$ if $\mathbb{P}(\boldsymbol{\nu} \in W_2^{ac}(\Omega)) = 1$, thanks to Proposition 2.3(*iv*).

2.3.4 Formulation of GPCA as an optimization problem in $L^2_\mu(\Omega)$

In this section we formulate the empirical GPCA, as an optimization problem in $L^2_\mu(\Omega)$. Then, in the next section, we use this formulation to compute principal geodesics in the case of the homothetic model. First we introduce some notation to be used throughout this section. For $\mathcal{U} = \{u_1, \ldots, u_k\} \subset L^2_\mu(\Omega)$, $k \geq 1$, we denote by $\operatorname{span}(\mathcal{U})$ the subspace spanned by $u_1, \ldots, u_k$, by $\Pi_{\operatorname{span}(\mathcal{U})} v$ the projection of $v \in L^2_\mu(\Omega)$ onto $\operatorname{span}(\mathcal{U})$ and by $\Pi_{\operatorname{span}(\mathcal{U}) \cap V_\mu} v$ the projection of v onto the closed convex set $\operatorname{span}(\mathcal{U}) \cap V_\mu$.

Definition 2.14. *Let $\boldsymbol{v}$ be a random element in $L^2_\mu(\Omega)$ such that $\mathbb{E}(\|\boldsymbol{v}\|^2) < \infty$ and let $\mathcal{U} = \{u_1, \ldots, u_k\} \subset L^2_\mu(\Omega)$, $k \geq 1$. We associate to $\mathcal{U}$ the cost given by the expected value of the square residual of projecting $\boldsymbol{\nu}$ onto* $\operatorname{span}(\mathcal{U}) \cap V_\mu$*, that is,*

$$\operatorname{cost}^*(\mathcal{U}) := \mathbb{E}\left(\|\boldsymbol{v} - \Pi_{\operatorname{span}(\mathcal{U}) \cap V_\mu} \boldsymbol{v}\|^2_\mu\right). \tag{2.17}$$

Note that cost^*, which depends on the random element $\boldsymbol{v}$, is necessarily finite as $\|\mathbb{E}(\boldsymbol{v})\|^2 \leq \mathbb{E}(\|\boldsymbol{v}\|^2) < \infty$.

Definition 2.15. *Let $\mu \in W_2^{ac}(\Omega)$ be a reference measure and $\mathcal{U} = \{u_1, \ldots, u_k\} \subset L^2_\mu(\Omega)$, $k \geq 1$. We define the geodesic set generated by $\mathcal{U}$ as*

$$G_{\mathcal{U}} := \exp_\mu(\operatorname{span}(\mathcal{U}) \cap V_\mu).$$

Observe that, by Corollary 2.1, $G_{\mathcal{U}}$ is indeed a geodesic set in $W_2(\Omega)$. In order to simplify the notation, we write $\operatorname{span}(u)$, $\operatorname{cost}^*(u)$ or G_u, in the definitions above, if $\mathcal{U} = \{u\}$.

The following proposition shows that the problem of finding global k-principal μ-geodesics, (see Definition 2.9), can be formulated as an optimization problem in $(L^2_\mu(\Omega))^k$.

Proposition 2.7. *Let $\mu \in W_2^{ac}(\Omega)$ be a reference measure, $\boldsymbol{\nu} \in W_2(\Omega)$ a random measure, $\boldsymbol{v} = \log_\mu(\boldsymbol{\nu})$ and $k \geq 1$ an integer. Let $\mathcal{U}^* = \{u_1^*, \ldots, u_k^*\}$ be a minimizer of* cost^**, given by* (2.17)*, over orthonormal sets $\mathcal{U} = \{u_1, \ldots, u_k\} \subset L^2_\mu(\Omega)$, then $G_{\mathcal{U}^*} \in \mathcal{G}_{\mu,k}$.*

Proof. Recall that, by Corollary 2.1, geodesic sets in $W_2(\Omega)$ correspond to the image under $\exp_\mu$ of convex sets in V_μ. Therefore, as $\exp_\mu$ is an homeomorphism, we have

$$G_{\mathcal{U}^*} \in \mathrm{CG}_{\mu,k}. \tag{2.18}$$

On the other hand, from (2.11), (2.17) and as $\exp_\mu$ is an isometry (see Proposition 2.2), we have

$$\operatorname{cost}^*(\mathcal{U}) = \operatorname{cost}(G_{\mathcal{U}}), \quad \mathcal{U} = \{u_1, \ldots, u_k\} \subset L^2_\mu. \tag{2.19}$$

Again by Corollary 2.1, given $G \in \mathrm{CG}_{\mu,k}$, there exists an orthonormal set $\mathcal{U} = \{u_1, \ldots, u_k\} \subset L^2_\mu(\Omega)$, such that $G \subset G_{\mathcal{U}}$. Thus, as cost is monotone, in the sense of (2.12), and from (2.19), we have

$$\operatorname{cost}(G) \geq \operatorname{cost}(G_{\mathcal{U}}) = \operatorname{cost}^*(\mathcal{U}) \geq \operatorname{cost}^*(\mathcal{U}^*) = \operatorname{cost}(G_{\mathcal{U}^*})$$

and the conclusion follows thanks to (2.18). $\square$

Similarly to the previous proposition, the next result shows that the problem of finding nested principal geodesics (see Definition 2.13) can be formulated as a sequence of optimization problem in $L^2_\mu(\Omega)$. The proof is based on the same arguments as the proof of Proposition 2.7, so it is omitted.

Proposition 2.8. *Let $\mu \in W_2^{ac}(\Omega)$ be a reference measure, $\boldsymbol{\nu} \in W_2(\Omega)$ a random measure, $\boldsymbol{v} = \log_\mu(\boldsymbol{\nu})$ and $k \geq 1$ an integer. Let $u_1^*, \ldots, u_k^* \in L^2_\mu(\Omega)$ such that*

$$u_1^* \in \arg\min\{\text{cost}^*(u) \mid u \in L^2_\mu(\Omega), \|u\|_\mu = 1\}$$

and for $j = 2, \ldots, k$,

$$u_j^* \in \arg\min\{\text{cost}^*(u) \mid u \in \text{span}(u_1^*, \ldots, u_{j-1}^*)^\perp, \|u\|_\mu = 1\}$$

where $\perp$ denotes orthogonal. Then $(G_{u_1^*}, G_{\{u_1^*, u_2^*\}} \ldots, G_{\{u_1^*, \ldots, u_k^*\}}) \in \mathcal{N}_{\mu,k}$.

Given data $\nu_1, \ldots, \nu_n \in W_2(\Omega)$, Propositions 2.7 and 2.8 can be applied to the random measure $\boldsymbol{\nu}^{(n)}$ of Definition 2.12. In this case, the empirical version of (2.17) can be written as

$$\text{cost}_n^*(\mathcal{U}) = \frac{1}{n}\sum_{i=1}^n \|v_i - \Pi_{\text{span}(\mathcal{U})\cap V_\mu} v_i\|_\mu^2 \tag{2.20}$$

and an optimal solution $\mathcal{U}^* = \{u_1^*, \ldots, u_k^*\}$ leads to the construction of empirical principal geodesics.

2.3.5 PCA on logarithms

Motivated by the method described in Section 1.2.4, we consider applying the standard PCA to the logarithms of the data. The next proposition provides conditions ensuring that such procedure leads to a solution of GPCA. For the sake of simplicity, we state the result only for global principal geodesics.

Proposition 2.9. *Let $\mu \in W_2^{ac}(\Omega)$ be a reference measure, $\boldsymbol{\nu} \in W_2(\Omega)$ a random measure, $\boldsymbol{v} = \log_\mu(\boldsymbol{\nu})$ and $k \geq 1$ an integer. Let $\tilde{\mathcal{U}} = \{\tilde{u}_1, \ldots, \tilde{u}_k\} \subset L^2_\mu(\Omega)$ be a set of orthonormal eigenvectors associated with the k largest eigenvalues of the covariance operator $K : L^2_\mu(\Omega) \to L^2_\mu(\Omega)$, given by*

$$Kv = \mathbb{E}\langle \boldsymbol{v} - \mathbb{E}(\boldsymbol{v}), v\rangle(\boldsymbol{v} - \mathbb{E}(\boldsymbol{v})),\ v \in L^2_\mu(\Omega). \tag{2.21}$$

If $\Pi_{\text{span}(\tilde{\mathcal{U}})}\boldsymbol{v} \in V_\mu$ a.s. then $G_{\tilde{\mathcal{U}}} \in \mathcal{G}_{\mu,k}$.

Proof. It is well known that $\tilde{\mathcal{U}}$ is minimizer of

$$\widetilde{\text{cost}}(\mathcal{U}) := \mathbb{E}\|\boldsymbol{v} - \Pi_{\text{span}(\mathcal{U})}\boldsymbol{v}\|_\mu^2, \tag{2.22}$$

over orthonormal sets $\mathcal{U} = \{u_1, \ldots, u_k\} \subset L^2_\mu(\Omega)$. On the other hand, from (2.17) and (2.22), it is clear that

$$\text{cost}^*(\mathcal{U}) \geq \widetilde{\text{cost}}(\mathcal{U}),\ \text{for all}\ \mathcal{U} = \{u_1, \ldots, u_k\} \subset L^2_\mu(\Omega). \tag{2.23}$$

As we have assumed that $\Pi_{\text{span}(\tilde{\mathcal{U}})}\boldsymbol{v} \in V_\mu$ a.s. then, from (2.17) and (2.22), we have

$$\text{cost}^*(\tilde{\mathcal{U}}) = \widetilde{\text{cost}}(\tilde{\mathcal{U}}). \tag{2.24}$$

From (2.23) and (2.24) we have that $\tilde{\mathcal{U}}$ is a minimizer of $\text{cost}^*(\mathcal{U})$ over orthonormal sets $\mathcal{U} = \{u_1, \ldots, u_k\} \subset L^2_\mu(\Omega)$. Finally, from Proposition 2.7 we obtain the result. □

Let $\nu_1, \ldots, \nu_n \in W_2(\Omega)$ and $\boldsymbol{\nu}^{(n)}$ from Definition 2.12. If in Proposition 2.9 we replace $\boldsymbol{\nu}$ by $\boldsymbol{\nu}^{(n)}$, we obtain the empirical version of this result. In this case, if $\tilde{\mathcal{U}} = \{\tilde{u}_1, \ldots, \tilde{u}_k\} \subset L^2_\mu(\Omega)$ are orthonormal eigenvectors associated with the k largest eigenvalues of the empirical covariance operator

$$Kv = \frac{1}{n}\sum_{i=1}^{n} \langle v_i - \bar{v}_n, v\rangle (v_i - \bar{v}_n), \; v \in L^2_\mu(\Omega), \tag{2.25}$$

with $v_i = \log_\mu \nu_i$ and if $\Pi_{\mathrm{span}(\tilde{\mathcal{U}})} v_i \in V_\mu$, $i = 1, \ldots, n$, then $G_{\tilde{\mathcal{U}}} \in \mathcal{G}_{\mu,k,n}$.

From Proposition 2.9, we can informally say that, if the data is sufficiently concentrated around the reference measure μ, then the GPCA in $W_2(\Omega)$ can be simply obtained from the standard PCA on logarithms.

2.4 Some numerical examples of GPCA in W_2

In Section 2.4.1 we present an example of concentrated data such that the conditions in Proposition 2.9 are satisfied and so, the problem of finding principal geodesics is reduced to the standard PCA on the data logarithms. On the other hand, in Section 2.4.2 we present an example of spread out data, where the GPCA cannot be obtained from the standard PCA of the logarithms.

2.4.1 The case of sufficiently concentrated data

To illustrate the empirical GPCA in W_2, let us consider the set of $n = 4$ probability measures $\nu_1, \ldots, \nu_4$ with densities $f_1, \ldots, f_4$, displayed in Figure 1. These measures satisfy the homothetic model (2.3), with μ_0 being the standard Gaussian measure (with zero mean and unit variance) and let the values of a_i's and b_i's be given in Table 2.1.

i	1	2	3	4
a_i	0.4	0.8	1.2	1.6
b_i	-1.8	-0.1	0.7	1.2

Table 2.1: Values of (a, b).

We first calculate the Fréchet mean $\bar{\nu}_4$ of $\nu_1, \ldots, \nu_4$. To that end, we apply the quantile average formula (2.15), from which we obtain the density $\bar{g}_4$ of $\bar{\nu}_4$ (displayed in Figure 1(f)), given by

$$\bar{g}_4(x) = f^{(\bar{a}_4, \bar{b}_4)}(x) = \frac{1}{\bar{a}_4} f_0\left(\frac{x - \bar{b}_4}{\bar{a}_4}\right), \; x \in \mathbb{R},$$

where $\bar{a}_4 = \frac{1}{4}\sum_{i=1}^4 a_i = 1$ and $\bar{b}_4 = \frac{1}{4}\sum_{i=1}^4 b_i = 0$. Hence, in this example, one has that $\bar{g}_4 = f_0$ and $\bar{\nu}_4 = \mu_0$.

Observe that $\nu_1, \ldots, \nu_4$ are concentrated around their empirical Fréchet mean $\bar{\nu}_4$, in the sense that their expectations and variances are not too far from those of $\bar{\nu}_4$, see Figure 1. Let us apply Proposition 2.9 to compute explicitly an empirical first principal geodesic, taking $\mu = \mu_0$ as reference measure. Let $\tilde{u}$ be the eigenvector associated with the largest eigenvalue of the empirical covariance operator

$$Kv = \frac{1}{4}\sum_{i=1}^{4} \langle v_i - \bar{v}, v \rangle (v_i - \bar{v}),\ v \in L^2_{\mu_0}(\mathbb{R})$$

where

$$v_i(x) = \log_{\mu_0}(\nu_i)(x) = \left(\frac{a_i}{\bar{a}_4} - 1\right) x + b_i - \bar{b}_4 \frac{a_i}{\bar{a}_4} = (a_i - 1)\, x + b_i,\ x \in \mathbb{R},$$

for $i = 1, \ldots, 4$. Using the fact that the v_i's are affine functions belonging to the space generated by the identity function and the constant function 1, which are orthonormal in $L^2_{\mu_0}(\mathbb{R})$, the operator K can be identified with the 2×2 matrix

$$M = \frac{1}{4}\sum_{i=1}^{4} V_i' V_i,$$

with $V_i = (a_i - 1, b_i)' \in \mathbb{R}^2$, $1 \leq i \leq 4$. Therefore, $\tilde{u}$ is also an affine function i.e. $\tilde{u}(x) = \tilde{\alpha}x + \tilde{\beta}, x \in \mathbb{R}$, where $\tilde{U} = (\tilde{\alpha}, \tilde{\beta})' \in \mathbb{R}^2$ (with $\tilde{\alpha} = 0.36$ and $\tilde{\beta} = 0.93$) is the eigenvector associated with the largest eigenvalue of M. In other words, computing $\tilde{U}$ simply amounts to calculating the first eigenvector associated with the standard PCA of $V_i \in \mathbb{R}^2, i = 1, \ldots, 4$, that represent the slope and intercept parameters of the affine functions v_i. In Figure 8, we display the vectors V_i (blue circles) for $1 \leq i \leq 4$, together with the linear space spanned by $\tilde{U}$ (red dash-dot line), which corresponds to the first principal direction of variation of this data set. Any affine function $u(x) = \alpha x + \beta$ in $L^2_{\mu_0}$ can be represented in $\mathbb{R}^2$ by the vector $U = (\alpha, \beta)'$. If $\alpha \geq -1$, then such function belongs to V_{μ_0}, which corresponds to the case where the point $(\alpha, \beta) \in \mathbb{R}^2$ is located to the right of the vertical green dashed line in Figure 8. Hence, it can be seen from the projections of the vectors V_i onto the linear space spanned by $\tilde{U}$, that $\Pi_{\mathrm{span}(\tilde{u})} v_i \in V_{\mu_0}$ for all $1 \leq i \leq 4$, and therefore, from Proposition 2.9, we conclude that the family of probability measures

$$G_{\tilde{u}} = \left\{ \tilde{\nu}_t := \exp_{\mu_0}(t\tilde{u}),\ \text{for all } t \in \mathbb{R} \text{ such that } 1 + t\tilde{\alpha} \geq 0 \right\},$$

is a first empirical principal geodesic. From (2.1) and (2.8), each $\tilde{\nu}_t$ in $G_{\tilde{u}}$ admits the density

$$\tilde{g}_t(x) = \frac{1}{\tilde{a}_t} f_0\left(\frac{x - \tilde{b}_t}{\tilde{a}_t}\right),\ x \in \mathbb{R}, \tag{2.26}$$

with $\tilde{a}_t = 1 + t\tilde{\alpha}$ and $\tilde{b}_t = t\tilde{\beta}$. In Figure 6, we display the first principal mode of geodesic variation $\tilde{g}_t$, for $-2 \leq t \leq 2$, of the densities displayed in Figure 1. As already mentioned, the GPCA in W_2 gives a better interpretation of the data variability when compared to the results given by the first principal mode of linear variation of these densities (in the Hilbert space $L^2(\mathbb{R})$), displayed in Figure 2.

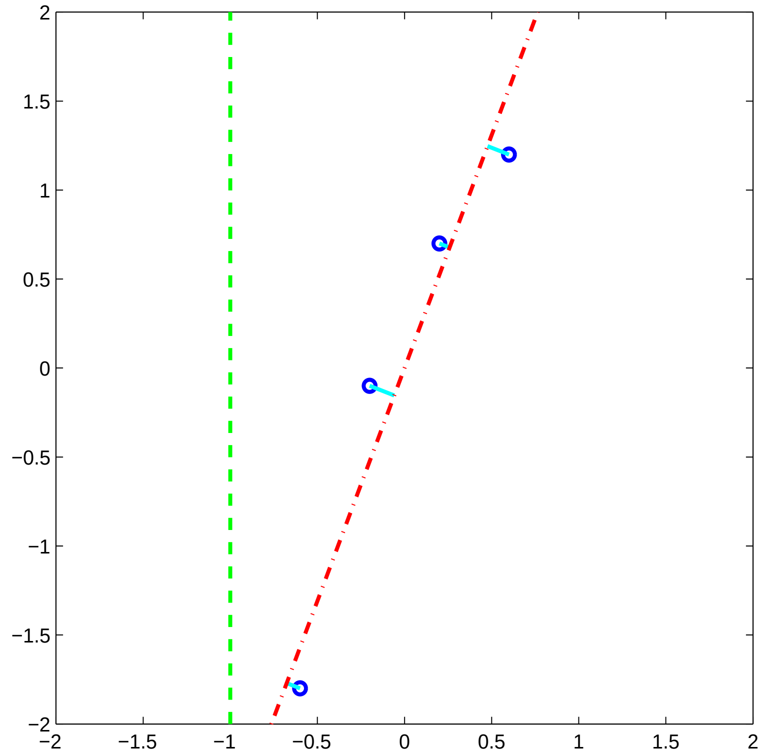


Figure 8: A two-dimensional representation of the affine functions $u(x) = \alpha x + \beta$ in $L^2_{\mu_0}$. The horizontal axis is the slope parameter α, and the vertical axis is the intercept parameter β. The points to the right of the vertical green dashed line at $\alpha = -1$ correspond to the affine functions belonging to V_{μ_0}. The blue circles correspond to the vectors $V_i = (a_i - 1, b_i)' \in \mathbb{R}^2$ that are associated with the affine functions $v_i(x) = (a_i - 1)\,x + b_i$, for $1 \leq i \leq 4$, corresponding to the measures with densities displayed in Figure 1. The dash-dot red line is the linear space spanned by the first eigenvector $\tilde{U} \in \mathbb{R}^2$ of a standard PCA of the vectors $V_1, \ldots, V_4$.

2.4.2 The case of spread-out data

In this section, we exhibit a situation where the GPCA of $\nu_1, \ldots, \nu_n$ in W_2 and the standard PCA of the functions $v_i = \log_{\bar{\nu}_n}(\nu_i), i = 1, \ldots, n$ in the Hilbert space $L^2_{\bar{\nu}_n}$, lead to different results. Let $\nu_1, \ldots, \nu_4$ from the homothetic model (2.3), with μ_0 again chosen as the standard Gaussian measure and the values of a_i's and b_i's be given in Table 2.2.

i	1	2	3	4
a_i	0.2	0.2	0.2	3.4
b_i	-3	-1	1	3

Table 2.2: Values of (a, b).

Since $\bar{a}_4 = 1$ and $\bar{b}_4 = 0$, the empirical Fréchet mean of these measures is $\bar{\nu}_4 = \mu_0$, as in the example in Section 2.4.1. From Figure 9 it can be seen that $\nu_1, \ldots, \nu_4$ are less concentrated around their Fréchet mean when compared with the previous example (see Figure 1).

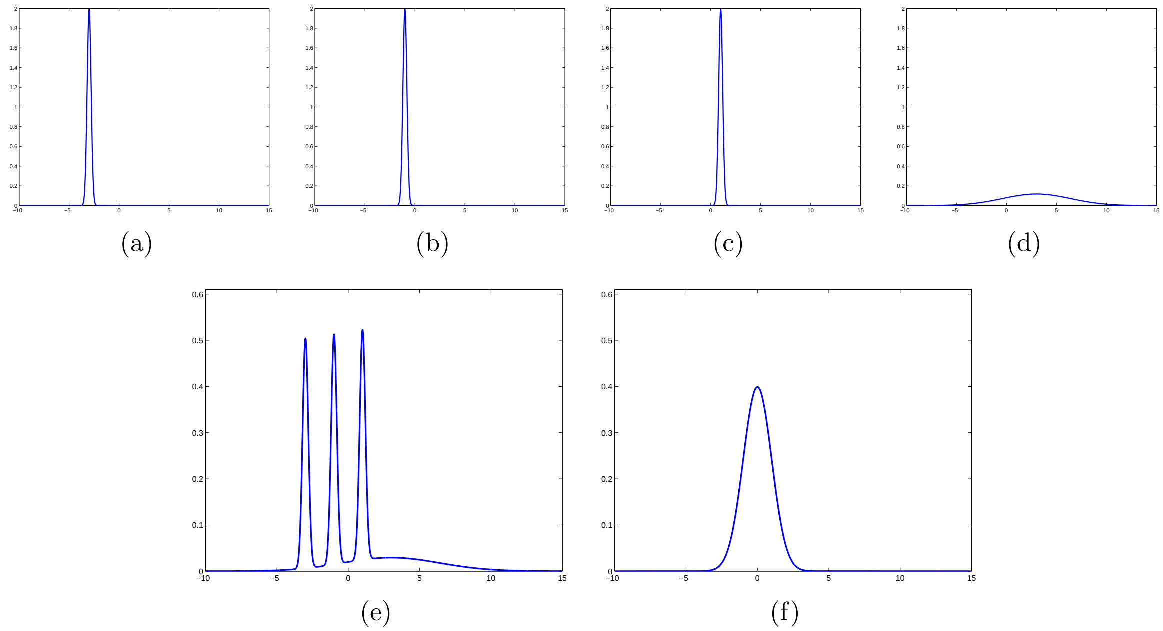


Figure 9: An example of $n = 4$ Gaussian densities $f_1, \ldots, f_4$ satisfying the homothetic model (2.3) with means and variances given in Table 2.2. (e) Euclidean mean of these densities in the Hilbert space $L^2(\mathbb{R})$. (f) Density of the barycenter $\bar{\nu}_4$ in the Wasserstein space W_2 of the probability measures $\nu_1, \ldots, \nu_4$ with densities $f_1, \ldots, f_4$.

We apply Proposition 2.7 to compute an emprirical first principal geodesic. Let

$$v_i(x) = \log_{\mu_0}(\nu_i)(x) = (a_i - 1)\, x + b_i,\ x \in \mathbb{R},$$

for $i = 1, \ldots, 4$. Recall that the affine functions v_i's belong to the space generated by the identity function and the constant function 1, which are orthonormal in $L^2_{\mu_0}(\mathbb{R})$. Recall also that an affine function $u(x) = \alpha x + \beta$ belongs to V_{μ_0} if and only if $\alpha \geq -1$. Therefore, if $S = \{(\alpha, \beta)' \in \mathbb{R}^2 \mid \alpha \geq -1\}$ then minimizing cost_4^* in (2.20) is equivalent to minimizing

$$U \in \mathbb{R}^2 \mapsto \frac{1}{4} \sum_{i=1}^{4} \|V_i - \Pi_{\mathrm{span}(U) \cap S} V_i\|^2, \tag{2.27}$$

with $V_i = (a_i - 1, b_i)' \in \mathbb{R}^2$, $1 \leq i \leq 4$. We have numerically found a unique minimizer $U^* = (\alpha^*, \beta^*)$ of (2.27), and so $u^*(x) = \alpha^* x + \beta^*$, is the unique minimizer of cost_4^*. From Proposition 2.7, the set of probability measures

$$G_{u^*} := \left\{ \nu_t^* := \exp_{\mu_0}(tu^*),\ \text{for all } t \in \mathbb{R} \text{ such that } 1 + t\alpha^* \geq 0 \right\},$$

is a first empirical principal geodesic, and, from (2.1) and (2.8), each ν_t^* in G_{u^*} admits the density

$$g_t^*(x) = \frac{1}{a_t^*} f_0 \left(\frac{x - b_t^*}{a_t^*} \right),\ x \in \mathbb{R}, \tag{2.28}$$

with $a_t^* = 1 + t\alpha^*$ and $b_t^* = t\beta^*$.

We have also computed $\tilde{U} = (\tilde{\alpha}, \tilde{\beta})' \in \mathbb{R}^2$, the eigenvector associated with the largest eigenvalue of the empirical covariance operator M in (**??**) of $V_1, \ldots, V_4$ and obtain that the affine function $\tilde{u}(x) = \tilde{\alpha}x + \tilde{\beta}$, is the eigenvector associated with the largest eigenvalue of the empirical covariance operator of $v_1, \ldots, v_4$. We see that $U^* \neq \tilde{U}$, so we have $\mathrm{cost}^*_n(u^*) < \mathrm{cost}^*_n(\tilde{u})$. Hence, from (2.19), we obtain

$$\mathrm{cost}_n(G_{u^*}) < \mathrm{cost}_n(G_{\tilde{u}}).$$

The previous inequality proves that computing a GPCA in W_2 of a set of measures $\nu_1, \ldots, \nu_n$ cannot always be achieved by performing a standard functional PCA in $L^2_{\mu_0}(\mathbb{R})$ of the functions $v_i = \log_{\mu_0}(\nu_i), i = 1, \ldots, 4$. Figure 10 shows the vectors V_i's and the convex sets $\mathrm{span}(U^*) \cap S$ and $\mathrm{span}(\tilde{U}) \cap S$. From this figure one can see that (2.27) is strictly smaller if evaluated at U^* than evaluated at $\tilde{U}$.

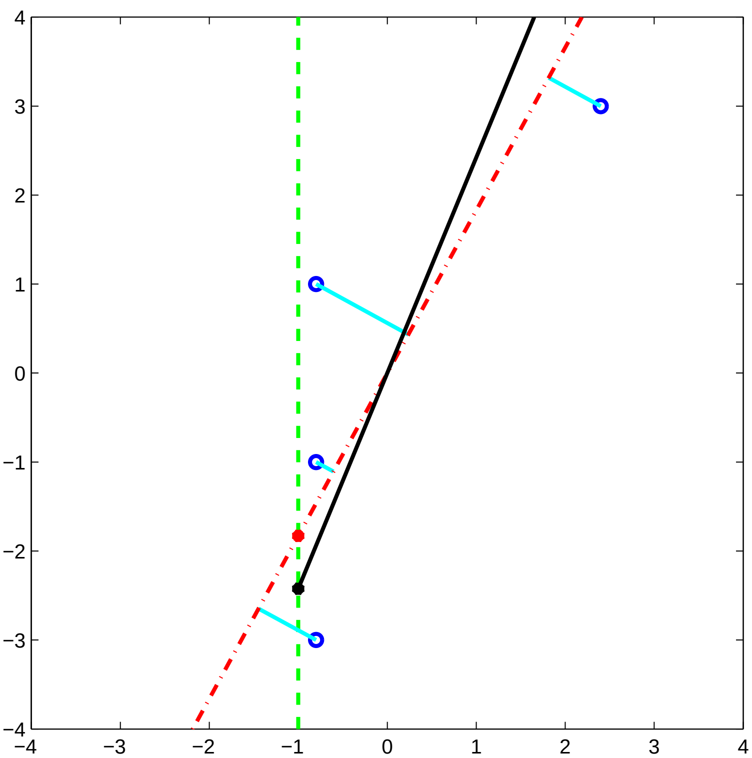


Figure 10: A two-dimensional representation of the affine functions $u(x) = \alpha x + \beta$ in $L^2_{\mu_0}(\Omega)$. The horizontal axis is the slope parameter α and the vertical axis is the intercept parameter β. The points that lie in S (the region to the right of the vertical green dashed line at $\alpha = -1$) correspond to the affine functions belonging to V_{μ_0}. The blue circles correspond to the vectors $V_i = (a_i - 1, b_i)' \in \mathbb{R}^2$ that are associated with the affine functions $v_i(x) = (a_i - 1)\, x + b_i$, for $1 \leq i \leq 4$, which are the logarithms of the measures with densities displayed in Figure 9. The dash-dot red line is the linear space spanned by the first eigenvector $U' \in \mathbb{R}^2$ of the standard PCA of $V_1, \ldots, V_4$. The black line is the convex set $\mathrm{span}(U^*) \cap S$, where $U^* = (\alpha, \beta^*)' \in \mathbb{R}^2$ is the minimizer of (2.27). The black dot is the projection of $V_1 = (0.2, -3)$ onto $\mathrm{span}(U^*) \cap S$, while the red dot is the projection of the vector V_1 onto $\mathrm{span}(U') \cap S$.

2.5 Consistency of the empirical GPCA

As in Section 2.3, we assume that Ω is compact and that $\boldsymbol{\nu}$ is a square-integrable random element in $W_2(\Omega)$. We will study the convergence of the empirical Fréchet mean and the empirical global principal geodesics to their population counterparts, when $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n$ are iid copies of $\boldsymbol{\nu}$.

Proposition 2.10. *Let $\boldsymbol{\nu}$ be a random measure in $W_2(\Omega)$. Let $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n$ be iid copies of $\boldsymbol{\nu}$ and denote by $\bar{\boldsymbol{\nu}}_n$ their empirical Fréchet mean. Then, as $n \to \infty$,*

$$d_{W_2}(\bar{\boldsymbol{\nu}}_n, \mathcal{M}(\boldsymbol{\nu})) \to 0 \ a.s.$$

Proof. Let $\mu \in W_2^{ac}(\Omega)$, then from Proposition 2.3(*ii*), $\log_\mu(\mathcal{M}(\boldsymbol{\nu})) = \mathbb{E}(\boldsymbol{v})$ and $\log_\mu(\bar{\boldsymbol{\nu}}_n) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{v}_i$, where $\boldsymbol{v} = \log_\mu(\boldsymbol{\nu})$ and $\boldsymbol{v}_i = \log_\mu(\boldsymbol{\nu}_i), i = 1, \ldots, n$. Observe that $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ are iid copies of $\boldsymbol{v}$ and that $\mathbb{E}(\|\boldsymbol{v}\|_\mu^2) = \mathbb{E}(d_W^2(\boldsymbol{\nu}, \mu)) < \infty$. Then, By Proposition 2.2, $d_W^2(\bar{\boldsymbol{\nu}}_n, \mathcal{M}(\boldsymbol{\nu})) = \|\frac{1}{n}\sum_{i=1}^n \boldsymbol{v}_i - \mathbb{E}(\boldsymbol{v})\|_\mu^2 \to 0$, a.s., by the strong law of large numbers in a Hilbert space (see e.g. [Ledoux and Talagrand, 2011]). $\square$

Observe that the previous lemma is also valid for non compact Ω, provided that $\boldsymbol{\nu}$ is square-integrable. In the following lemma we show that the indicators of $\mathrm{CG}_{\mu_n,k}$, Γ-converge to the indicator of $\mathrm{CG}_{\mu,k}$ when μ_n converges to μ in $W_2^{ac}(\Omega)$. We refer to Section A.4 in the Appendix for the definitions of Γ-convergence and of indicator function.

Lemma 2.3. *Let (μ_n) be a sequence in $W_2^{ac}(\Omega)$ converging to $\mu \in W_2^{ac}(\Omega)$, then*

$$\Gamma\text{-}\lim_{n\to\infty} \chi_{\mathrm{CG}_{\mu_n,k}} = \chi_{\mathrm{CG}_{\mu,k}}. \tag{2.29}$$

Proof. By Lemma A.4 with $(X, d) = (CL, h_{W_2})$, it is sufficient to show that $\mathrm{CG}_{\mu_n,k}$ converges to $\mathrm{CG}_{\mu,k}$ in the sense of Kuratowski. That is, we have to show that

(a) for every $G \in \mathrm{CG}_{\mu,k}$ there exists a sequence (G_n) converging to G such that $G_n \in \mathrm{CG}_{\mu_n,k}$, for every $n \geq 1$, and

(b) if G is an accumulation point of a sequence (G_n), with $G_n \in \mathrm{CG}_{\mu_n,k}, n \geq 1$, then $G \in \mathrm{CG}_{\mu,k}$.

For (a) let $R_n : L^2_\mu(\Omega) \to L^2_{\mu_n}(\Omega)$ be defined by $R_n(u) = u \circ F^- \circ F_n$, where F and F_n are the cdf of μ and μ_n respectively. Let also $T_n : W_2(\Omega) \to W_2(\Omega)$ be given by $T_n = \exp_{\mu_n} \circ R_n \circ \log_\mu$, $n \geq 1$. Take $G \in \mathrm{CG}_{\mu,k}$ and define $G_n = T_n(G)$, $n \geq 1$. From Corollary 2.1 and as R_n is linear, it is easy to check that $G_n \in \mathrm{CG}_{\mu_n,k}, n \geq 1$. Denote by d_μ the distance in $L^2_\mu(\Omega)$. As the logarithmic map is an isometry and after some calculation we get that the deviation from G_n to G (see Definition A.7), satisfies

$$d_{W_2}(G, G_n) = d_\mu(\log_\mu(G), \log_\mu(G_n)) \leq \|\log_\mu(\mu_n)\|_\mu = d_{W_2}(\mu, \mu_n),\ n \geq 1.$$

Similarly, $d_{W_2}(G_n, G) \leq d_{W_2}(\mu, \mu_n)$, $n \geq 1$, and we conclude that

$$h_{W_2}(G, G_n) \leq d_{W_2}(\mu, \mu_n) \to 0,\ \text{as } n \to \infty.$$

For (b) take $G, G_n, n \geq 1$ as stated above. Since $\mu_n \in G_n, n \geq 1$, and $\mu_n \to \mu$, it follows that $\mu \in G$, by (ii) in Definition A.6.

On the other hand, let $\mathrm{CG}_k = \{G \in \mathrm{CL} \mid G \text{ is a geodesic set and } \dim(G) \leq k\}$, which is shown to be compact, following the same arguments in Proposition 2.5. Then, as $G_n \in \mathrm{CG}_k, n \geq 1$, we have $G \in \mathrm{CG}_k$ and recalling that $\mu \in G$, we conclude that $G \in \mathrm{CG}_{\mu,k}$. $\square$

As CL is compact, every sequence (G_n) with G_n belonging to the set of empirical global k-principal $\bar{\boldsymbol{\nu}}_n$-geodesics $\mathcal{G}_{\bar{\boldsymbol{\nu}}_n,k,n}, n \geq 1$ has a convergent subsequence. The following theorem ensures that the limit set belongs almost surely to the population global k-principal $\mathcal{M}(\boldsymbol{\nu})$-geodesics $\mathcal{G}_{\mathcal{M}(\boldsymbol{\nu}),k}$. This is the main result about consistency of global GPCA.

Theorem 2.4. *Let $\boldsymbol{\nu}$ be a random measure in $W_2^{ac}(\Omega)$ with Fréchet mean $\mu := \mathcal{M}(\boldsymbol{\nu})$ and let $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n$ be iid copies of $\boldsymbol{\nu}$, with empirical Fréchet mean $\mu_n := \bar{\boldsymbol{\nu}}_n$. Let $\mathcal{G}_{\mu,k}$ be the global k-principal μ-geodesics of $\boldsymbol{\nu}$ and $\mathcal{G}_{\mu_n,k,n}$ be the empirical global k-principal μ_n-geodesics of $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n$.*

Then for every sequence (G_n), with $G_n \in \mathcal{G}_{\mu_n,k,n}$, one has

$$\lim_{n\to\infty} \mathrm{cost}_n(G_n) = \min\{\mathrm{cost}(G) \mid G \in \mathrm{CG}_{\mu,k}\} \ a.s. \tag{2.30}$$

Moreover, the accumulation points of (G_n) belong to $\mathcal{G}_{\mu,k}$ a.s. In other words, subsequential limits of (G_n) are global principal population geodesics.

Proof. Observe that $\mu, \mu_n \in W_2^{ac}(\Omega), n \geq 1$, by Proposition 2.3(*iv*). Also, by Proposition 2.10, $\mu_n \to \mu$ a.s. On the other hand, one may remark that the set of global principal geodesics (2.14) can be characterized as

$$\mathcal{G}_{\mu,k} = \arg\min\{\mathrm{cost}(G) + \chi_{\mathrm{CG}_{\mu,k}}(G) \mid G \in \mathrm{CL}\}, \tag{2.31}$$

where $\chi_{\mathrm{CG}_{\mu,k}} : \mathrm{CL} \to \mathbb{R} \cup \{+\infty\}$ is the indicator of $\mathrm{CG}_{\mu,k}$ according to Definition A.11. Similarly,

$$\mathcal{G}_{\mu_n,k,n} = \arg\min\{\mathrm{cost}_n(G) + \chi_{\mathrm{CG}_{\mu_n,k(W)}}(G) \mid G \in \mathrm{CL}\}. \tag{2.32}$$

By applying Lemma 2.3, we have that

$$\Gamma\text{-}\lim_{n\to\infty} \chi_{\mathrm{CG}_{\mu_n,k}} = \chi_{\mathrm{CG}_{\mu,k}} \ \text{ a.s.}, \tag{2.33}$$

where the Γ-convergence holds in the space CL. From Proposition 2.4 and recalling that $W_2(\Omega)$ is compact, we have that $d^2_{W_2}(\nu, G)$ is separately continuous in $\nu \in W_2(\Omega)$ and $G \in \mathrm{CL}$. Hence $d^2_{W_2}(\nu, G)$ is measurable on the product space $W_2(\Omega) \times \mathrm{CL}$; see [Johnson, 1969] or [Rudin, 1981]. Also, from Theorem 2.3 in [Artstein and Wets, 1995], we have the following Γ-convergence in CL,

$$\Gamma\text{-}\lim_{n\to\infty} \mathit{cost}_n = \mathit{cost} \text{ a.s.} \tag{2.34}$$

On the other hand, as $W_2(\Omega)$ is compact, there exists a constant $R > 0$ such that $d^2_{W_2}(\nu, G) \leq R$, for all $\nu \in W_2(\Omega)$ and $G \in \mathrm{CL}$. Also, by Proposition 2.5, CL is a compact set. Therefore, by the uniform strong law of large number (see Lemma 2.4 in [Newey and McFadden, 1994]), $\mathrm{cost}_n(G) \to \mathrm{cost}(G)$ uniformly in CL a.s., that is,

$$\lim_{n\to\infty} \sup_{G\in\mathrm{CL}} |\mathrm{cost}(G) - \mathrm{cost}_n(G)| = 0 \text{ a.s.} \tag{2.35}$$

From (2.33) to (2.35) and by Proposition 6.24 in [Dal Maso, 1993], we obtain

$$\Gamma\text{-}\lim_{n\to\infty} \mathrm{cost}_n + \chi_{\mathrm{CG}_{\mu_n,k}} = \mathrm{cost} + \chi_{\mathrm{CG}_{\mu,k}} \text{ a.s.} \tag{2.36}$$

Therefore, from (2.31), (2.32) and (2.36); the compactness of CL and Theorem A.1, the conclusions follow. □

The following theorem ensures that the limit set of sequences $(\boldsymbol{G}_n)$ with $\boldsymbol{G}_n \in \mathcal{N}_{\bar{\boldsymbol{\nu}}_n,k,n}, n \geq 1$, belongs almost surely to the population nested k-principal $\mathcal{M}(\boldsymbol{\nu})$-geodesics $\mathcal{N}_{\mathcal{M}(\boldsymbol{\nu}),k}$. This is the main result about consistency of nested GPCA.

Theorem 2.5. *Let $\boldsymbol{\nu}$ be a random measure in $W_2^{ac}(\Omega)$, with Fréchet mean $\mu := \mathcal{M}(\boldsymbol{\nu})$, and let $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n$ be iid copies of $\boldsymbol{\nu}$, with empirical Fréchet mean $\mu_n := \bar{\boldsymbol{\nu}}_n$. Let $\mathcal{N}_{\mu,k}$ be the nested k-principal μ-geodesics of $\boldsymbol{\nu}$ and $\mathcal{N}_{\mu_n,k,n}$ the empirical nested k-principal μ_n-geodesics of $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n$. If $\boldsymbol{G}_n = (G_{1,n}, \ldots, G_{k,n}) \in \mathcal{N}_{\mu_n,k,n}$ and $(\boldsymbol{G}_{n'})$ is a subsequence converging to $\boldsymbol{G} = (G_1, \ldots, G_k)$, then $\boldsymbol{G} \in \mathcal{N}_{\mu,k}$ and $\mathrm{cost}_{n'}(G_{j,n'}) \to \mathrm{cost}(G_j)$, $j = 1, \ldots, k$.*

Proof. Let us show the result by induction on k. First observe that $\mathcal{N}_{\mu_n,1,n} = \mathcal{G}_{\mu_n,1,n}$ and $\mathcal{N}_{\mu,1} = \mathcal{G}_{\mu,1}$, then the case $k = 1$ follows from Theorem 2.4. Let us assume that the result is valid for $k-1$ and show that it is also true for $k \geq 2$. Observe that set of population and empirical nested principal geodesic can be expressed as

$$\mathcal{N}_{\mu,k} = \{(G_1, \ldots, G_k) \mid (G_1, \ldots, G_{k-1}) \in \mathcal{N}_{\mu,k-1}, G_k \in \mathcal{G}_{\mu,k} \text{ and } G_{k-1} \subset G_k\} \tag{2.37}$$

and

$$\mathcal{N}_{\mu_n,k,n} = \{(G_1, \ldots, G_k) \mid (G_1, \ldots, G_{k-1}) \in \mathcal{N}_{\mu_n,k-1,n}, G_k \in \mathcal{G}_{\mu_n,k,n} \text{ and } G_{k-1} \subset G_k\}, \tag{2.38}$$

where $\mathcal{G}_{\mu,k}$ and $\mathcal{G}_{\mu_n,k,n}$ are the sets of population and empirical global principal geodesic, respectively. Let $\boldsymbol{G}_n = (G_{1,n}, \ldots, G_{k,n}) \in \mathcal{N}_{\mu_n,k,n}$ and let $(\boldsymbol{G}_{n'})$ be a subsequence of $(\boldsymbol{G}_n)$ converging to an element $\boldsymbol{G} = (G_1, \ldots, G_k)$. From (2.38) we have $(G_{1,n}, \ldots, G_{k-1,n}) \in \mathcal{N}_{\mu_n,k-1,n}$ and therefore, by hypothesis of induction, $(G_1, \ldots, G_{k-1}) \in \mathcal{N}_{\mu,k-1}$ and $\mathrm{cost}_{n'}(G_{j,n'}) \to \mathrm{cost}(G_j)$, $j = 1, \ldots, k-1$. Also from (2.38), $G_{k,n} \in \mathcal{G}_{\mu_n,k,n}$, $n \geq 1$, hence $G_k \in \mathcal{G}_{\mu,k}$ and $\mathrm{cost}_n(G_{k,n}) \to \mathrm{cost}(G_k)$, thanks to Theorem 2.4. Finally, from Lemma A.3, $G_{k-1} \subset G_k$ and since $(G_1, \ldots, G_{k-1}) \in \mathcal{N}_{\mu,k-1}$ and $G_k \in \mathcal{G}_{\mu,k}$, the result follows from (2.37). □

The interpretation of Theorems 2.4 and 2.5 is that the empirical GPCA is strongly consistent in its two variants. From the proofs of these theorems, one can immediately see that $\mu, \mu_n, n \geq 1$, can be arbitrary measures in $W_2^{ac}(\Omega)$ provided that $\mu_n \to \mu$. However, it seems natural to use $\mu_n = \bar{\boldsymbol{\nu}}_n$ as a reference measures to compute the empirical GPCA in the Wasserstein space $W_2(\Omega)$. In this case, we obtain convergence to the population GPCA, related to the reference measure $\mu = \mathcal{M}(\boldsymbol{\nu})$.

2.6 Conclusions and discussion of this chapter

In this concluding section we make a comparison between GPCA in W_2 and analogs of PCA on Riemannian manifolds. As already mentioned in the introduction, nonlinear analogs of

PCA have been proposed in the literature [Fletcher et al., 2004; Huckemann et al., 2010] for the analysis of data belonging to curved Riemannian manifolds. To perform a PCA-like analysis, two popular approaches are: (1) standard PCA of the data projected onto the tangent space at their Fréchet mean, with back projection onto the manifold, as presented in Section 1.2.4 and (2) PGA along geodesics, as described in Section 1.2.3. These two approaches lead generally to different directions of geodesic variability in a curved manifold [Sommer et al., 2010].

In this chapter we consider the analysis of data in the Wasserstein space $W_2(\Omega)$, which is not a Riemannian manifold but has pseudo-Riemannian structure, rich enough to allow the definition of a notion of geodesic PCA. By means of the analogs of the logarithmic and of the exponential maps, we also introduce the corresponding version of the standard PCA in the tangent space, with back projection onto $W_2(\Omega)$, thus establishing a parallel to the methodological duality available for data in Riemannian manifolds, described in the introduction. Also, as could be expected, these two approaches yield, in general, different forms of geodesic variability.

There is however a significant distinguishing feature of our methodology, namely the possibility of performing a PCA on the tangent space under convexity restrictions, which is equivalent (after projection) to the geodesic PCA in $W_2(\Omega)$. This restricted PCA on the logarithms of the data is interesting because it is formally simpler than the geodesic PCA in $W_2(\Omega)$ although more complex than standard PCA. In this respect it is also worth noticing that if the data are "sufficiently concentrated", the standard and the restricted PCA in the tangent space yield the same results.

Finally, it should be mentioned that the terminology geodesic PCA (GPCA) was used previously in [Huckemann et al., 2010] to denote a Riemannian manifold generalization of linear PCA. Their approach shares similarities with the PGA method introduced in [Fletcher et al., 2004], but optimizes additionally for the placement of the center point (not necessarily equal to the Fréchet mean). Furthermore, it does not use a linear approximation of the manifold and is only suited for Riemannian manifolds, where explicit formulas for geodesics exist. However, it is difficult to compare our approach to the GPCA in [Huckemann et al., 2010] since the notion of principal geodesic, that we propose in this chapter, is defined with respect to a given reference measure μ (chosen to be either the population or the empirical Fréchet mean). For a precise comparison it would be necessary to carry out the optimization (2.14) with respect to the reference measure μ, a task which is beyond the scope of this thesis.

2.7 Extensions and related problems

In this section we discuss possible extensions and related problems of the geodesic PCA. In this thesis we have not explored such issues in depth, as they can be the object of future investigations.

2.7.1 Consistency of GPCA in a statistical deformable model

In Section 2.5 we study the consistency of the empirical GPCA of $\nu_1, \dots, \nu_n$, when $n \to \infty$. In the current section we consider the consistency in the context of the following statistical model:

$$\boldsymbol{\nu}_i^\epsilon = \boldsymbol{\nu}_i * \boldsymbol{\kappa}_i^\epsilon, \; i = 1, \dots, n, \tag{2.39}$$

where

- $\boldsymbol{\nu}_i = \exp_{\mu_0}(\boldsymbol{u}_i), i = 1, \dots, n$ and $\boldsymbol{u}_1, \dots, \boldsymbol{u}_n$ are iid, zero mean, V_{μ_0}-valued random variables, such that $\|\boldsymbol{u}_1\|_{\mu_0} \le r$ a.s. with $r > 0$,
- μ_0 is a mean pattern belonging to $W_2^{ac}(\Omega)$,
- the symbol $*$ denotes the convolution operator,
- $\epsilon \ge 0$ is a noise level parameter.
- Furthermore, for every $\epsilon \ge 0$, $\boldsymbol{\kappa}_1^\epsilon, \dots, \boldsymbol{\kappa}_n^\epsilon$ are iid $W_2^{ac}(\Omega)$-valued random variables such that, $\boldsymbol{\kappa}_1^0 = \delta_0$, $d_{W_2}(\boldsymbol{\kappa}_1^\epsilon, \delta_0) \to 0$ a.s. as $\epsilon \to 0$, with δ_0 the Dirac measure on 0 and $d_{W_2}(\boldsymbol{\kappa}_1^\epsilon, \delta_0) \le s$ a.s., for all $\epsilon > 0$, with $s > 0$.
- $(\boldsymbol{\nu}_1, \dots, \boldsymbol{\nu}_n)$ and $(\boldsymbol{\kappa}_1^\epsilon, \dots, \boldsymbol{\kappa}_n^\epsilon)$ are mutually independent, for all $\epsilon > 0$.
- We either assume that ϵ belongs to a denumerable set or that $\{\boldsymbol{\kappa}_1^\epsilon : \epsilon \ge 0\}$ is a separable process.

Let $\boldsymbol{\nu}^\epsilon, \boldsymbol{\nu}, \boldsymbol{u}, \boldsymbol{\kappa}^\epsilon$ be generic random variables distributed as $\boldsymbol{\nu}_1^\epsilon, \boldsymbol{\nu}_1, \boldsymbol{u}_1, \boldsymbol{\kappa}_1^\epsilon$ respectively, for $\epsilon \ge 0$. In model (2.39), the mean pattern μ_0 is randomly warped by means of $\boldsymbol{u}$ and randomly blurred by means of $\boldsymbol{\kappa}^\epsilon$. An example of random blur is given by $\boldsymbol{\kappa}^\epsilon = \exp_\kappa((Y^\epsilon - 1)id)$, where $\kappa \in W_2^{ac}(\Omega)$ is such that $\int x^2 d\kappa(x) \le 1$ and $Y^\epsilon, \epsilon > 0$, is a family of positive valued random variables such that, as $\epsilon \to 0$, $Y^\epsilon \to 0$ a.s. and $Y^\epsilon \le s$ a.s. In fact, as $\log_\kappa(\delta_0) = -id$ and by Proposition 2.2, $d_{W_2}(\boldsymbol{\kappa}^\epsilon, \delta_0) = \|\log_\kappa(\boldsymbol{\kappa}^\epsilon) - \log_\kappa(\delta_0)\|_\kappa = \|(Y^\epsilon - 1)id + id\|_\kappa = \|Y^\epsilon id\|_\kappa \le s(\int x^2 d\kappa(x))^{\frac{1}{2}} \le s$.

The following proposition provides a technical tool to deal with the convolution operator in model 2.39. The proof follows from [Villani, 2003], Proposition 7.17.

Proposition 2.11. *Let $\mu, \kappa \in W_2(\Omega)$ and δ_0 be the Dirac measure on 0, then $d_{W_2}(\mu * \kappa, \mu) \le d_{W_2}(\kappa, \delta_0)$.*

Observe that $\boldsymbol{\nu}$ and $\boldsymbol{\nu}^\epsilon$ are square-integrable according to Definition 2.5. In fact, by Propositions 2.2, $d_{W_2}(\boldsymbol{\nu}, \mu_0) = \|\boldsymbol{u}\|_{\mu_0} \le r < \infty$ and, by Proposition 2.11, $d_{W_2}(\boldsymbol{\nu}^\epsilon, \mu_0) \le d_{W_2}(\boldsymbol{\nu}^\epsilon, \boldsymbol{\nu}) + d_{W_2}(\boldsymbol{\nu}, \mu_0) \le d_{W_2}(\boldsymbol{\kappa}^\epsilon, \delta_0) + r \le s + r < \infty$. Thus $\mathcal{M}(\boldsymbol{\nu}) = \mathcal{M}(\exp_{\mu_0}(\boldsymbol{u})) = \exp_{\mu_0}(\mathbb{E}(\boldsymbol{u})) = \exp_{\mu_0}(0) = \mu_0$ and $\boldsymbol{\nu}^\epsilon$ has a unique Fréchet mean $\mathcal{M}(\boldsymbol{\nu}^\epsilon)$, thanks to Proposition 2.3.

We are interested in the problem of analyzing the consistency of GPCA when $n \to \infty$ and $\epsilon \to 0$. The following Proposition is a first step towards such analysis.

Proposition 2.12. *Let $\boldsymbol{\nu}_1^\epsilon, \dots, \boldsymbol{\nu}_n^\epsilon$ from the model (2.39) and let $\bar{\boldsymbol{\nu}}_n^\epsilon$ be the corresponding empirical Frechet mean. Then $d_{W_2}(\bar{\boldsymbol{\nu}}_n^\epsilon, \mu_0) \to 0$ a.s. as $n \to \infty$ and $\epsilon \to 0$.*

Proof. Let us denote $\boldsymbol{u}^\epsilon = \log_{\mu_0}(\boldsymbol{\nu}^\epsilon)$ and $\boldsymbol{u}_i^\epsilon = \log_{\mu_0}(\boldsymbol{\nu}_i^\epsilon), i = 1, \ldots, n$. From Proposition 2.10,

$$d_{W_2}(\bar{\boldsymbol{\nu}}_n^\epsilon, \mathcal{M}(\boldsymbol{\nu}^\epsilon)) \to 0, \quad \text{a.s., as } n \to \infty. \tag{2.40}$$

Propositions 2.2 and 2.3 imply that

$$d_{W_2}(\mathcal{M}(\boldsymbol{\nu}^\epsilon), \mu_0) = \|\mathbb{E}(\boldsymbol{u}^\epsilon)\|_{\mu_0} = \|\mathbb{E}(\boldsymbol{u}^\epsilon) - \mathbb{E}(\boldsymbol{u})\|_{\mu_0} \leq \mathbb{E}(\|\boldsymbol{u}^\epsilon - \boldsymbol{u}\|_{\mu_0}) = \mathbb{E}(d_{W_2}(\boldsymbol{\nu}^\epsilon, \boldsymbol{\nu})),$$

hence, by Propositions 2.11 and the dominated convergence theorem,

$$d_{W_2}(\mathcal{M}(\boldsymbol{\nu}^\epsilon), \mu_0) \leq \mathbb{E}(d_{W_2}(\boldsymbol{\kappa}^\epsilon, \delta_0)) \to 0, \quad \text{as } \epsilon \to 0. \tag{2.41}$$

From (2.40) and (2.41) we conclude that $\lim_{\epsilon\to 0} \lim_{n\to\infty} \bar{\boldsymbol{\nu}}_n^\epsilon = \mu_0$.
Let $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n$, from model (2.39). Using the same arguments as above, we have

$$\begin{aligned} d_{W_2}(\bar{\boldsymbol{\nu}}_n^\epsilon, \bar{\boldsymbol{\nu}}_n) &= \left\| \frac{1}{n}\sum_{i=1}^n \boldsymbol{u}_i^\epsilon - \frac{1}{n}\sum_{i=1}^n \boldsymbol{u}_i \right\| \leq \frac{1}{n}\sum_{i=1}^n \|\boldsymbol{u}_i^\epsilon - \boldsymbol{u}_i\| = \frac{1}{n}\sum_{i=1}^n d_{W_2}(\boldsymbol{\nu}_i^\epsilon, \boldsymbol{\nu}_i) \\ &\leq \frac{1}{n}\sum_{i=1}^n d_{W_2}(\boldsymbol{\kappa}_i^\epsilon, \delta_0) \to 0 \ \text{ a.s., as } \epsilon \to 0. \end{aligned}$$

Thus, by Proposition 2.10 we conclude $\lim_{n\to\infty} \lim_{\epsilon\to 0} \bar{\boldsymbol{\nu}}_n^\epsilon = \mu_0$. □

2.7.2 GPCA in the case of non-compactly supported measures

As we have remarked, our proofs of existence of principal geodesics in Theorems 2.2 and 2.3, rely on the assumption that Ω is compact. However, this assumption is not a necessary condition as seen in Section 2.4, where we provide an example with $\Omega = \mathbb{R}$ and compute a first principal geodesic. In this section we prove the existence of first principal geodesics, for the case of $\Omega = \mathbb{R}$; see Theorem 2.6. The proof is considerable more complicated and "less elegant" than the proof of Theorem 2.2, so in this thesis we didn't further develop the techniques and arguments involved.

Before stating Theorem 2.6, let us introduce some definitions and prove some intermediate results. Recall that a subset C of a vector space X is said to be absorbing if for every $x \in X$ there exists $t > 0$ such that $x \in tC$. It is clear that C is absorbing if and only if $p(x) < \infty$ for all $x \in H$, where p is the Minkowski functional of C (see Definition A.12). Let $\mu \in W_2(\Omega)^{ac}$ and denote by p the Minkowski functional of V_μ. Let $v \in C^1(\Omega)$ and observe that $v \in tV_\mu$ if and only if $(1 + v'(x)/t) \geq 0$, for all $x \in \Omega$, μ-a.s., therefore

$$p(v) = \begin{cases} \sup_{x\in\Omega, v'(x)\leq 0} |v'(x)|, & \text{if } \{x \in \Omega, v'(x) \leq 0\} \neq \emptyset \\ 0, & \text{if } \{x \in \Omega, v'(x) \leq 0\} = \emptyset \end{cases} \tag{2.42}$$

and

$$\max(p(-v), p(v)) = \|v\|_{1,\infty} := \sup_{x\in\Omega} |v'(x)| < \infty. \tag{2.43}$$

We show that V_μ (see Definition 2.3) is not, in general, absorbing in $L^2_\mu(\Omega)$. Let $\Omega = [-\pi, \pi]$ and μ_0 be the probability on $[-\pi, \pi]$, with density $f_0(x) = 1/(2\pi)$, $x \in [-\pi, \pi]$. Let also

$u_j(x) = \sqrt{2}\sin(jx), x \in [-\pi, \pi], j \geq 1$ and observe that $u_j'(x) = \sqrt{2}j\cos(jx)$, so from (2.42) we have, $p(u_j) = \sqrt{2}j$. Let $w_j = \frac{1}{j}u_j$ and note that $w_j \to 0$ in $L^2_{\mu_0}$ as $j \to \infty$. But, by Lemma A.6(i), $p(w_j) = \frac{1}{j}p(u_j) = \sqrt{2}$, which does not converge to 0. Therefore p is not continuous and, by Lemma A.7, V_{μ_0} is not absorbing in $\text{Ł}^2_{\mu_0}$.

Based on the concept of absorbing set, we introduce the following definition.

Definition 2.16. *Let C be a subset of a Hilbert space H and let p be its Minkowski functional (see Definition A.12). We say that C is quasi-absorbing if there exists an orthonormal set $\mathcal{U} = \{x_i\}_{i\geq 1} \subset H$ such that $C \subset \operatorname{span}(\mathcal{U})$ and $p(-x_i), p(x_i) < \infty$ for all $i \geq 1$.*

Observe that if C is an absorbing set in a Hilbert space, then it is quasi-absorbing. The next proposition, shows that V_μ (see Definition 2.3) is quasi-absorbing, when $\mu \in W_2^{ac}(\Omega)$ has bounded density.

Proposition 2.13. *If $\mu \in W_2^{ac}(\Omega)$ has bounded density f, then V_μ is quasi-absorbing in $L^2_\mu(\Omega)$.*

Proof. Without loss of generality assume that $[-\pi, \pi] \subset \Omega$ and let μ_0 be the probability on Ω, with density $f_0(x) = (1/(2\pi))\mathbb{1}_{[-\pi,\pi]}(x)$, $x \in \Omega$ and cdf denoted by F_0. Let also $u_0(x) = 1$, $u_j(x) = \sqrt{2}\sin(jF_0(x))$, $v_j(x) = \sqrt{2}\cos(jF_0(x))$, $x \in \Omega, j \geq 1$, be the Fourier basis of $L^2_{\mu_0}(\Omega)$. Let us define

$$\begin{aligned} R : L^2_{\mu_0}(\Omega) &\to L^2_\mu(\Omega) \\ u &\to u \circ F_0^- \circ F, \end{aligned}$$

where F_0^- is the quantile of μ_0 (see Definition A.1) and F is the cdf of μ. Observe that R is invertible with inverse given by $R^{-1}(v) = u \circ F^- \circ F_0$, linear and $\langle Ru, Rv\rangle_\mu = \langle u, v\rangle_{\mu_0}$, $u, v \in L^2_{\mu_0}(\Omega)$. Therefore, $\tilde{u}_0(x) := R(u_0)(x) = 1$, $\tilde{u}_j := R(u_j)(x) = \sqrt{2}\sin(jF(x))$, $\tilde{v}_j(x) := R(v_j)(x) = \sqrt{2}\cos(jF(x))$, $j \geq 1$, is an orthonormal basis of $L^2_\mu(\Omega)$. From (2.43) we have $\max(p(-\tilde{u}_0), p(\tilde{u}_0)) = 0$, $\max(p(-\tilde{u}_j), p(\tilde{u}_j)) = \sup_{x\in\Omega}\sqrt{2}jf(x)|\cos(jF(x))| \leq \sqrt{2}j\|f\|_\infty$ and $\max(p(-\tilde{u}_j), p(\tilde{u}_j)) \leq 2j\|f\|_\infty$, where $\|f\|_\infty = \sup_{x\in\Omega} f(x) < \infty$. ☐

Now, let us denote by $\mathcal{T}$ the weak topology of a Hilbert space H. Given a sequence (x_n) in H we write $x_n \rightharpoonup x$ if (x_n) weakly converges to $x \in H$. We say that $g : H \to \mathbb{R}$ is $\mathcal{T}$-continuous if it is continuous with respect to $\mathcal{T}$. Similarly, we say that g is $\mathcal{T}$-lsc, if it is lower semicontinuous with respect to $\mathcal{T}$.

Lemma 2.4. *Let H be a Hilbert space and $C \subset H$ a closed and convex set containing 0. Denote by q the reciprocal Minkowski functional of C (see Definition A.12) and let $f : H \times H \to \mathbb{R}$ be defined by*

$$f(x, y) = \|y\|^2 - \langle x, y\rangle^2 + (\langle x, y\rangle - \lambda(x, y))^2. \tag{2.44}$$

where

$$\lambda(x, y) = \begin{cases} -q(-x), & \text{if } \langle x, y\rangle < -q(-x) \\ \langle x, y\rangle, & \text{if } -q(-x) \leq \langle x, y\rangle \leq q(x) \\ q(x), & \text{if } \langle x, y\rangle > q(x). \end{cases} \tag{2.45}$$

Then

(i) $f(\cdot,y)$ *is* $\mathcal{T}$*-lsc, for all* $y\in H$*, and*

(ii) for $x\in B$ *and* $y\in H$*, the function* $t\to f(tx,y)$ *is decreasing for* $t\geq 0$*. Moreover, if* $q(x)>0$ *and* $\langle x,y\rangle>0$*, then it is strictly decreasing.*

Proof. (i) From (2.45),

$$(\langle x,y\rangle-\lambda(x,y))^2=(h^2\mathbb{1}_{h>0})(-x,y)+(h^2\mathbb{1}_{h>0})(x,y),$$

where $h(x,y)=\langle x,y\rangle-q(x)$. Hence

$$f(x,y)=\|y\|^2-\langle x,y\rangle^2+(h^2\mathbb{1}_{h>0})(-x,y)+(h^2\mathbb{1}_{h>0})(x,y). \tag{2.46}$$

Observe that $h(\cdot,y)$ is $\mathcal{T}$-lsc because $-q$ is $\mathcal{T}$-lsc, by Lemma A.8, and $\langle\cdot,y\rangle$ is $\mathcal{T}$-continuous, for any $y\in H$. Recall that the maximum of two lsc functions is lsc and that the square of a positive lsc function is also lsc. Therefore $h^2(\cdot,y)\mathbb{1}_{h(\cdot,y)>0}=(h(\cdot,y)\mathbb{1}_{h(\cdot,y)>0})^2=(\max(h(\cdot,y),0))^2$ is $\mathcal{T}$-lsc. As the sum of two lsc functions is also lsc, then, from (2.46), we conclude that $f(\cdot,y)$ is $\mathcal{T}$-lsc.

(ii) Since $f(x,y)=f(-x,y)$, we can assume, without loss of generality, that $\langle x,y\rangle\geq 0$. So, from (2.46), we have

$$\begin{aligned}f(x,y)&=\|y\|^2-\langle x,y\rangle^2+(h^2\mathbb{1}_{h>0})(x,y)\\&=\|y\|^2-\langle x,y\rangle^2\mathbb{1}_{h\leq 0}(x,y)-\langle x,y\rangle^2\mathbb{1}_{h>0}(x,y)+(h^2\mathbb{1}_{h>0})(x,y)\\&=\|y\|-\langle x,y\rangle^2\mathbb{1}_{h\leq 0}(x,y)-(\langle x,y\rangle^2-h^2(x,y))\mathbb{1}_{h>0}(x,y).\\&=\|y\|-\langle x,y\rangle^2\mathbb{1}_{h\leq 0}(x,y)-(q^2(x)-2q(x)\langle x,y\rangle)\mathbb{1}_{h>0}(x,y).\end{aligned}$$

Observe that $q(tx)=\frac{1}{t}q(x)$, for $t>0$, hence,

$$f(tx,y)=\|y\|-t^2\langle x,y\rangle^2\mathbb{1}_{h\leq 0}(tx,y)-(\frac{1}{t^2}q^2(x)-2q(x)\langle x,y\rangle)\mathbb{1}_{h>0}(tx,y),$$

for $t>0$. From the previous equation,

$$f(tx,y)=\begin{cases}\|y\|^2-t^2\langle x,y\rangle^2, & \text{if } t\leq(\frac{q(x)}{\langle x,y\rangle})^{\frac{1}{2}}\\ \|y\|^2+\frac{1}{t^2}q^2(x)-2q(x)\langle x,y\rangle, & \text{if } t>(\frac{q(x)}{\langle x,y\rangle})^{\frac{1}{2}},\end{cases}$$

which is decreasing in $t\geq 0$ and strictly decreasing if $q(x)>0$ and $\langle x,y\rangle>0$. □

Lemma 2.5. *Let* H *be a Hilbert space and* $C\subset H$ *a quasi-absorbing (see Definition 2.16), closed and convex set containing* 0*. Let* ξ *be an* H*-valued random variable such that* $\xi\in C$ *a.s. and* $\mathbb{E}(\|\xi\|^2)<\infty$*. Let* $\mathrm{cost}^*:H\to\mathbb{R}_+$ *be defined by*

$$\mathrm{cost}^*(x)=\mathbb{E}(\|\xi-\Pi_{\mathrm{span}(x)\cap C}(\xi)\|^2), \tag{2.47}$$

where $\Pi_{\mathrm{span}(x)\cap C}y:=\min_{y'\in\mathrm{span}(x)\cap C}\|y-y'\|^2$ *is the projection of* $y\in H$ *onto the closed convex set* $\mathrm{span}(x)\cap C$ *and* $\mathrm{span}(x)=\{tx:t\in\mathbb{R}\}$*. Let* $B=\{x\in H:\|x\|\leq 1\}$ *and* $\partial B=\{x\in H:\|x\|=1\}$*. Then*

(i) the function $F:B\to\mathbb{R}_+$*, defined by* $F(x)=\mathbb{E}(f(x,\xi))$ *with* f *given in* (2.44)*, is* $\mathcal{T}$*-lsc.*

(ii) $F(x)=\mathrm{cost}^*(x)$*, for all* $x\in\partial B$*.*

(iii) $F(x) < \mathbb{E}(\|\xi\|^2)$ *if and only* $x \in Q$ *or* $-x \in Q$, *where*

$$Q = \{x \in B : q(x) > 0 \text{ and } \mathbb{P}(\langle x, \xi\rangle > 0) > 0\}. \tag{2.48}$$

(iv) If $\|\xi\| > 0$ *a.s, then* Q *is nonempty.*

(v) $\arg\min_{x\in\partial B} \operatorname{cost}^*(x) = \arg\min_{x\in B} F(x)$.

(vi) $\arg\min_{\partial B} \operatorname{cost}^*(x)$ *is nonempty and closed in the weak topology* $\mathcal{T}$ *of* H.

Proof. (i) It follows from Lemma 2.4(i) and Fatou's Lemma.

(ii) Take $x \in \partial B$ and $y \in H$. By Lemma A.5(ii), $\Pi_{\mathrm{span}(x)\cap C}(y) = \lambda(x, y)x$, therefore

$$\begin{aligned}\operatorname{cost}^*(x) &= \mathbb{E}(\|\xi - \lambda(x, \xi)x\|^2) = \mathbb{E}(\|\xi\|^2 - 2\lambda^2(x, \xi)\langle x, \xi\rangle + \lambda^2(x, \xi))\\ &= \mathbb{E}(\|\xi\|^2 - \langle x, \xi\rangle^2 + (\langle x, \xi\rangle - \lambda(x, \xi))^2 = \mathbb{E}(f(x, \xi)),\end{aligned}$$

so we obtain the result.

(iii) Let $x \in Q$ and $A_x := \{y \in H : \langle x, y\rangle > 0\}$. From (2.46), for any $y \in A_x$,

$$\begin{aligned}f(x, y) &= \|y\|^2 - \langle x, y\rangle^2 + (h^2\mathbb{1}_{h>0})(x, y)\\ &= \|y\|^2 - \langle x, y\rangle^2\mathbb{1}_{h\le 0}(x, y) - \langle x, y\rangle^2\mathbb{1}_{h>0}(x, y) + (h^2\mathbb{1}_{h>0})(x, y)\\ &= \|y\| - \langle x, y\rangle^2\mathbb{1}_{h\le 0}(x, y) - (\langle x, y\rangle^2 - h^2(x, y))\mathbb{1}_{h>0}(x, y).\end{aligned}$$

As $\langle x, y\rangle > 0$ and $q(x) > 0$, we have

$$f(x, y) = \begin{cases}\|y\|^2 - \langle x, y\rangle^2 < \|y\|^2, & \text{if } h(x, y) \le 0\\ \|y\|^2 - (\langle x, y\rangle^2 - h^2(x, y)) < \|y\|^2, & \text{if } h(x, y) > 0,\end{cases}$$

therefore $f(x, y) < \|y\|^2$. Let $\mathbb{P}_\xi(A) = \mathbb{P}(\xi \in A), A \in \mathcal{B}(H)$. Since $\mathbb{P}_\xi(A_x) > 0$ we have

$$\int_{A_x} f(x, y)d\mathbb{P}_\xi(y) < \int_{A_x} \|y\|^2 d\mathbb{P}_\xi(y).$$

From Lemma (2.4)(ii), we have that $f(x, y) \le f(0, y) = \|y\|^2$ for all $y \in H$, so we obtain

$$\int_{H} f(x, y)d\mathbb{P}_\xi(y) < \int_{H} \|y\|^2 d\mathbb{P}_\xi(y),$$

i.e. $F(x) < \mathbb{E}(\|\xi\|^2)$. Now, if we assume that $-x \in Q$ then $F(-x) < \mathbb{E}(\|\xi\|^2)$, but $F(x) = F(-x)$, so we obtain $F(x) < \mathbb{E}(\|\xi\|^2)$.To prove the other implication, assume that $-x, x \notin Q$. From (2.46) it is easy to check that $f(x, y) = \|y\|^2$ for all $y \in H$ a.s.

(iv) As C is quasi-absorbing, there exists an orthonormal set $\{x_i\}_{i\ge 1} \subset H$ such that $\|y\|^2 = \sum_{i\ge 1}\langle x_i, y\rangle^2$ and $q(-x_i), q(x_i) > 0$ for all $i \ge 1$. By hypothesis $\mathbb{P}_\xi(y \in H : \|y\|^2 > 0) > 0$, hence

$$0 < \mathbb{P}_\xi(y \in H : \sum_{i\ge 1}\langle x_i, y\rangle^2 > 0) \le \sum_{i\ge 1} \mathbb{P}_\xi(y \in H : \langle x_i, y\rangle^2 > 0).$$

From the previous inequality, there exists $j \ge 1$ such that $\mathbb{P}_\xi(y \in H : \langle x_j, y\rangle > 0) > 0$ or $\mathbb{P}_\xi(y \in H : \langle -x_j, y\rangle > 0) > 0$. As $q(x_j), q(-x_j) > 0$, we obtain that x_j or $-x_j$ belongs to Q.

(v) Take $x^* \in \arg\min_{x \in B} F(x)$ and let us show that $x^* \in \partial B$. If $\xi = 0$ a.s, then $F(x) = \mathrm{cost}^*(x) = 0$, for all $x \in H$ and the result holds. So, let us assume that $\|\xi\| > 0$ a.s. From (iii) and (iv), there exists $x \in B$ such that $F(x) < \mathbb{E}(\|\xi\|^2)$, hence $F(x^*) < \mathbb{E}(\|\xi\|^2)$. As $F(x^*) = F(-x^*)$, from (iii) we can assume, without loss of generality that $x^* \in Q$, so $q(x^*) > 0$ and $\mathbb{P}_\xi(A_{x^*}) > 0$, where $A_{x^*} = \{y \in H : \langle x^*, y \rangle > 0\}$. Observe that $t \to f(tx^*, y)$ is decreasing for all $y \in H$, and strictly decreasing for all $y \in A_{x^*}$, thanks to Lemma 2.4(ii). As $\mathbb{P}_\xi(A_{x^*}) > 0$, we conclude that $t \to F(tx)$ is strictly decreasing, for $t \geq 0$, which implies that $x^* \in \partial B$. Finally, the result follows from (ii).

(vi) As F is $\mathcal{T}$-lsc and as B is $\mathcal{T}$-compact, we obtain the result. $\square$

Let us state the main result of this section.

Theorem 2.6. *Let $\Omega = \mathbb{R}$, $\mu \in W_2^{ac}(\Omega)$ a reference measure with bounded density, $\boldsymbol{\nu} \in W_2(\Omega)$ a square-integrable random measure (see Definition 2.5) and $k \geq 1$ an integer. Then $\mathcal{G}_{\mu,1}$, from Definition 2.9, is nonempty.*

Proof. The idea is to apply Lemma 2.5(vi), with $H = L^2_\mu(\Omega)$, $C = V_\mu$ and $\xi = \log_\mu(\boldsymbol{\nu})$. From Proposition 2.2 and as $\boldsymbol{\nu}$ is square-integrable, we have that $\mathbb{E}\left(\|\log_\mu(\boldsymbol{\nu})\|_\mu^2\right) = \mathbb{E}\left(d^2_{W_2}(\boldsymbol{\nu}, \mu)\right) < \infty$. On the other hand, from Propositions 2.1 and 2.13, V_μ is a quasi-absorbing, closed and convex set containing 0. Then the result follows from Proposition 2.7 and Lemma 2.5(vi). $\square$

2.7.3 GPCA based on kernel density estimation

So far we have assumed that the input data consist of n probability measures $\nu_1, \ldots, \nu_n$ belonging to $W_2(\Omega)$. However, in many applications we have access only to random observations from these measures. In such case we propose to carry out a density estimation step before performing the GPCA. For $i = 1, \ldots, n$, let $X_{i,j}, j = 1, \ldots, m_i$ be $m_i \in \mathbb{N}$ iid random variables from ν_i, with density f_i. Let $\hat{\nu}_i$ be the probability measure with density $\hat{f}_i$, obtained, for example, by kernel smoothing as

$$\hat{f}_i(x) = \frac{1}{m_i h} \sum_{j=1}^{m_i} K\left(\frac{x - X_{i,j}}{h}\right), \; x \in \mathbb{R}, \; i = 1, \ldots, n,$$

where $K : \mathbb{R} \to \mathbb{R}$ is probability density kernel and $h > 0$ is a bandwidth parameter. We define the method of GPCA on $X_{i,j}, j = 1, \ldots, m_i, i = 1, \ldots, n$, as the GPCA on $\hat{\nu}_1, \ldots, \hat{\nu}_n$. In this context, the main question we formulate has to do with the consistency of this method in the asymptotic setting $m_i \to \infty$, $i = 1, \ldots, n$ and $n \to \infty$.

Chapter 3

Geometric principal components analysis of images

3.1 Introduction

In many applications, observations are in the form of a set of n gray-level images $y_1, \ldots, y_n$ (e.g. in geophysics, in biomedical imaging, or in signal processing for neurosciences), which can be considered as square-integrable functions on a domain Ω, a convex subset of $\mathbb{R}^d$. Such data are generally two or three dimensional images. In many situations the observed images share the same structure. This may lead to the assumption that these observations are random elements, which vary around the same mean pattern. Estimating such a mean pattern and characterizing the modes of individual variations around this common shape, is of fundamental interest. Principal component analysis (PCA) is a widely used method for estimating the variations in intensity of images around the usual Euclidean mean $\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$. However, such data typically exhibit not only a classical source of photometric variability (a pixel intensity changes from one image to another) but also a (less standard) geometric source of variability which cannot be recovered by standard PCA (see Section 1.1.2).

The goal of this chapter is to provide a general framework for analyzing the geometric variability of images through the use of deformation operators that can be parametrized by elements in a Hilbert space. This setting leads to a simple algorithm for estimating the main modes of geometric variability of images, and we prove the consistency of this approach in statistical deformable models.

3.1.1 An overview of PCA-like methods for analyzing geometric variability

A standard approach for analyzing geometric variability is to use registration. This well-known approach consists in computing geometric transforms of a set of images $y_1, \ldots, y_n$, so that they can be compared. Then, the main idea for estimating the geometric variability of such data is to apply classical PCA to the resulting transformation parameters after registration and not to the images themselves. This approach is at the core of several methods for estimating the geometrical variability of images, that we choose to call *geometric PCA methods* in what follows.

In [Rueckert et al., 2003], a linear and finite dimensional space of non-rigid transformations is considered as the admissible space of deformations onto which a standard PCA is carried out. This is the so-called statistical deformation model that is inspired by Cootes active shape models [Cootes et al., 1995]. An important limitation of this approach is the lack of invertibility of the deformations in such models. In several cases, the invertibility is a desirable property from the point of view of physical (for instance when analyzing geometric variability of a determined organ) and mathematical modelling. Within the context of a linear space of deformations, the non-invertibility issue had been addressed by enforcing the positiveness of the Jacobian determinant [Haber and Modersitzki, 2006; Sdika, 2008]. However, such methods are unsuited for further statistical analysis, as statistical procedures on resulting transformation parameters (such the empirical Euclidean mean), do not lead to invertible transforms. Moreover, the inverse transforms do not belong to the initial space of transformations.

More recently, diffeomorphisms have been used for modeling geometric transformations between images in the context of Grenander's pattern theory [Beg et al., 2005; Hernandez et al., 2009; Ashburner, 2007]. In this framework, the set of admissible diffeomorphic transformations is considered as a Riemaniann manifold, and thus first and second order statistics analysis on manifolds [Fletcher et al., 2004; Pennec, 2006] can be applied for performing statistical analysis of diffeomorphic deformations [Arsigny et al., 2006; Bossa et al., 2007; Hernandez et al., 2007; Wang et al., 2007]. Such diffeomorphisms are constructed as solution of an ordinary differential equation (ODE), governed by a time dependent vector field belonging to a linear space, see e.g. [Beg et al., 2005]. In this way, it is possible to build sets of diffeomorphisms that have mathematical properties very similar to Lie groups. This approach, which leads to the representation of the geometric variability of images through the use of standard PCA on the elements in the "Lie algebra" of vector fields, is discussed in details in [Trouvé and Younes, 2005; Trouvé and Younes, 2005; Younes, 2010]. In particular, the optimal diffeomorphism after the registration of two images can be fully characterized by the initial point in time (or equivalently by the initial momentum) of the associated time dependent vector field. This key property, called momentum conservation, allows to carry out PCA on the Hilbert space of initial momentums, see e.g. [Wang et al., 2007].

A particular sub-class of diffeomorphic deformations is the set of diffeomorphisms generated by an ODE governed by stationary vector fields. In this way, diffeomorphisms are directly characterized by vector fields belonging to a Hilbert space, and thus a standard statistical analysis such as PCA can be carried out on the vector fields computed after image registration. Compared to the case of diffeomorphisms generated by non-stationary vector fields, the resulting deformations do not have the same desirable properties in term of group structure. Nevertheless, the natural parametrization of these deformations by a linear space, make them well suited for the purpose of geometric PCA. Moreover, the computational cost of the registration step when using stationary vector fields is considerably smaller, while keeping comparable accuracy according to the experimental results reported in [Ashburner, 2007; Hernandez et al., 2009]. The properties of diffeomorphisms generated by stationary vector fields also allow simple and fast computations of the diffeomorphism associated with a vector field and vice versa. Hence, PCA methods for manifolds can be implemented for analyzing geometric variability of diffeomorphic deformations generated by stationary vector fields, see [Arsigny et al., 2006; Bossa et al., 2007; Hernandez et al., 2007].

3.1.2 Main contributions and organization of this chapter

In Section 3.2, we propose for analyzing geometric PCA methods using a general framework where the spatial deformation operators (to represent geometric variability) are invertible and can be parametrized by elements of a Hilbert space. For estimating geometric variability, as it is done in geometric PCA methods, we use a preliminary registration step. Then, we apply classical PCA on Hilbert spaces to the resulting parameters representing the deformations after registration. The main contributions of this chapter are then the following ones. First, for the application considered in this chapter and for algorithmic purposes, we use diffeomorphic deformations parametrized by stationary vector fields that are expanded into a finite dimensional basis of a linear space of functions. We show that such deformations are well suited for the analysis of handwritten digits. In this setting, an important and new contribution is to provide an automatic method for choosing the regularization parameter that represents the usual balance between the regularity of the spatial deformations and the quality of images alignment. Secondly, in Section 3.3, we consider the problem of building geometric PCA methods that are consistent. To the best of our knowledge, this issue has not been very much studied in the literature on geometric PCA. We discuss the appropriate asymptotic setting for such an analysis and prove the consistency of our approach in statistical deformable models. We conclude the chapter in Section 3.4 by a short discussion. Almost all proofs are gathered in Appendix B.

3.2 Geometric PCA

For convenience, we prefer to present the ideas of geometric PCA under the assumption that the images are observed on a continuous domain Ω. In practice, such data are obviously observed on a discrete set of time-points or pixels. However, assuming that the data are random elements of $L^2(\Omega)$ is more convenient for dealing with the statistical aspects of an inferential procedure, as it avoids the treatment of the bias introduced by any discretization of the domain Ω. We refer to Section 3.3 for a detailed discussion on this point.

3.2.1 Grenander's pattern theory of deformable templates

Following the ideas of Grenander's pattern theory (see [Grenander and Miller, 2007] for a recent overview), one may consider that the data $y_1, \ldots, y_n$ are obtained through the deformation of the same reference image. In this setting, images are treated as points in $L^2(\Omega)$ and the geometric variations of the images are modeled by the action of Lie groups on the domain Ω. Recently, there has been a growing interest in Lie groups of transformations for modeling the geometric variability of images (see e.g. [Beg et al., 2005; Trouvé and Younes, 2005; Trouvé and Younes, 2005; Younes, 2010] and references therein), and applications are numerous, in particular, in biomedical imaging, see e.g. [Fletcher et al., 2004; Joshi et al., 2004].

Grenander's pattern theory leads to the construction of non-Euclidean distances between images. In this chapter we propose for modeling geometric variability through the use of deformation operators (acting on Ω), that are parameterized by a separable Hilbert space $\mathcal{V}$, with inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$. We also assume that Ω is equipped with a metric denoted by d_Ω.

Definition 3.1. *Let $\mathcal{V}$ be a Hilbert space. A deformation operator parameterized by $\mathcal{V}$ is a mapping $\varphi : \mathcal{V} \times \Omega \to \Omega$ such that, for any $v \in \mathcal{V}$, the function $x \mapsto \varphi(v, x)$ is a homeomorphism on Ω. Moreover, $\varphi(0, \cdot)$ is the identity on Ω and, for any $v \in \mathcal{V}$, there exists $v^* \in \mathcal{V}$ such that $\varphi^{-1}(v, \cdot) = \varphi(v^*, \cdot)$.*

In this chapter we study, as illustrative examples of deformation operators, the cases of translations, rigid deformations and non-rigid deformations generated by stationary vector fields.

Translations: Let $\Omega = [0, 1)^d$, for some integer $d \geq 1$ and $\mathcal{V} = \mathbb{R}^d$. Let also

$$\varphi(v, x) = (\operatorname{mod}(x_1 + v_1, 1), \ldots, \operatorname{mod}(x_d + v_d, 1)), \tag{3.1}$$

with

$$\varphi^{-1}(v, x) = (\operatorname{mod}(x_1 - v_1, 1), \ldots, \operatorname{mod}(x_d - v_d, 1)),$$

for all $v = (v_1, \ldots, v_d) \in \mathbb{R}^d$ and $x = (x_1, \ldots, x_d) \in \Omega$, where $\operatorname{mod}(a, 1)$ denotes the modulo operation between $a \in \mathbb{R}$ and 1. Clearly, $\varphi(0, \cdot)$ is the identity in Ω and $v^* = -v$. Moreover, it can be shown (see Section 3.3.3) that $\varphi(v, \cdot)$ is an homeomorphism.

Rigid deformations of 2D images: Let $\Omega = \mathbb{R}^2$ and $\mathcal{V} = \mathbb{R} \times \mathbb{R}^2$. Let also

$$\varphi(v, x) = R_\alpha x + b,$$

with

$$\varphi^{-1}(v, x) = R_{-\alpha}(x - b),$$

for all $v = (\alpha, b) \in \mathbb{R} \times \mathbb{R}^2$, and $x \in \mathbb{R}^2$, where R_α is the rotation matrix of angle α and $b \in \mathbb{R}^2$ defines a translation. Observe that, $\varphi((0, 0), \cdot)$ is the identity in Ω and $v^* = (-\alpha, -R_{-\alpha} b)$. Clearly, $\varphi(v, \cdot)$ is an homeomorphism.

Diffeomorphic deformations generated by stationary vector fields: Let $\Omega = [0, 1]^d$, for some integer $d \geq 1$, and $\mathcal{V}$ a separable Hilbert space of smooth vector fields, such that $\mathcal{V}$ is continuously embedded on $C_0^1(\Omega)$, the space of functions $v : \Omega \to \mathbb{R}^d$ which are continuously differentiable and such v and its derivatives vanish at the boundary of Ω. For $x \in \Omega$ and $v \in \mathcal{V}$, define $\varphi(v, x)$ as the solution at time $t = 1$ of the following ODE

$$\frac{d\phi_t}{dt} = v(\phi_t), \tag{3.2}$$

with initial condition $\phi_0 = x \in \Omega$. It is well known (see e.g. [Younes, 2010]) that, for any $v \in \mathcal{V}$, the function $x \mapsto \varphi(v, x)$ is a C^1 diffeomorphism on Ω. The inverse of $x \mapsto \varphi(v, x)$ is given by $x \mapsto \varphi(-v, x)$, and thus $v^* = -v$. Hence, $\varphi(v, \cdot)$ satisfies the conditions in Definition 3.1.

3.2.2 Registration

Registration is a widely used method in image processing that consist in geometric transforms of a set of images $y_1, \ldots, y_n \in L^2(\Omega)$, so that they be compared. This method can be described as an optimization problem, which amounts to minimizing a dissimilarity functional between images.

Definition 3.2 (Dissimilarity functional)**.** *Let φ be a deformation operator, as described in Definition 3.1, $\boldsymbol{v} = (v_1, \ldots, v_n) \in \boldsymbol{\mathcal{V}} := \mathcal{V}^n$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$, with $y_i \in L^2(\Omega)$, $i = 1, \ldots, n$.*
(a) The template dissimilarity functional corresponding to $\boldsymbol{v}$, $\boldsymbol{y}$ and $f \in L^2(\Omega)$ is defined as

$$M^t(\boldsymbol{v}, \boldsymbol{y}, f) := \frac{1}{n} \sum_{i=1}^{n} \int_{\Omega} \Big(y_i(\varphi(v_i, x)) - f(x) \Big)^2 dx. \tag{3.3}$$

(b) The groupwise dissimilarity functional corresponding to $\boldsymbol{v}$ and $\boldsymbol{y}$ is defined as

$$M^g(\boldsymbol{v}, \boldsymbol{y}) := \frac{1}{n} \sum_{i=1}^{n} \int_{\Omega} \Big(y_i(\varphi(v_i, x)) - \frac{1}{n} \sum_{j=1}^{n} y_j(\varphi(v_j, x)) \Big)^2 dx. \tag{3.4}$$

Template registration of the images $y_1, \ldots, y_n$, onto some known template $f \in L^2(\Omega)$, is defined as the problem of minimizing the criterion given by the dissimilarity functional (3.3), with respect to $\boldsymbol{v}$ in

$$\boldsymbol{\mathcal{V}}_\mu := \{\boldsymbol{v} = (v_1, \ldots, v_n),\ v_i \in \mathcal{V}_\mu\},$$

where $\mathcal{V}_\mu := \{v \in \mathcal{V} : \|v\| \leq \mu\}$, for some regularization parameter $\mu \geq 0$. Note that imposing the constraint $\|v_i\| \leq \mu$ allows to explicitly control the norm of the vector v_i, which is generally proportional to the distance between the deformation $\varphi(v_i, \cdot)$ and the identity. The choice of μ is obviously of primary importance. A data-based procedure for its calibration is thus discussed in details in Section 3.2.3.

On the other hand, groupwise registration of $y_1, \ldots, y_n$ is defined as the problem of minimizing the functional (3.4) with respect to v in $\boldsymbol{\mathcal{U}} \subseteq \boldsymbol{\mathcal{V}}_\mu$. Two possible choices for $\boldsymbol{\mathcal{U}}$, defined in terms of linear constraints on $\boldsymbol{v}$, are

$$\boldsymbol{\mathcal{U}}_0 := \Big\{\boldsymbol{v} \in \boldsymbol{\mathcal{V}}_\mu,\ \sum_{m=1}^{n} v_m = 0\Big\} \quad \text{and} \quad \boldsymbol{\mathcal{U}}_1 := \Big\{\boldsymbol{v} \in \boldsymbol{\mathcal{V}}_\mu,\ v_1 = 0\Big\}. \tag{3.5}$$

Choosing $\boldsymbol{\mathcal{U}} = \boldsymbol{\mathcal{U}}_0$ amounts to imposing that the deformation parameters $(v_1, \ldots, v_n)$, used to align the data, have an empirical mean equal to zero, while taking $\boldsymbol{\mathcal{U}} = \boldsymbol{\mathcal{U}}_1$ corresponds to choosing y_1 as a reference template onto which $y_2, \ldots, y_n$ will be aligned.

Geometric PCA applied to the images $\boldsymbol{y} = (y_1, \ldots, y_n)$ is the following two step procedure. In the first step, one applies either a template or a groupwise registration, which leads to the computation of

$$\hat{\boldsymbol{v}} = \hat{\boldsymbol{v}}_\mu \in \arg\min_{\boldsymbol{v}\in\boldsymbol{\mathcal{V}}_\mu} M^t(\boldsymbol{v}, \boldsymbol{y}, f) \text{ or } \hat{\boldsymbol{v}} = \hat{\boldsymbol{v}}_\mu \in \arg\min_{\boldsymbol{v}\in\boldsymbol{\mathcal{U}}} M^g(\boldsymbol{v}, \boldsymbol{y}). \tag{3.6}$$

In the second step, a standard PCA (see Section 1.1) is carried out on $\hat{\boldsymbol{v}} = (\hat{v}_1, \dots, \hat{v}_n)$, based on the following covariance operator

$$\hat{K}_n v = \frac{1}{n}\sum_{i=1}^{n} \langle \hat{v}_i - \overline{v}_n, v\rangle(\hat{v}_i - \overline{v}_n), \text{ for } v \in \mathcal{V}, \tag{3.7}$$

with $\overline{v}_n = \frac{1}{n}\sum_{i=1}^{n} \hat{v}_i$. This operator is compact and so it admits the decomposition

$$\hat{K}_n v = \sum_{k\in\mathbb{K}} \hat{\lambda}_k \langle v, \hat{\phi}_k\rangle \hat{\phi}_k, \tag{3.8}$$

where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq 0$ and $(\hat{\phi}_k)_{k\in\mathbb{K}}$ are the eigenvalues and orthonormal eigenvectors of $\hat{K}_n$ and $\mathbb{K} = \{1, \dots, dim(\mathcal{V})\}$, if $dim(\mathcal{V}) < \infty$ or $\mathbb{K} = \mathbb{N}$ otherwise. We now state the definition of geometric PCA of a set of images.

Definition 3.3 (Geometric PCA). *Let φ be a deformation operator parametrized by $\mathcal{V}$, as described in Definition 3.1. Let $(\hat{\lambda}_k, \hat{\phi}_k)_{k\in\mathbb{K}}$ be the eigenvalues and orthonormal eigenvectors of the operator $\hat{K}_n$ in (3.7). For $k \in \mathbb{K}$, the k-th empirical mode of geometric variation of the data $y_1, \dots, y_n$ is the homeomorphism $\hat{\psi}_k : \Omega \to \Omega$ defined by*

$$\hat{\psi}_k(x) = \varphi^{-1}(\overline{v}_n + \sqrt{\hat{\lambda}_k}\hat{\phi}_k, x), \; x \in \Omega. \tag{3.9}$$

We also define $\hat{\psi}_{k,\rho}(x) = \varphi^{-1}(\overline{v}_n + \rho\sqrt{\hat{\lambda}_k}\hat{\phi}_k, x)$, where $\rho \in \mathbb{R}$ is a weighting value.

After the registration step we obtain a set of deformed images $y_1 \circ \varphi(\hat{v}_1, \cdot), \dots, y_n \circ \varphi(\hat{v}_n, \cdot)$, each of them aligned either with respect to f, in the case of template registration, or with respect to $\hat{f} := \frac{1}{n}\sum_{j=1}^{n} y_j(\varphi(\hat{v}_j, x))$, in the case of groupwise registration. Hence, in the case of template registration, $f \circ \hat{\psi}_k$ can be used to visualize the k-th mode of geometric variation of the data. Similarly, in the case of groupwise registration, one uses $\hat{f} \circ \hat{\psi}_k$. Note that $\hat{f}$ can be interpreted as a mean pattern image. Moreover, the computation of $\hat{f}$ is closely related to the notion of Fréchet mean of images, recently studied in [Bigot and Charlier, 2011], from a statistical point of view.

We conclude this section by providing sufficient conditions ensuring the existence of a solution of the registration problem and thus the existence of $(\hat{\lambda}_k, \hat{\phi}_k)_{k\in\mathbb{K}}$ in Definition 3.3. In the following proposition we say that a condition holds for all $x \in \Omega$ a.e., if it holds for all $x \in \Omega$ up to a set of null Lebesgue-measure.

Proposition 3.1. *Let φ be a deformation operator, as described in Definition 3.1 and $f, y_1, \dots, y_n \in L^2(\Omega)$. Assume also that*

(i) if (v_k) is a sequence in $\mathcal{V}_\mu$ that weakly converges to $v \in \mathcal{V}_\mu$, then $\varphi(v_k, x)$ converges to $\varphi(v, x)$, for all $x \in \Omega$ a.e.

(ii) $y_i, \dots, y_n$ are continuous in every $x \in \Omega$ a.e.

Then $\hat{\boldsymbol{v}}$ defined in (3.6) *exists, for the cases of template and groupwise registration.*

Proof. We prove the case of template registration of an image $y \in L^2(\Omega)$ onto f. The case of template registration, with $n > 1$ and groupwise registration, can be shown using similar arguments. Take a minimizing sequence (v_k) in $\mathcal{V}_\mu$ for the template dissimilarity functional $M^t(\cdot, y, f)$, i.e. $M^t(v_k, y, f)$ decrease to $\inf_{u \in V} M^t(u, y, f)$. Since (v_k) is bounded, we can extract a subsequence, that we will name as the original, that weakly converges to some $v \in \mathcal{V}_\mu$. From conditions (i),(ii) and Fatou's lemma, we have $M^t(v, y, f) \leq \liminf_{k \to \infty} M^t(v_k, y, f) = \inf_{u \in V} M^t(v, y, f)$ and we conclude that $M^t(v, y, f) = \inf_{u \in V} M^t(u, y, f)$. □

Let us check that condition (i) in Proposition 3.1 is satisfied for the 3 examples of deformation operators defined in Section 3.2.1. Recall that in finite-dimensional spaces, weak and strong convergence coincide, thus (i) follows directly in the case of rigid deformations and from the fact that $\mathrm{mod}(\cdot, 1)$ is continuous in $\mathbb{R} \backslash \mathbb{Z}$, in the case of translations. In the case of non-rigid deformations, generated by stationary vector fields, such condition is guaranteed by Theorem 8.11 in [Younes, 2010].

3.2.3 Numerical implementation and application of geometric PCA to handwritten digits data

In this section we explain in detail the implementation of geometric PCA, in the case of groupwise registration, using the class of diffeomorphic deformations described in Section 3.2.1. The method is applied to a set of $n = 30$ images, defined on the domain $\Omega = [0, 1]^2$, taken from the Mnist database of handwritten digits [LeCun et al., 1998].

Specification of the Hilbert space of parameter $\mathcal{V}$

We choose $\mathcal{V}$ as the vector space of functions from Ω to $\mathbb{R}^2$, generated by a B-Splines basis of functions, because they have good properties for approximating continuous functions and implementing efficient computations [Unser et al., 1993a,b]. Let $\{b_k : \Omega \to \mathbb{R}, k = 1, \ldots, p\}$ denote a set of bi-dimensional tensor product B-Splines, with knots defined on a regular grid of Ω, and p some integer whose choice has to be discussed. We define $\mathcal{V}$ as the space of vector fields of the form $v = \sum_{k=1}^{p} \tilde{v}_k b_k$, where $\tilde{v}_k = (\tilde{v}_k^{(1)}, \tilde{v}_k^{(2)}) \in \mathbb{R}^2, k = 1, \ldots, p$. We denote by $v^{(1)}, v^{(2)} : \Omega \to \mathbb{R}$ the coordinates of $v \in \mathcal{V}$, i.e., $v(x) = (v^{(1)}(x), v^{(2)}(x))$ for $x \in \Omega$. Note that the dimension of $\mathcal{V}$ is $2p$ and that a basis is given by

$$\{(b_1, 0), \ldots, (b_p, 0), (0, b_1), \ldots, (0, b_p)\}. \tag{3.10}$$

We endow $\mathcal{V}$ with the inner product

$$\langle u, v \rangle := \langle u^{(1)}, v^{(1)} \rangle_L + \langle u^{(2)}, v^{(2)} \rangle_L, \ u, v \in \mathcal{V},$$

where $\langle u^{(1)}, v^{(1)} \rangle_L := \int_\Omega L u^{(1)}(x) L v^{(1)}(x) dx$, $\langle u^{(2)}, v^{(2)} \rangle_L := \int_\Omega L u^{(2)}(x) L v^{(2)}(x) dx$ and L is a differential operator. As suggested in [Beg et al., 2005] we take $L = \gamma I + \alpha \Delta$, where I is the identity operator, Δ is the Laplacian operator and γ, α are positive scalars. By using the basic properties of differentiation and integration of B-Splines [Unser et al., 1993a], we derive an explicit formula for computing the inner product in $\mathcal{V}$, that can be implemented using convolution filters. By an adequate design of the B-Spline grid, we ensure that the values of v and its derivatives are zero at the boundary of Ω.

Minimization of the dissimilarity functional M^g

In the case of groupwise registration, we minimize the dissimilarity functional (3.4) over the set $\mathcal{U}_0$ defined in (3.2.2). Thanks to the above choice for $\mathcal{V}$, the minimization of the criterion (3.4) has to be performed over a subset of $\mathbb{R}^{2p}$. In order to take into account the constrains $\|v_i\| \leq \mu,\ 1 \leq i \leq n$, we use a logarithmic barrier approach to obtain an approximate solution. Then, for the minimization, we use a gradient descent algorithm, with an adaptive step. Such an algorithm requires the computation of the deformation operator $\varphi : \mathcal{V} \times \Omega \to \Omega$ and its gradient, with respect to the coefficients $\tilde{v}_k = (\tilde{v}_k^{(1)}, \tilde{v}_k^{(2)})$ that parameterize the vector field v. In the case of diffeomorphic deformation operators, $\varphi(v, x)$ corresponds to the solution at time $t = 1$ of the ODE (3.2). We solve the ODE using a forward Euler integration scheme. For a comparison of different methods for solving such ODE, we refer to [Bossa et al., 2008]. It can be shown (see Lemma 2.1 in [Beg et al., 2005]), that the gradient of φ with respect to the $\tilde{v}_k = (\tilde{v}_k^{(1)}, \tilde{v}_k^{(2)})$'s has a closed-form expression.

Spectral decomposition of the empirical covariance operator

Let $\hat{v}_1, \ldots, \hat{v}_n$ be the vector fields in $\mathcal{V}$ obtained after the registration step described above. Recall that the empirical covariance operator of the $\hat{v}_i$'s is defined as $\hat{K}_n v = \frac{1}{n}\sum_{i=1}^{n} \langle \hat{v}_i - \overline{v}_n, v\rangle(\hat{v}_i - \overline{v}_n),\ v \in \mathcal{V}$. In what follows, we describe how for performing the spectral decomposition of $\hat{K}_n$.

Let $\tilde{v}_i = (\tilde{v}_i^{(1)}, \tilde{v}_i^{(2)})$ with $\tilde{v}_i^{(1)} = (\tilde{v}_{i,1}^{(1)}, \ldots, \tilde{v}_{i,p}^{(1)})$ and $\tilde{v}_i^{(2)} = (\tilde{v}_{i,1}^{(2)}, \ldots, \tilde{v}_{i,p}^{(2)})$ being the coefficients of $\hat{v}_i$ with respect to the base (3.10), i.e. $\hat{v}_i = \sum_{k=1}^{p} (\tilde{v}_{i,k}^{(1)}, \tilde{v}_{i,k}^{(2)}) b_k$. We identify the Hilbert space $\mathcal{V}$ with $\mathbb{R}^{2p}$, endowed by the inner product

$$\langle \tilde{u}, \tilde{v}\rangle := \tilde{u}\Sigma\tilde{u}^t,\ \tilde{u}, \tilde{v} \in \mathbb{R}^{2p}, \tag{3.11}$$

where Σ is a $2p \times 2p$ matrix with entries $\Sigma_{j,k} = \Sigma_{j+p,k+p} := \langle b_j, b_k\rangle_L$ for $j, k = 1, \ldots, p$ and $\Sigma_{j,k} := 0$ in the other cases. Hence, the operator $\hat{K}_n$ can be identified with a $2p \times 2p$ matrix $\tilde{K}_n$, given by

$$\tilde{K}_n := \frac{1}{n}\tilde{v}\tilde{v}^t\Sigma,$$

where $\tilde{v}$ is the $2p \times n$ matrix with i-th column equal to $\tilde{v}_i^t - \frac{1}{n}\sum_{j=1}^{n} \tilde{v}_j^t$. The matrix Σ is symmetric, hence admits a diagonalization $\Sigma = P\Lambda P^t$ with Λ a diagonal matrix and $P^tP = PP^t = I$. The idea now is to reduce the problem to a standard diagonalization of the symmetric matrix $M := \frac{1}{n}\Lambda^{\frac{1}{2}} P^t\tilde{v}\tilde{v}^t P\Lambda^{\frac{1}{2}}$, namely we find the decomposition $M = WDW^t$ with D a diagonal matrix and $W^tW = WW^t = I$. We obtain the following spectral decomposition of $\tilde{K}_n$ relative to the inner product (3.11)

$$\tilde{K}_n = UDU^t\Sigma,$$

where $U := P^t W$. Remark that the columns of U are orthonormal vectors in $\mathbb{R}^{2p}$ relative to the inner product (3.11). Indeed, it holds that $U^t \Sigma U = I$. Finally, we define $\hat{\lambda}_k$ as the k-th elements of the diagonal matrix D and we let $\hat{\phi}_k := \sum_{k=1}^{p} (U_{k,k}, U_{k+p,k}) b_k$, for $k = \{1, \ldots, 2p\}$. It can be checked that $(\hat{\phi}_k)_{k=1}^{2p}$ are orthonormal vectors of $\mathcal{V}$, and we thus obtain that $\hat{K}_n v = \sum_{k=1}^{2p} \hat{\lambda}_k \langle v, \hat{\phi}_k \rangle_L \hat{\phi}_k$. If we assume that $\hat{\lambda}_1 \geq, \ldots, \geq \hat{\lambda}_{2p}$, then

$$\hat{\psi}_{k,\rho} = \varphi^{-1}(\overline{v}_n + \rho \sqrt{\hat{\lambda}_k} \hat{\phi}_k, \cdot)$$

is the k-th empirical mode of geometric variation, according to Definition 3.3.

Choice of the regularization parameter μ and application of geometric PCA to handwritten digits data

We now describe the application of geometric PCA to handwritten digits, taken from the Mnist database [LeCun et al., 1998], based on the numerical framework we have described so far. We also discuss the problem of automatically selecting the regularization parameter μ and we finally illustrate the benefits of geometric PCA over standard PCA.

For the B-Spline base of $\mathcal{V}$, we choose a B-Spline degree equal to 3, as it provides a good trade-off between smoothness and the size of the support. The number of B-Spline knots is $p = 81$ arranged in a 9×9 regular grid. Such value of p provides a fine B-Spline grid with respect to an image size of 28×28. For defining the differential operator L, we take $\gamma = 100$ and $\alpha = 1$. Note that γ is much larger than α in order to compensate for the scaling factor associated with the inter knot spacing.

For each available digit (from 0 to 9) we determine the regularization parameter μ experimentally, by trying to obtain a good balance between the regularity of the vector fields and the matching of the images during the preliminary registration step. Our approach is inspired by the classical L-curve method in inverse problems. More precisely, for each digit, we took $n = 30$ images and we carried out registration on each of these image sets. In this database, for each digit, one observes a large source of geometrical variability that can be modeled by diffeomorphic deformations. We proceed by groupwise registration, as there are no reference images available. For each digit and for a given $\mu > 0$, we define $r(\mu)$ as the relative percentage value between:

- the dissimilarity $M^g(\hat{\boldsymbol{v}}_\mu, \boldsymbol{y})$ of the images after registration, with regularization parameter μ, see equation (3.6),

and

- the dissimilarity $M^g(\hat{\boldsymbol{v}}_0, \boldsymbol{y})$ of the images before registration (i.e., with regularization parameter $\mu = 0$) ,

that is

$$r(\mu) = 100 * \frac{M^g(\hat{\boldsymbol{v}}_\mu, \boldsymbol{y})}{M^g(\hat{\boldsymbol{v}}_0, \boldsymbol{y})}.$$

We also define the finite difference derivative $\Delta_h r(\mu) = -(r(\mu + h) - r(\mu))/h$, for $h > 0$. For $h = 2$ and for all digits, we observed that the curves $\mu \to r(\mu)$ (with $\mu = 0, 0+h, \ldots, 30$) have an approximate convex shape and that the curves $\mu \to \Delta_h r(\mu)$, with $\mu = 0, 0 + h, \ldots, 28$, have a decreasing trend to 0. We display the curves $\mu \to r(\mu)$ in Figure 11, for digits 0 and 1. It is reasonable to say that taking μ such that $\Delta_h r(\mu)$ is large, corresponds to a situation of underfitting, whereas the cases such that $\Delta_h r(\mu - h)$ is small corresponds to a situation of overfitting.

Hence, an automatic choice for the regularization parameter is to take

$$\mu^* = \max\{0, 0 + h, \ldots, 28 : \Delta r(\mu) > t\},$$

where $0 \leq t \leq 100/h$ is a threshold value. One thus has the following interpretation: for μ larger than the selected value μ^* the rate of decrease of $r(\mu)$ is less than $t\%$. By setting $t = 2$, we have obtained $\mu^* = 12, 16, 14, 14, 18, 18, 12, 12, 10, 10$ for digit $0, 1, \ldots, 9$ respectively. We observed that $r(\mu^*)$ range from 0.1 to 0.3 among all digits, that is, the dissimilarity between the images after registration, using the regularization parameter μ^*, corresponds to $10\% - 30\%$ of the dissimilarity between the images before registration.

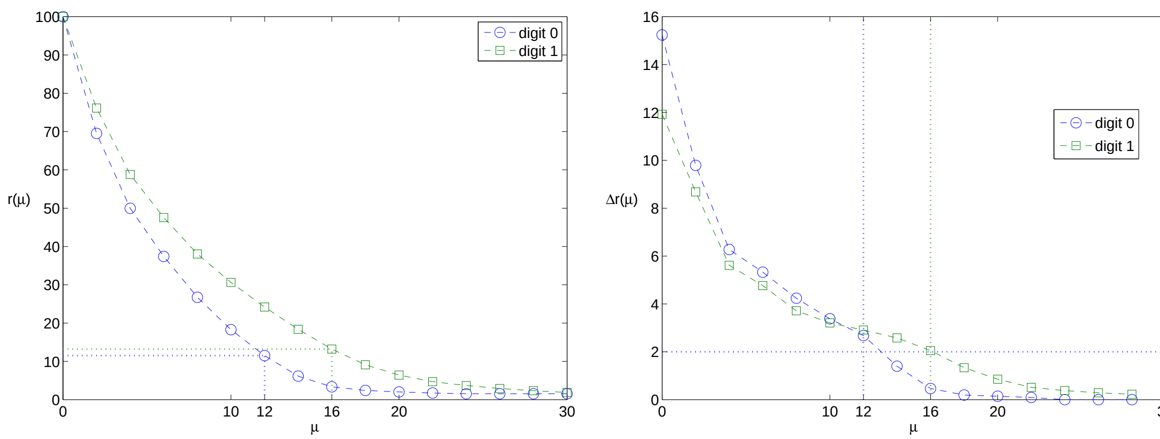


Figure 11: Choice of the regularization parameter μ^* for digit 0 and digit 1 though the analysis of the curves $\mu \to r(\mu)$ (figure on left-hand side) and $\mu \to \Delta r(\mu)$ (figure on right-hand side) with threshold $t = 2\%$.

For each digit, we carried out a geometric PCA with a preliminary registration step, as described in the previous paragraph, and with regularization parameter μ^*. To illustrate the advantages of our procedure we have also carried out a standard PCA of each digit, which amounts to analyzing the photometric variability of the data. Thus, we have computed

$$\bar{y}_n + \rho\sqrt{\hat{\gamma}_k}\hat{u}_k,$$

the k-th standard empirical mode of photometric variation of the data as described in Section 1.1. In Figure 12, we show the geometric modes of variations by displaying the images

$$\hat{f} \circ \hat{\psi}_{k,\rho},$$

where $\hat{f}(x) = \frac{1}{n}\sum_{j=1}^{n} y_j(\varphi(\hat{v}_j, x))$, $k = 1, 2$ and $\rho = 2, -2$. Results using the standard PCA are also displayed in Figure 12. We observe that geometric PCA better reflects the main modes of variability of the digits. To the contrary, standard PCA fails in several cases in representing the geometric variability of some digits and it results, sometimes, in a blurring of the images. Also, it can be seen that $\hat{f}$ is a much better mean pattern of the data than the Euclidean mean $\bar{y}_n$.

We also use the learned Fréchet mean $\hat{f}$ and the learned empirical eigenvalues $(\hat{\lambda}_k)_{k\in\mathbb{K}}$ and eigenvectors $(\hat{\phi}_k)_{k\in\mathbb{K}}$ to produce simulated images consisting of a random warp of the Fréchet mean. More precisely, we generate a new image $Y \in L^2(\Omega)$ from the model

$$Y(x) = \hat{f}(\varphi^{-1}(V, x)), \; x \in \Omega, \tag{3.12}$$

where $V = \bar{v}_n + \sum_{k=1}^{q} \rho_k \sqrt{\hat{\lambda}_k}\hat{\phi}_k$, $q \leq 2p$ and $\rho_k, k = 1, \ldots, q$, are independent standard normal random variables. Observe that $\mathbb{E}(V) = \bar{v}_n$ and that the covariance operator $\tilde{K}_n$ of V is the projection of $\hat{K}_n$ onto the space generated by $\hat{\phi}_1, \ldots, \hat{\phi}_q$. In Figure 13, we display for each digit, five independent random images obtained from (3.12), with $q = 8$. For such choice of q we have that the ratio between the trace of $\tilde{K}_n$ and the trace of $\hat{K}_n$, that is $\sum_{k=1}^{2p} \hat{\lambda}_k / \sum_{k=1}^{q} \hat{\lambda}_k$, ranges from 0.75 to 0.9, among digits $0, \ldots, 9$.

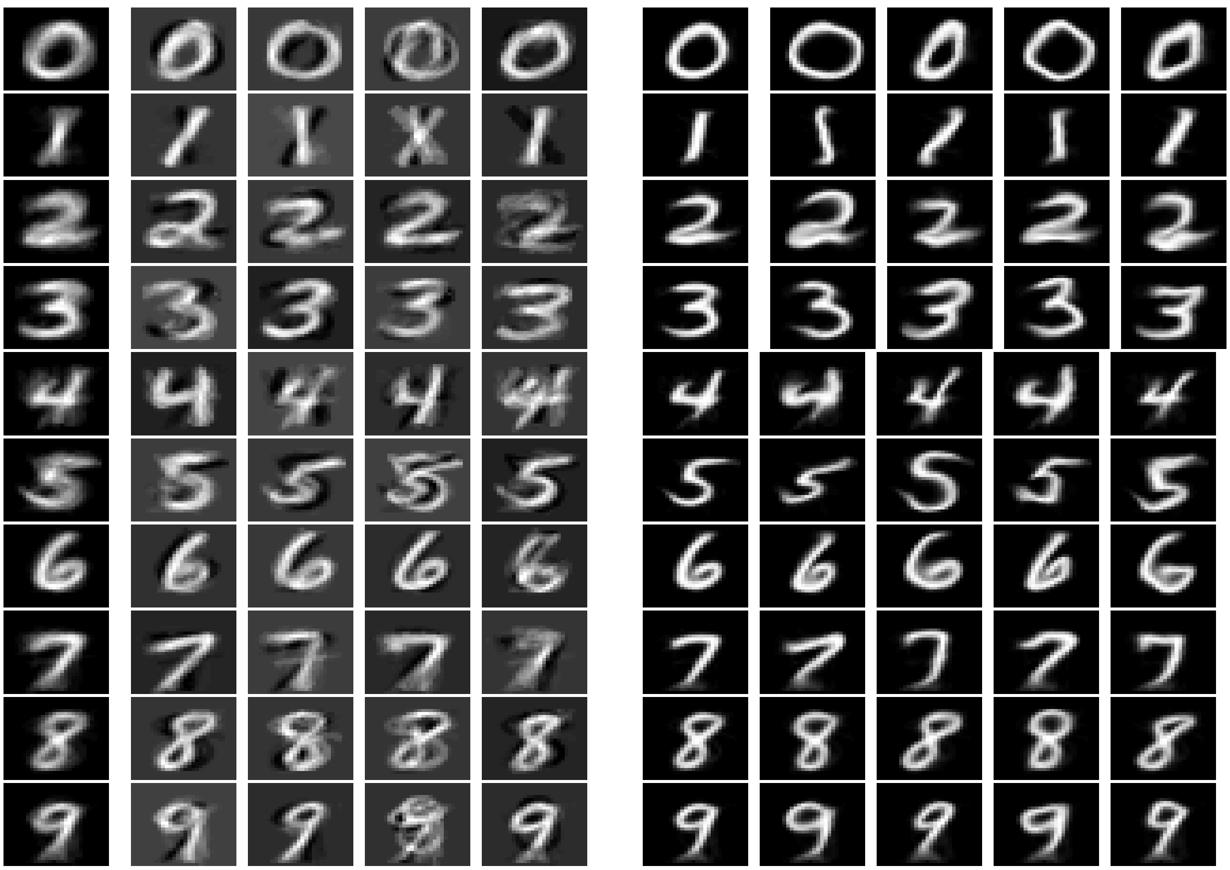


Figure 12: Visualization of the standard PCA (block of five images in the left-hand side) and geometric PCA (block of five images in the right-hand side) for digits 0 to 9. For each digit, from left to right, the images correspond to $\bar{y}$, $\bar{y}_n - 2\sqrt{\hat{\gamma}_1}\hat{u}_1$, $\bar{y}_n + 2\sqrt{\hat{\gamma}_1}\hat{u}_1$, $\bar{y}_n - 2\sqrt{\hat{\gamma}_2}\hat{u}_2$ and $\bar{y}_n + 2\sqrt{\hat{\gamma}_2}\hat{u}_2$, and then to $\hat{f}$, $\hat{f} \circ \hat{\psi}_{1,-2}$, $\hat{f} \circ \hat{\psi}_{1,2}$, $\hat{f} \circ \hat{\psi}_{2,-2}$ and $\hat{f} \circ \hat{\psi}_{2,2}$. We observe, in several cases, that standard PCA results do not recover well the shape of the digits and that they produce a blurring of the images. In contrast, geometric PCA results recover the geometric features of the digits.

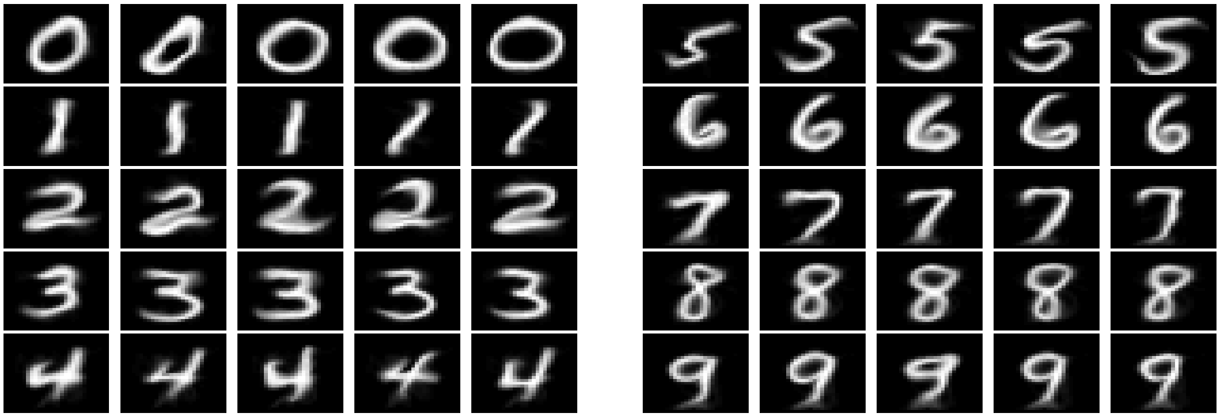


Figure 13: Simulated images for each digit $0, \dots, 9$ from model (3.12) based on the learning of the Fréchet mean $\hat{f}$ and the eigenvalues/eigenvectors of the empirical covariance operator $\hat{K}_n$.

3.3 Consistency of geometric PCA in statistical deformable models

In the last decades, in the framework of Grenander's pattern theory, there has been a growing interest in the use of first order statistics for the computation of mean pattern from a set of images [Allassonière et al., 2007, 2010; Bigot and Charlier, 2011; Bigot et al., 2009b,a] and in the construction of consistent procedures. However, there is not so much work in the statistical literature on the consistency of second order statistics for the analysis of the geometric variability of images. In particular, the convergence of such procedures in simple statistical models has generally not been established.

We study the consistency of geometric PCA, in the context of the following statistical deformable model

$$Y_i(x) = f^*(\varphi^{-1}(V_i, x)) + \epsilon W_i(x),\ x \in \Omega,\ i = 1, \dots, n, \tag{3.13}$$

where

- f^* is an unknown mean pattern belonging to $L^2(\Omega)$,
- φ is a deformation operator associated with a Hilbert space $\mathcal{V}$, (in the sense of Definition 3.1), equipped with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \|$,
- $V_1, \dots, V_n$ are independents copies of V, a zero-mean, square-integrable $\mathcal{V}$-valued random variable (i.e. $\mathbb{E}V = 0$ and $\mathbb{E}\|V\|^2 < \infty$),
- There exists $\mu > 0$ (regularization parameter) such that $\mathbb{P}(V \in \mathcal{V}_\mu) = 1$,
- $\epsilon > 0$ is a noise level parameter,
- $W_1, \dots, W_n$ are independents copies of a zero mean Gaussian process $W \in L^2(\Omega)$, such that $\mathbb{E}\|W\|_2^2 = 1$,
- $(V_1, \dots, V_n)$ and $(W_1, \dots, W_n)$ are mutually independent.

Additionally, we assume that the eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ of the population covariance operator K, defined as

$$Kv = \mathbb{E}\langle V, v\rangle V \ , \ v \in \mathcal{V}, \tag{3.14}$$

have algebraic multiplicity 1, i.e. $\lambda_1 > \lambda_2 > \ldots \geq 0$. This implies that the k-th eigen-gap, defined as $\delta_k := \min_{k' \in \mathbb{K}\setminus\{k\}} |\lambda_k - \lambda_{k'}|$, is strictly positive, for any $k \in \mathbb{K}$.

Observe that the function f^* in (3.13) models the common shape of the Y_i's. The W_i's represent the individual variations in intensity of the data around the mean pattern f^* and thus correspond to a classical source of variability, that could be analyzed by standard PCA. On the contrary, the random elements $\varphi^{-1}(V_i, \cdot)$ model deformations of the domain Ω, and thus correspond to a source of geometric variability in the data.

Model (3.13) is somewhat ideal since images are never observed in a continuous domain but rather on a discrete set of pixels. A detailed discussion on this issue can be found in [Allassonière et al., 2007, 2010] where it is proposed to deform a template model and not the observed discrete images themselves for the purpose of template estimation. However, to study the asymptotic properties of a statistical procedure, it is simpler to assume that the data are random elements of $L^2(\Omega)$ thus avoiding the treatment of the bias introduced by any discretization scheme.

Definition 3.4 (Population geometric modes of variations)**.** *Let K be the population covariance operator, defined in* (3.14)*, with (population) eigenvalues $\lambda_1 > \lambda_2 > \ldots \geq 0$ and (population) orthonormal eigenvectors $\phi_1, \phi_2, \ldots$. For $k \in \mathbb{K}$, the k-th population mode of geometric variation of the random variable V is the homeomorphism $\psi_k : \Omega \to \Omega$ defined by*

$$\psi_k(x) = \varphi^{-1}(\sqrt{\lambda_k}\phi_k, x), \ x \in \Omega.$$

In this chapter we say that geometric PCA is a consistent procedure if, for data $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ following model (3.13), and for all $k \in \mathbb{K}$, the k-th empirical mode of geometric variation $\hat{\psi}_k$ (see equation (3.9)) tends to the k-th population mode of geometric variation ψ_k, as $n \to +\infty$ and $\epsilon \to 0$, in a sense to be made precise later on. In this context, the empirical modes of geometric variation are obtained from the eigenvalues $\hat{\lambda}_k$ and the eigenvectors $\hat{\phi}_k$ of the empirical covariance operator

$$\hat{K}_n v = \frac{1}{n}\sum_{i=1}^{n} \langle \hat{V}_i - \overline{V}_n, v\rangle (\hat{V}_i - \overline{V}_n), \ \text{for } v \in \mathcal{V}, \tag{3.15}$$

where $(\hat{V}_1, \ldots, \hat{V}_n)$ belongs to $\arg\min_{\boldsymbol{v} \in \boldsymbol{\mathcal{V}}_\mu} M^t(\boldsymbol{v}, \boldsymbol{Y}, f^*)$ or $\arg\min_{\boldsymbol{v} \in \boldsymbol{\mathcal{U}}} M^g(\boldsymbol{v}, \boldsymbol{Y})$. Consequently, from now on, empirical eigenvalues, eigenvectors and modes of geometric variations will be considered as random elements.

Remark 3.1. *Observe that, for template registration, where $\hat{\boldsymbol{v}} \in \arg min_{\boldsymbol{v} \in \boldsymbol{\mathcal{V}}_\mu} M^t(\boldsymbol{v}, \boldsymbol{y}, f^*)$, each coordinate $\hat{v}_i$ depends only on v_i. This fact implies that $\hat{V}_1, \ldots, \hat{V}_n$ are independent and identically distributed (i.i.d.). However, this is not the case for groupwise registration, where $\hat{v}_i$ may depend on all the v_i's.*

The asymptotic $n \to +\infty$ is rather natural and it corresponds to the setting of a growing number of images. On the other hand, the setting $\epsilon \to 0$ corresponds to the analysis of the influence of the additive term ϵW_i in model (3.13). In the statistical literature, see e.g.

[Brown and Low, 1996], it has been shown that $\epsilon \to 0$, in a white noise model such as (3.13), corresponds to the setting where the number $N \sim \epsilon^{-2}$ of pixels would tend to infinity in a related model of images sampled on a discrete grid of size N. Therefore, one may interpret $\epsilon \to 0$ as the asymptotic setting where one observes images with a growing number $N \sim \epsilon^{-2}$ of pixels.

The main result of this section is that geometric PCA is consistent only in the double asymptotic setting $n \to +\infty$ and $\epsilon \to 0$. This result illustrates the fact that the photometric perturbations ϵW_i in model (3.13) have to be sufficiently small in order to recover the geometric modes of variation. One may argue that the main interest for practical purposes is the asymptotic setting where $n \to +\infty$ and ϵ is fixed. However, recent results show that, in the setting where ϵ is fixed, it is not possible to recover the random variables V_i encoding the deformations in model (3.13), by any statistical procedure, see e.g. [Bigot and Charlier, 2011; Bigot and Gadat, 2010]. In the double asymptotic setting $n \to +\infty$ and $\epsilon \to 0$, a detailed analysis of the problem of recovering the template f^* in model (3.13) has been carried out in [Bigot and Gendre, 2013]. In particular, some answers are given in [Bigot and Gendre, 2013] on the relative rate between n and ϵ that is needed to guarantee a consistent estimation of f^* via the use of the Fréchet mean. However, it is out of the scope of this thesis to discuss such issues for the problem of consistent estimation of the main modes of geometric variations.

Definition 3.5. *A deformation operator φ (see Definition 3.1) is said to be μ-regular if there exists $\mu > 0$ such that*

(i)

$$\int_\Omega f^2(\varphi^{-1}(v,x))dx \leq A_\mu \int_\Omega f^2(x)dx, \tag{3.16}$$

for all $f \in L^2(\Omega)$, $v \in \mathcal{V}_\mu$ and some constant $A_\mu > 0$; and

(ii) the mapping $v \to \varphi(v,\cdot)$ from $\mathcal{V}_\mu$ to $C(\Omega,\Omega)$ is continuous,

where $C(\Omega,\Omega)$ is the space of continuous functions from Ω to Ω, endowed with the metric $d_C(\psi,\phi) := \sup_{x\in\Omega} d_\Omega(\psi(x),\phi(x))$.

Note that if $\varphi(v,\cdot)$ is sufficiently smooth, such that the determinant of its Jacobian matrix is bounded, that is $|\det(J(\varphi(v,x))| \leq A_\mu$, for all $v \in \mathcal{V}_\mu$ and $x \in \Omega$, then (3.16) follows from a change of variable.

Finally, before stating our consistency results, we define convergence in probability in the double asymptotic setting $n \to \infty, \epsilon \to 0$. Let $X_{n,\epsilon}, X_n, X_\epsilon, X, n = 1,2,\ldots,\epsilon > 0$ random variables with values on a metric space (S,d). The notation $\operatorname{plim}_\epsilon X_{n,\epsilon} = X_n$ stands for $d(X_{n,\epsilon}, X_n) \to 0$ in probability as $\epsilon \to 0$; $\operatorname{plim}_n X_{n,\epsilon} = X_\epsilon$ denotes $d(X_{n,\epsilon}, X_\epsilon) \to 0$ in probability as $n \to \infty$. Finally, $\operatorname{plim}_{n,\epsilon} X_{n,\epsilon} = X$ means that $\operatorname{plim}_n \operatorname{plim}_\epsilon X_{n,\epsilon} = \operatorname{plim}_\epsilon \operatorname{plim}_n X_{n,\epsilon} = X$. In this chapter, all equalities and inequalities involving random variables are understood in the almost sure sense. We require the following definition: for $u, v \in \mathcal{V}$, $\sin(u,v) := \sqrt{1 - \langle u/\|u\|, v/\|v\| \rangle^2}$.

3.3.1 Case of template registration

Theorem 3.1. *Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ be i.i.d. observations of model* (3.13)*, with deformation operator φ and regularization parameter μ. Let $\hat{\lambda}_k$ and $\hat{\phi}_k, k \in \mathbb{K}$, be the empirical eigenvalues and eigenvectors corresponding to template registration of $\boldsymbol{Y}$, with $f = f^*$. Suppose that φ is μ-regular and that $\varphi^* : \mathcal{V}_\mu \to L^2(\Omega)$, defined by $\varphi^*(v) := f^* \circ \varphi^{-1}(v, \cdot)$, for $v \in \mathcal{V}_\mu$, is one-to-one and its inverse $\varphi^{*-1} : \varphi^*(\mathcal{V}_\mu) \to \mathcal{V}_\mu$ is continuous. Then* $\operatorname{plim}_{n,\epsilon} \hat{\lambda}_k = \lambda_k$ *and* $\operatorname{plim}_{n,\epsilon} \sin^2(\hat{\phi}_k, \phi_k) = 0$*, for all $k \in \mathbb{K}$.*

The injectivity condition of φ^* in the previous theorem, implies that, if there were no additive noise in model (3.13), then the registration of the observations $\boldsymbol{Y}$ onto the template f^* would lead exactly to the non-observed deformation parameters $V_1, \ldots, V_n$. In other words, if $\epsilon = 0$ in (3.13), then the template dissimilarity functional $M^t(\boldsymbol{v}, \boldsymbol{Y}, f^*)$ (see (3.3)) has a unique minimizer over $\boldsymbol{\mathcal{V}}_\mu$ given by $\boldsymbol{v} = (V_1, \ldots, V_n)$. The condition of continuity of φ^{*-1} ensures that the registration problem with noise level ϵ will converge to the registration problem with no noise as $\epsilon \to 0$. In the following proposition, we provide sufficient conditions to ensure that if φ^* is one-to-one, then its inverse is continuous.

Proposition 3.2. *Let φ be a μ-regular deformation operator, with Ω compact and $\mathcal{V}$ finite dimensional. If f^* is continuous and φ^*, defined in Theorem 3.1, is one-to-one, then its inverse is continuous.*

Proof. Let us show first that φ^* is continuous, so take a sequence (v_n) in $\mathcal{V}_\mu$ converging to $v \in \mathcal{V}_\mu$. By the continuity of f^* and condition (ii) of Definition 3.5, we have $f^*(\varphi^{-1}(v_n, x)) \to f^*(\varphi^{-1}(v, x))$, for all $x \in \Omega$. On the other hand, $f^*(\varphi^{-1}(v_n^*, x)) \leq \sup_{x' \in \Omega} |f^*(x')| < \infty$, for all $x \in \Omega$. Then, by the dominated convergence theorem, we conclude that φ^* is continuous. Finally, recall that the inverse of a one-to-one and continuous function from a compact space onto a topological space, is also continuous. Therefore, as $\mathcal{V}_\mu$ is compact, we obtain the result. □

In Section 3.3.3, we analyze the case where φ is the translation operator defined in Section 3.2.1 and we provide some conditions on f^* and φ ensuring that the hypotheses of Theorem 3.1 are satisfied. In the case where φ is the diffeomorphic deformation operator, defined in Section 3.2.1, it is necessary to impose much stronger assumptions on the template f^* and the space of vector fields $\mathcal{V}$ to ensure that $\varphi^* : \mathcal{V}_\mu \to \varphi^*(\mathcal{V}_\mu)$ has a continuous inverse; see Section 3.3.4.

Remark 3.2. *Observe that, under the hypotheses of Theorem 3.1, the k-th empirical mode of geometric variation $\hat{\psi}_k$ converges in probability the the k-th population mode of geometric variation ψ_k, when $n \to +\infty$ and $\epsilon \to 0$, as elements of $(C(\Omega, \Omega), d_C)$. Indeed, this result follows from the continuity of the mapping $v \to \varphi(v, \cdot)$, which is guaranteed by the μ-regularity of φ.*

We show below how a stronger regularity assumption on φ allows one to obtain rates of convergence for $\hat{\lambda}_k$ and $\hat{\phi}_k$, via a concentration inequality that depends explicitly on n and ϵ.

Theorem 3.2. *Under the hypotheses of Theorem 3.1 and if φ^{*-1} is uniformly Lipschitz (in the sense that $\|u - v\|^2 \leq L(f^*, \mu)\|\varphi^*(u) - \varphi^*(v)\|_2^2$, for every $u, v \in \mathcal{V}_\mu$ and some constant*

$L(f^*, \mu) > 0$ depending only on f^* and μ), then

$$\mathbb{P}\left(|\hat{\lambda}_k - \lambda_k|^2 > C(f^*, \mu)\max(h(u,n,\epsilon) + \sqrt{h(u,n,\epsilon)}; g(u,n))\right) \le \exp(-u),$$

for any $u > 0$, where $C(f^*, \mu) > 0$ is a constant depending only on f^* and μ; $h(u,n,\epsilon) = \epsilon^2\left(1 + 2\frac{u}{n} + 2\sqrt{\frac{u}{n}}\right)$ and $g(u,n) = \left(\frac{u}{n} + \sqrt{\frac{u^2}{n^2} + \frac{u}{n}}\right)^2$.

Take now $u^* > 0$ such that

$$C(f^*, \mu)\max(h(u^*,n,\epsilon) + \sqrt{h(u^*,n,\epsilon)}; g(u^*,n)) < (\delta_k/2)^2,$$

then, for any $0 < u \le u^*$,

$$\mathbb{P}\left(\sin^2(\hat{\phi}_k, \phi_k) > (2/\delta_k)^2 C(f^*, \mu)\max(h(u,n,\epsilon) + \sqrt{h(u,n,\epsilon)}; g(u,n))\right) \le 2\exp(-u).$$

3.3.2 Case of groupwise registration

In order to prove consistency for groupwise registration, we require model (3.13) to satisfy the set of identifiability assumptions, shown below. For $\boldsymbol{u}, \boldsymbol{v} \in \boldsymbol{\mathcal{V}} = \mathcal{V}^n$, let

$$D^g(\boldsymbol{u}, \boldsymbol{v}) := M^g(\boldsymbol{u}, (f_1^*, \dots, f_n^*)) = \frac{1}{n}\sum_{i=1}^n \int_\Omega \left(f_i^*(\varphi(u_i, x)) - \frac{1}{n}\sum_{j=1}^n f_j^*(\varphi(u_j, x))\right)^2 dx, \tag{3.17}$$

where $f_i^*(x) := f^*(\varphi^{-1}(v_i, x)), x \in \Omega,\ i = 1, \dots, n$. Observe that, $D^g(\boldsymbol{u}, \boldsymbol{V}) = M^g(\boldsymbol{u}, \boldsymbol{Y})$ when $\boldsymbol{Y} = (Y_1, \dots, Y_n)$ follows model (3.13) with $\epsilon = 0$.

Definition 3.6 (g-identifiability)**.** *Model* (3.13) *is said to be g-identifiable if*

(i) there exists a measurable function $\boldsymbol{u}^* : \boldsymbol{\mathcal{V}}_\mu \to \boldsymbol{\mathcal{U}}$ such that for every $\eta > 0$ there exists a constant $C > 0$, not depending on n, with $D^g(\boldsymbol{u}, \boldsymbol{v}) - D^g(\boldsymbol{u}^*, \boldsymbol{v}) > C$, for every $\boldsymbol{u} \in \boldsymbol{\mathcal{U}}$ satisfying $\bar{d}^2(\boldsymbol{u}^*, \boldsymbol{u}) > \eta$, and

(ii) $\operatorname{plim}_n \bar{d}^2(\boldsymbol{u}^*(\boldsymbol{V}), \boldsymbol{V}) = 0$,

where $\bar{d}^2(\boldsymbol{u}, \boldsymbol{v}) := \frac{1}{n}\sum_{i=1}^n \|u_i - v_i\|^2$, for $\boldsymbol{u}, \boldsymbol{v} \in \boldsymbol{\mathcal{V}}$.

Observe that condition (i) above implies that, for every $\boldsymbol{v} \in \boldsymbol{\mathcal{V}}_\mu$, $D^g(\boldsymbol{u}, \boldsymbol{v})$ has a unique measurable minimizer $\boldsymbol{u}^*(\boldsymbol{v})$ on $\boldsymbol{\mathcal{U}}$.

Theorem 3.3. *Let $\boldsymbol{Y} = (Y_1, \dots, Y_n)$ be i.i.d. observations of model* (3.13)*, with deformation operator φ and regularization parameter μ. Let $\hat{\lambda}_k$ and $\hat{\phi}_k, k \in \mathbb{K}$, be the empirical eigenvalues and eigenvectors corresponding to the groupwise registration of $\boldsymbol{Y}$. Suppose that φ is μ-regular and that* (3.13) *is g-identifiable. Then $\operatorname{plim}_{n,\epsilon} \hat{\lambda}_k = \lambda_k$ and $\operatorname{plim}_{n,\epsilon} \sin^2(\hat{\phi}_k, \phi_k) = 0$, for all $k \in \mathbb{K}$.*

Remark 3.3. *Observe that, as in the case of template registration, it can be shown that under the hypotheses of Theorem 3.3, $\hat{\psi}_k$ converges to ψ_k in probability, as $n \to +\infty$ and $\epsilon \to 0$.*

3.3.3 Translation operator

We study the applicability of Theorems 3.1, 3.2 and 3.3 to the translation operator φ given by (3.1). In this case, $\Omega = [0,1)^d$, for some integer $d \geq 1$, is equipped with the distance $d_\Omega(x,y) := \sum_{k=1}^{d} \min\{|x_k - y_k|, 1 - |x_k - y_k|\}$, for $x = (x_1, \dots, x_d), y = (y_1, \dots, y_d) \in \Omega$. Let also $\mathcal{V} = \mathbb{R}^d$ be equipped with the usual Euclidean inner product.

Now, let us show that φ is a deformation operator in the sense of Definition 3.1. It holds that $\varphi(0,\cdot)$ is the identity in Ω and $\varphi^{-1}(v,\cdot) = \varphi(-v,\cdot)$. Last, from (i) in Lemma B.5, it follows that the mapping $\varphi(v,\cdot)$, is continuous for all $v \in \mathcal{V}$. Moreover, we prove that φ is μ-regular, for all $\mu > 0$: for (i) in Definition 3.5, take $f \in L^2(\Omega)$ and consider its periodic extension f_{per} to $\mathbb{R}^d$. Then (3.16) is a consequence of $f(\varphi(v,x)) = f_{per}(x+v)$ which holds for all $v \in \mathbb{R}^d, x \in \Omega$. Finally, condition (ii) in Definition 3.5, follows from (ii) in Lemma B.5.

We impose further conditions on model (3.13) implying that φ^{*-1}, defined in Theorem 3.1, is Lipschitz. Let θ_k, $k = 1, \dots, d$, be the low frequency Fourier coefficients of the template f^*, that is

$$\theta_k := \int_\Omega f^*(x) e^{-i2\pi x_k} dx \neq 0 \text{ for all } 1 \leq k \leq d. \tag{3.18}$$

Lemma 3.1. *Suppose that f^* is such that $\theta_k \neq 0$, for all $1 \leq k \leq d$, and let $\mu < 1/2$. Then φ^* is one-to-one and φ^{*-1} is uniformly Lipschitz.*

Hence, if $\theta_k \neq 0$, for all $1 \leq k \leq d$ and $\mu < 1/2$, the hypotheses of Theorems 3.1 and 3.2 are verified. Thus, the geometric PCA is consistent in the case of template registration with translation operator. Observe that the hypotheses of Lemma 3.1 imply that translation invariant templates f^* are excluded.

We now turn our attention to groupwise registration. We have to impose further conditions on model (3.13) ensuring g-identifiability, so that Theorem 3.3 applies. The set of deformation parameters $\boldsymbol{\mathcal{U}} \subset \boldsymbol{\mathcal{V}}_\mu$ over which $M^g(\boldsymbol{v}, \boldsymbol{y})$ will be minimized, is $\boldsymbol{\mathcal{U}} = \boldsymbol{\mathcal{U}}_0$, given in (3.2.2). We have the following.

Proposition 3.3. *Suppose $\theta_k \neq 0$, for all $1 \leq k \leq d$, and that $\mathbb{P}(V \in [-\rho,\rho]^d) = 1$, with $\rho = \min(\frac{\mu}{2}, \frac{\mu}{\sqrt{d}})$ and $0 < \mu < \frac{1}{12}$. Then*

$$D^g(\boldsymbol{u}, \boldsymbol{v}) - D^g(\boldsymbol{u}^*(\boldsymbol{v}), \boldsymbol{v}) \geq C(f^*, \mu) \bar{d}^2(\boldsymbol{u}, \boldsymbol{u}^*(\boldsymbol{v})), \ \textit{for all } \boldsymbol{u} \in \boldsymbol{\mathcal{U}}, \tag{3.19}$$

where $\boldsymbol{u}^*(\boldsymbol{v}) := \left(v_1 - \frac{1}{n}\sum_{i=1}^n v_i, \dots, v_n - \frac{1}{n}\sum_{i=1}^n v_n\right)$ and $C(f^*, \mu) > 0$ is a constant depending only on f^* and μ.

Remark that, in Proposition 3.3, $D^g(\boldsymbol{u}^*(\boldsymbol{v}), \boldsymbol{v}) = 0$. This shows that $D^g(\boldsymbol{u}, \boldsymbol{v})$ is bounded below by a quadratic functional.

We are now ready to prove g-identifiability under the hypotheses of the previous proposition. Observe that (i) in Definition 3.6 follows at once from (3.19). For (ii) note that $\bar{d}^2(\boldsymbol{u}^*(\boldsymbol{V}), \boldsymbol{V}) = \|\frac{1}{n}\sum_{i=1}^n V_i\|$. Hence, given that $\mathbb{E}V = 0$, from Bernstein's inequality for bounded random variables in a Hilbert space (see e.g. [Bosq, 2000], Theorem 2.6) we conclude that, for any $\eta > 0$,

$$\mathbb{P}\left(d(\boldsymbol{u}^*(\boldsymbol{V}), \boldsymbol{V}) > \eta\right) \leq 2\exp\left(-\frac{n\eta^2}{2\mathbb{E}\|V\|^2 + \frac{\mu}{3}\eta}\right).$$

Therefore $\bar{d}^2(\boldsymbol{u}^*(\boldsymbol{V}), \boldsymbol{V})$ converges in probability to 0 as $n \to +\infty$.

Finally, having checked the g-identifiability of the model, we conclude that the geometric PCA is consistent, in the case of groupwise registration with translation operator.

3.3.4 Non-rigid diffeomorphic operator

We study the applicability of Theorem 3.1 to the non-rigid diffeomorphic operator φ, defined in Section 3.2.1, in the case of dimension $d = 1$. In this case, $\Omega = [0, 1]$ is equipped with the distance $d_\Omega(x, y) := |x - y|$, for $x, y \in \Omega$. Let $\mathcal{V}$ a separable Hilbert space, continuously embedded in the set $C_0^1(\Omega)$ of functions $v : \Omega \to \mathbb{R}$, which are class C^1 on Ω and such v and its derivative vanish at the boundary of Ω. That is, there exists a positive constant c_0 such that

$$\|v\|_{1,\infty} \leq c_0 \|v\|, \quad \forall v \in \mathcal{V}, \tag{3.20}$$

where $\|v\|_{1,\infty} = \sup_{x \in \Omega} |v(x)| + |\frac{\partial v}{\partial x}(x)|$ is the sup-norm. For $x \in \Omega$ and $v \in \mathcal{V}$, define $\varphi(v, x)$ as the solution at time $t = 1$ of the ODE (3.2), with initial condition $\phi_0 = x \in \Omega$.

Now, we show that φ is a deformation operator in the sense of Definition 3.1. The smoothness condition (3.20) implies that, for any $v \in \mathcal{V}$, the function $x \mapsto \varphi(v, x)$ is a diffeomorphism on Ω (see e.g. [Younes, 2010], Theorem 8.7 or [Grenander and Miller, 2007], Theorem 11.3) with inverse given by $x \mapsto \varphi(-v, x)$ (see e.g. the proof of Theorem 8.14 in [Younes, 2010] or Theorem 11.3 in [Grenander and Miller, 2007]). Moreover, we prove that φ is μ-regular, for all $\mu > 0$. From Theorem 8.9 in [Younes, 2010], there exist constants c_1, c_2, such that

$$\|\varphi(v, \cdot) - id\|_{1,\infty} \leq c_1 e^{c_2 \|v\|_{1,\infty}}, \ v \in \mathcal{V}, \tag{3.21}$$

from where we obtain that

$$\sup_{x \in \Omega} \left| \frac{\partial \varphi(v, x)}{\partial x} \right| \leq 2 + c_1 e^{c_2 \mu}, \ v \in \mathcal{V}_\mu. \tag{3.22}$$

Then (3.16) follows from (3.22) and a change of variable. Finally, condition (ii) in Definition 3.5 follows from [Younes, 2010], Theorem 8.11.

We impose further conditions on model (3.13) implying that φ^*, defined in Theorem 3.1, is one-to-one and continuous. We show first that the map $v \mapsto \varphi(v, \cdot)$ is one-to-one, under appropriate assumptions over the space $\mathcal{V}$.

Definition 3.7. *We say that $v \in \mathcal{V}$ is non-oscillatory if there exists a positive integer p (depending on v) such that the zero-level set of v defined by $L(v) = \{x \in \Omega : v(x) = 0\}$ can be written as*

$$L(v) = \bigcup_{i=1,\ldots,p} A_i, \tag{3.23}$$

where $A_i = [a_i, b_i]$, for $i = 1, \ldots, p$ with $0 = a_1 \leq b_1 < a_2 \leq b_2 < \ldots a_p \leq b_p = 1$. In words, $L(v)$ is the union of disjoint closed subintervals of Ω.

Remark 3.4. *The previous definition is satisfied for the following two spaces, which are a common election for numerical implementation. First, the space generated by a finite basis of Fourier functions. In this case it can be shown that any function in $\mathcal{V}$ has a finite number* *of roots [Boyd, 2006]. Second, the space generated by a finite number of B-Splines, which are piecewise polinomious with compact support. In this case, the previous assumption can be easily checked.*

Proposition 3.4. *If every $v \in \mathcal{V}$ is non-oscillatory, according to Definition 3.7, then the mapping $v \in \mathcal{V} \mapsto \varphi(v, \cdot)$ is one-to-one.*

We impose conditions over the template f^*, to discard invariance under the action of diffeomorphisms.

Definition 3.8. *We say that a function $f : \Omega \to \mathbb{R}$ has no flat regions if*

$$\forall x \in \Omega,\ \forall N_x \text{ neighborhood of } x \text{ in } \Omega,\ \exists x' \in N_x \text{ such that } f(x) \neq f(x'). \tag{3.24}$$

Recall that the mapping $\varphi^* : \mathcal{V}_\mu \to L^2(\Omega)$ is defined by $\varphi^*(v) := f^* \circ \varphi^{-1}(v, \cdot)$, for $v \in \mathcal{V}_\mu$.

Proposition 3.5. *Assume that f^* is continuous and has no flat regions and let $u, v \in \mathcal{V}$ such that $\varphi^*(u) = \varphi^*(v)$. Then $\varphi(u, x) = \varphi(v, x)$, for all $x \in \Omega$.*

So, if f^* is continuous with no flat regions and $\mathcal{V}$ is a finite dimensional space of non-oscillatory functions, then, by Propositions 3.2, 3.4 and 3.5, φ^* is one-to-one, with continuous inverse and so, the hypotheses of Theorem 3.1 are verified. Thus, the geometric PCA is consistent in the case of template registration, with the non-rigid diffeomorphic operator φ, previously defined.

3.4 Conclusions and discussion of this chapter

The contribution of this thesis, related to the geometric PCA, is twofold. First, the use of deformation operators (as introduced in this chapter) provides a general framework for modeling and analyzing the geometric variability of images. As a particular case, it allows the use of diffeomorphic deformations parametrized by stationary vector fields. In the case of diffeomorphisms computed with nonstationary vector fields, as in [Beg et al., 2005], the link with our framework is not straightforward. Indeed, in this setting, there are two possibilities for defining the deformation operators. One can parameterize them either by the Hilbert space of time-dependent vector fields, or by the Hilbert space of initial velocities. Both cases are rather complex from the analytical and the computational points of view and treating them is beyond the scope of this thesis. In contrast, due to its analytical and numerical tractability, we have preferred to focus on diffeomorphic deformation operators, parametrized by stationary vector fields belonging to a finite-dimensional Hilbert space.

The second contribution in this chapter is the study of the consistency of geometric PCA methods in statistical deformable models which, to the best of our knowledge, has not been investigated so far. One can remark that our consistency results rely on strong assumptions on the template f^* and the deformation operator φ. For the case of translations, we have provided (see Section 3.3.3) verifiable conditions to satisfy such assumptions. A similar analysis, in the case of diffeomorphic deformations, is much more complex (see Section 3.3.4). For the case of template registration, one of our main assumptions is that the mapping $\varphi^* : \mathcal{V}_\mu \to L^2(\Omega)$, defined by $\varphi^*(v) := f^* \circ \varphi^{-1}(v, \cdot)$, for $v \in \mathcal{V}_\mu$, is one-to-one. Such condition together with some regularity conditions on φ ensure that, if $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ is sampled from model (3.13), with noise level $\epsilon = 0$, then the registration problem $\min_{\boldsymbol{v} \in \boldsymbol{\mathcal{V}}_\mu} M^t(\boldsymbol{v}, \boldsymbol{Y}, f^*)$ has a unique solution, given by the non-observed deformation parameters $\boldsymbol{V} = (V_1, \ldots, V_n)$, where M^t is the template dissimilarity functional defined in (3.3). In particular, in the case where φ is a diffeomorphic deformations operator, the injectivity condition requires necessarily the template f^* not to be constant in any open region of Ω, which is a quite restrictive assumption that does not hold in applications. One possibility to remove this injectivity condition could be, for instance, estimating deformation parameters in the set

$$\arg\min\{\|v\| : v \in \mathcal{V}, \varphi^*(v) = \varphi^*(V_i)\}, \tag{3.25}$$

instead of estimating V_i, for all $i = 1, \ldots, n$. In Section 3.5.1 we discuss an approach based on Tikhonov regularization to address this issue. However, such an approach makes much more difficult the analysis on the consistency of our procedure.
We hope that the methods presented here will stimulate further investigation into the development of consistent statistical procedures for the analysis of geometric variability.

3.5 Extensions and related problems

In this section we discuss possible extensions and related problems of the geometric PCA. In this thesis we have not investigated such issues in depth, as they can be the object of future investigations.

3.5.1 Tikhonov regularization

In this section we discuss the use of Tikhonov regularization in the registration step of the geometric PCA. For simplicity we discuss only the case of template registration. Instead of the constrained regularization we have proposed (see (3.6)), it is possible to use Tikhonov regularization with parameter $\tau > 0$. That is, given $f \in L^2(\Omega)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$, with $y_i \in L^2(\Omega)$, $i = 1, \ldots, n$, define $\hat{\boldsymbol{v}}_\tau$ as a solution of

$$\min_{\boldsymbol{v} \in \mathcal{V}^n} M^t(\boldsymbol{v}, \boldsymbol{y}, f) + \tau R(\boldsymbol{v}), \tag{3.26}$$

where M^t is the template dissimilarity functional defined in (3.3) and $R(\boldsymbol{v}) := \frac{1}{n}\sum_{i=1}^n \|v_i\|^2$ is a regularization term. Minimizing the cost functional (3.26) is a widely used approach in many image registration problems. In order for analyzing the consistency of the geometric PCA, using the previous registration method, consider $Y_1, \ldots, Y_n$ from model (3.13), with associated random parameters $V_1, \ldots, V_n \in \mathcal{V}$. Based on the guidelines in [Droske, 2005], we believe that if $\hat{\boldsymbol{V}}_\tau$ is a solution of (3.26), with $\boldsymbol{y} := (Y_1, \ldots, Y_n)$, then the distance from $\hat{\boldsymbol{V}}_\tau$ to the set

$$\arg\min\{\|v\| \ : \ v \in \mathcal{V}, \varphi^*(v) = \varphi^*(V_i)\}, \tag{3.27}$$

converges to zero, when $n \to \infty$, $\epsilon \to 0$ and $\tau \to 0$.

3.5.2 Discrete model

For convenience, we have presented the ideas of geometric PCA under the assumption that the images $y_1, \dots, y_n \in L^2(\Omega)$ are observed on the continuous domain Ω. In this section, we assume that we observe $y_{i,\ell} = y_i(x_\ell), \ell = 1, \dots, p, i = 1, \dots, n$, where $x_1, \dots, x_p \in \Omega$ is a discrete set of pixels. We propose to interpolate/smooth the data before calculating deformation parameters. Denote by $(e_k)_{k=1}^{\infty}$ an orthonormal basis of $L^2(\Omega)$ (e.g. a Fourier or a wavelet basis). Then define, for $1 \leq i \leq n$, the following estimators obtained by smoothing the data

$$\hat{y}_i = \sum_{k=1}^{k_0} \hat{\beta}_{i,k} e_k, \quad \text{with} \quad \hat{\beta}_{i,k} = \frac{1}{p} \sum_{\ell=1}^{p} y_{i,\ell} e_k(x_\ell),$$

where $k_0 \in \mathbb{N}$ is a regularization parameter whose choice has to be discussed. We define the method of geometric PCA on the data $y_{i,\ell} = y_i(x_\ell), \ell = 1, \dots, p, i = 1, \dots, n$ as the geometric PCA (see Definition 3.3) applied to $\hat{y}_1, \dots, \hat{y}_n$.

Analogously to the analysis made in Section 3.3, we are interested in studying the consistency of discrete geometric PCA, when data comes from the following statistical deformable model

$$Y_{i,\ell} = f^*(\varphi^{-1}(V_i, x_\ell)) + \epsilon W_i(x_\ell),\ x_\ell \in \Omega,\ i = 1, \dots, n,\ \ell = 1, \dots, p, \tag{3.28}$$

which is the discrete analog of model 3.13.

Conclusions

In this thesis we define and analyze two different adaptations of the Principal Components Analysis (PCA) for analyzing the variability of two instances of infinite dimensional data: measures on the line and images. Each of the proposed methods is analyzed independently, however, there exists a common basis and several connections between them, as we point out in the Introduction. In Chapter 2 we define the Geodesic PCA of measures, a method that has a clear parallel with respect to recently proposed PCA adaptations to Riemaniann manifolds, (see discussion in Section 2.6). In Chapter 3 we define the geometric PCA of images, a method that relies on well known procedures of image analysis, but it also can be interpreted as as PCA on the manifold of deformations (see discussion in Section 3.1.1). Our main mathematical result is the statistical consistency of our procedures, that is, the convergence of the empirical characteristics to their population counterparts. We remark that these results are original contributions that make use, from our point of view, of advanced mathematical tools and techniques.

There exists another interesting link between the two approaches we propose, namely a possible application of Geodesic PCA to analyze the variability of an image set. Modeling image variability using Wasserstein distance is based on the idea that images represent a material density over the domain Ω and that the source of image variability is due to a transportation of material. Therefore, the Wasserstein distance is well suited, for instance, when the gray level values represent the density of a tissue, as in the case of MRI cardiac applications [Zhu and Tannenbaum, 1998]. However, the extension of our methodology to dimension $d > 1$ is necessary to deal with applications in image analysis, but is clearly not straightforward. In this context we mention recent work on image registration based on Wasserstein distance: [Zhu and Tannenbaum, 1998; Haker et al., 2001; Museyko et al., 2009; Haker et al., 2004; Sulman et al., 2009].

Finally we remark that the subject of this thesis belongs primarily to the fields of probability and statistics, however it also has components of others fields: differential geometry, metric geometry, functional analysis, convex analysis, optimal transportation theory, shape optimization and image processing. Last, as closing remark, we would like to mention that our ambition in this research project was to strike a good balance between theory and numerical implementations. We hope this work partially fulfills such aspirations.

References

M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM J. Math. Anal.*, 43 (2):904–924, 2011.

S. Allassonière, Y. Amit, and A. Trouvé. Toward a coherent statistical framework for dense deformable template estimation. *Journal of the Statistical Royal Society (B)*, 69(1):3–29, 2007.

S. Allassonière, E. Kuhn, and A. Trouvé. Bayesian deformable models building via stochastic approximation algorithm: a convergence study. *Bernoulli*, 16(3):641–678, 2010.

S. Amari and M. Nagaoka. *Methods of Information Geometry.* American Mathematical Society. Oxford University Press, Oxford, 2000.

L. Ambrosio and G. Savaré. Gradient Flows of Probability Measures. In *Handbook of differential equations: evolutionary equations. Vol. III*, Handb. Differ. Equ., pages 1–136. Elsevier, North-Holland, Amsterdam, 2007.

L. Ambrosio, N. Gigli, and G. Savaré. Gradient flows with metric and differentiable structures, and applications to the Wasserstein space. *Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Mat. Appl.*, 15(3-4):327–343, 2004.

T.W. Anderson. Asymptotic theory for principal component analysis. *Ann. Math. Statist.*, 34(1):122–148, 1963.

V. Arsigny, O. Commowick, X. Pennec, and N. Ayache. A log-euclidean framework for statistics on diffeomorphisms. In *Proc. of the 9th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'06), Part I*, number 4190 in LNCS, pages 924–931, 2006.

Z. Artstein and R. J. B. Wets. Consistency of minimizers and the SLLN for stochastic programs. *J. Convex Anal.*, 2(1-2):1–17, 1995.

J. Ashburner. A fast diffeomorphic image registration algorithm. *NeuroImage*, 38(1):95–113, 2007.

H. Attouch. *Variational Convergence for Functions and Operators.* Applicable Mathematic series. Pitman, London, 1984.

G. Beer. On convergence of closed sets in a metric space and distance functions. *Bulletin of the Australian Mathematical Society*, 31(03):421–432, 1985.

M. F. Beg, M. I. Miller, A. Trouvé, and L. Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *Int. J. Comput. Vision*, 61(2):139–157, 2005.

R. Bhatia, C. Davis, and A. McIntosh. Perturbation of spectral subspaces and solution of linear operator equations. *Linear Algebra Appl.*, 52/53:45–67, 1983.

R. Bhattacharya and V. Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifolds. *Annals of statistics*, 31(1):1–29, 2003.

J. Bigot and B. Charlier. On the consistency of Fréchet means in deformable models for curve and image analysis. *Electronic Journal of Statistics*, 5:1054–1089, 2011.

J. Bigot and S. Gadat. A deconvolution approach to estimation of a common shape in a shifted curves model. *Ann. Statist.*, 38(4):2422–2464, 2010.

J. Bigot and X. Gendre. Minimax properties of Fréchet means of discretely sampled curves. *Ann. Statist.*, 41(2):923–956, 2013.

J. Bigot and T. Klein. Consistent estimation of a population barycenter in the Wasserstein space. *Preprint, ArXiv/1212.2562*, 2012.

J. Bigot, S. Gadat, and J.M. Loubes. Statistical M-estimation and consistency in large deformable models for image warping. *Journal of Mathematical Imaging and Vision*, 34 (3):270–290, 2009a.

J. Bigot, F. Gamboa, and M. Vimond. Estimation of translation, rotation and scaling between noisy images using the Fourier Mellin transform. *SIAM Journal on Imaging Sciences*, 2 (2):614–645, 2009b.

S. Bonaccorsi and E. Priola. From brownian motion to stochastic differential equations. In *10th Internet Seminar*, 2006.

D. Bosq. *Nonparametric Statistics for Stochastic Processes: Estimation and prediction*, volume 110 of *Lecture Notes in Statistics*. Springer-Verlag, New York, second edition, 1998.

D. Bosq. *Linear Processes in Function Spaces: Theory and applications*, volume 149 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 2000.

M. Bossa, M. Hernandez, and S. Olmos. Contributions to 3D diffeomorphic atlas estimation: application to brain images. In *Proceedings of the 10th international conference on Medical image computing and computer-assisted intervention*, volume 10 of *MICCAI'07*, pages 667–674. Springer-Verlag, Berlin, Heidelberg, 2007.

M. Bossa, E. Zacur, and S. Olmos. Algorithms for computing the group exponential of diffeomorphisms: Performance evaluation. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on*, pages 1–8, 2008.

J. Boyd. Computing the zeros, maxima and inflection points of Chebyshev, Legendre and Fourier series: solving transcendental equations by spectral interpolation and polynomial rootfinding. *Journal of Engineering Mathematics*, 56(3):203–219, 2006.

Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math.*, 44(4):375–417, 1991.

L. D. Brown and M. G. Low. Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.*, 24(6):2384–2398, 1996.

M. Carrasco. Probabilidades en Espacios Abstractos: una introducción. Technical report, Facultad de Ciencias, Universidad de la República, Uruguay, 2005.

O. Chodosh. Optimal transport and Ricci curvature: Wasserstein space over the interval. *Preprint, ArXiv/1105.2883*, 2011.

T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models - their training and application. *Comput. Vis. Image Underst.*, 61(1):38–59, 1995.

G. Dal Maso. *An Introduction to* Γ*-convergence*, volume 8 of *Progress in Nonlinear Differential Equations and their Applications*. Birkhäuser, Boston, MA, 1993.

J. Dauxois, A. Pousse, and Y. Romain. Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *J. Multivariate Anal.*, 12(1):136–154, 1982.

C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.*, 7(1):1–46, 1970.

M. Droske. *On Variational Problems and Gradients Flows in Image Processing*. Ph.D. Thesis, Universitat Duisburg-Essen, 2005.

P. Embrechts and M. Hofert. A note on generalized inverses. *Mathematical Methods of Operations Research*, 77(3):423–432, 2013.

M. Fabian, P. Habala, P. Hájek, V. Montesinos, and V. Zizler. Hahn–Banach and Banach Open Mapping Theorems. In *Banach Space Theory*, CMS Books in Mathematics, pages 53–81. Springer, New York, 2011.

P. T. Fletcher, C. Lu, Stephen 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.

S. Gallón, J.-M. Loubes, and E. Maza. Statistical properties of the quantile normalization method for density curve alignment. *Mathematical Biosciences*, 242(2):129–142, 2013.

U. Grenander and M. Miller. *General Pattern Theory*. Oxford Univ. Press, Oxford, 1993.

U. Grenander and M. Miller. Computational anatomy: An emerging discipline. *Quart. Appl. Math.*, LVI(4):617–694, 1998.

U. Grenander and M. Miller. *Pattern Theory: From Representation to Inference*. Oxford Univ. Press, Oxford, 2007.

E. Haber and J. Modersitzki. Image registration with guaranteed displacement regularity. *Int. J. on Comp. Vision*, 71(3):361–372, 2006.

N. N. Hai and P. T. An. A generalization of Blaschke's convergence theorem in metric spaces. *Journal of Convex Analysis*, to be published, 2013.

Steven Haker, Allen Tannenbaum, and Ron Kikinis. Mass preserving mappings and image registration. In *Proceedings of the 4th International Conference on Medical Image Com-*

puting and Computer-Assisted Intervention, MICCAI '01, pages 120–127. Springer-Verlag, London, UK, 2001.

Steven Haker, Lei Zhu, Allen Tannenbaum, and Sigurd Angenent. Optimal mass transport for registration and warping. *International Journal on Computer Vision*, 60(3):225–240, 2004.

M. Hernandez, M. Bossa, and S. Olmos. Estimation of statistical atlases using groups of diffeomorphisms. Technical report, Aragon Institute of Engineering Research, University of Zaragoza, Spain, 2007.

M. Hernandez, M. Bossa, and S. Olmos. Registration of anatomical images using paths of diffeomorphisms parameterized with stationary vector field flows. *International Journal of Computer Vision*, 85(3):291–306, 2009.

S. Huckemann, T. Hotz, and A. Munk. Intrinsic shape analysis: Geodesic PCA for riemannian manifolds modulo isometric Lie group actions. *Statistica Sinica*, 20:1–100, 2010.

B. E. Johnson. Separate continuity and measurability. *Proceedings of the American Mathematical Society*, 20(2):420–422, 1969.

I. Jolliffe. *Principal Component Analysis.* Springer, New York, 2002.

S. Joshi, B. Davis, M. Jomier, and G. Gerig. Unbiased diffeomorphic atlas construction for computational anatomy. *Neuroimage*, 23:151–160, 2004.

A. Kneip and K. J. Utikal. Inference for density families using functional principal component analysis. *J. Amer. Statist. Assoc.*, 96(454):519–542, 2001.

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.

M. Ledoux and M. Talagrand. *Probability in Banach spaces: Isoperimetry and processes.* Classics in Mathematics. Springer-Verlag, Berlin, 2011.

M. I. Miller and L. Younes. Group actions, homeomorphisms, and matching: A general framework. *International Journal of Computer Vision*, 41(1-2):61–84, 2001.

M. K. Murray and J. W. Rice. *Differential Geometry and Statistics.* Monographs on Statistics & Applied Probability. Chapman and Hall/CRC, 1993.

O. Museyko, M. Stiglmayr, K. Klamroth, and G. Leugering. On the application of the Monge–Kantorovich problem to image registration. *SIAM J. Imaging Sciences*, 2(4):1068–1097, 2009.

W. K. Newey and D. McFadden. Large Sample Estimation and Hypothesis Testing. In *Handbook of econometrics, Vol. IV*, volume 2 of *Handbooks in Econom.*, pages 2111–2245. North-Holland, Amsterdam, 1994.

X. Pennec. Intrinsic statistics on Riemannian manifolds: basic tools for geometric measurements. *J. Math. Imaging Vision*, 25(1):127–154, 2006.

B. Price. On the completeness of a certain metric space with an application to Blaschke's selection theorem. *Bull. Amer. Math. Soc.*, 46:278–280, 1940.

J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer, New York, 2005.

James O. Ramsay and Bernard W. Silverman. *Applied functional data analysis : methods and case studies*. Springer series in statistics. Springer, New York, Berlin, Heidelberg, 2002.

W. Rudin. Lebesgue's first theorem. In *Mathematical analysis and applications, Part B*, volume 7b of *Advances in Math. Suppl. Stud*, pages 741–747. Academic Press, 1981.

D. Rueckert, A. F. Frangi, and J. A. Schnabel. Automatic construction of 3D statistical deformation models using non-rigid registration. *IEEE Transactions on Medical Imaging*, 22:77–84, 2003.

M. Sdika. A fast nonrigid image registration with constraints on the Jacobian using large scale constrained optimization. *IEEE Transactions on Medical Imaging*, 27(2):271 –281, 2008.

H.L. Shang. A survey of functional principal component analysis. *AStA Advances in Statistical Analysis (Submitted)*, 2013.

X. Shi, M. Styner, J. Lieberman, J. G. Ibrahim, W. Lin, and H. Zhu. Intrinsic regression models for manifold-valued data. In *Proceedings of the 12th International Conference on Medical Image Computing and Computer-Assisted Intervention: Part II*, MICCAI '09, pages 192–199. Springer-Verlag, Berlin, Heidelberg, 2009.

B. W. Silverman. Smoothed functional principal components analysis by choice of norm. *Ann. Statist.*, 24(1):1–24, 1996.

S. Sommer, F. Lauze, S. Hauberg, and M. Nielsen. Manifold valued statistics, exact principal geodesic analysis and the effect of linear approximations. In *Proceedings of the 11th European conference on Computer vision: Part VI*, ECCV'10, pages 43–56. Springer-Verlag, Berlin, Heidelberg, 2010.

M. Sulman, J.F. Williams, M. F. Beg, and R. D. Russel. Volumetric image registration methods based on solving the monge ampere equation. *Canadian Applied Math Quarterly*, 17(3), 2009.

A. Trouvé and L. Younes. Local geometry of deformable templates. *SIAM J. Math. Anal.*, 37(1):17–59, 2005.

A. Trouvé and L. Younes. Metamorphoses through Lie group action. *Found. Comput. Math.*, 5(2):173–198, 2005.

M. Unser, A. Aldroubi, and M. Eden. B-Spline signal processing: Part I—Theory. *IEEE Transactions on Signal Processing*, 41(2):821–833, 1993a.

M. Unser, A. Aldroubi, and M. Eden. B-Spline signal processing: Part II—Efficient design and applications. *IEEE Transactions on Signal Processing*, 41(2):834–848, 1993b.

C. Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studiesm in Mathematics*. American Mathematical Society, 2003.

L. Wang, F. Beg, T. Ratnanather, C. Ceritoglu, L. Younes, J.C. Morris, J.G. Csernansky, and M.I. Miller. Large deformation diffeomorphism and momentum based hippocampal shape discrimination in dementia of the Alzheimer type. *IEEE Transactions on Medical Imaging*, 26(4):462 –470, 2007.

L. Younes. *Shapes and Diffeomorphisms*, volume 171 of *Applied Mathematical Sciences*. Springer-Verlag, Berlin, 2010.

Z. Zhang and H.-G. Müller. Functional density synchronization. *Comput. Stat. Data Anal.*, 55(7):2234–2249, 2011.

L. Zhu and A. Tannenbaum. Mass preserving registration for heart MR images. In *Department of Biomedical Engineering, Tsinghua University*, pages 147–154, 1998.

Appendix A

Appendix of Chapter 2

A.1 Quantile functions

Let us recall the definition of the quantile function of a probability measure, and let us state some of its properties.

Definition A.1. *Let μ be a probability measure on $\mathbb{R}$ with cdf F. The quantile function of μ is defined as $F^{-}(y) = \inf\{x \in \mathbb{R} : F(x) \geq y\},\ y \in (0, 1)$.*

The quantile function F^{-} corresponds to the so called generalized inverse of F.

Definition A.2. *let $T : \mathbb{R} \to \mathbb{R}$ be an increasing and left-continuous map. For $x \in \mathbb{R}$ define $T^{\dagger}(x) = \sup\{z \in \mathbb{R} : T(z) \leq x\}$.*

Proposition A.1. *Let μ be a probability measure on $\mathbb{R}$ and $T : \mathbb{R} \to \mathbb{R}$ be an increasing and left-continuous map. Denote by F the cdf of μ and by G the cdf of $T\#\mu$. Then*

(i) F^{-} is increasing, left-continuous and has a limit from the right.

(ii) $F(x) \geq y$ if and only if $x \geq F^{-}(y)$.

(iii) If μ is absolutely continuous with respect to the Lebesgue measure, then F^{-} is strictly increasing and $F(F^{-}(y)) = y$, for all $y \in (0, 1)$.

(iv) $T(z) \leq x$ if and only if $z \leq T^{\dagger}(x)$.

(v) $(F^{-})^{\dagger} = F$ and $(T^{\dagger})^{-} = T$.

(vi) $G^{-} = T \circ F^{-}$.

Proof. Properties (i), (ii) and (iii) are well known, see for instance [Embrechts and Hofert, 2013]. (iv) is analogous to (ii). (v) is direct from (ii) and (iv). Let us prove (vi). From (iv), $G(x) = \mu(z : T(z) \leq x) = \mu(z : z \leq T^{\dagger}(x)) = F(T^{\dagger}(x))$. Therefore, from (ii) and (iv), $G^{-}(y) = \inf\{x \in \mathbb{R} : F(T^{\dagger}(x)) \geq y\} = \inf\{x \in \mathbb{R} : T^{\dagger}(x) \geq F^{-}(y)\} = \inf\{x \in \mathbb{R} : x \geq T(F^{-}(y))\} = T(F^{-}(y))$. $\square$

A.2 Geodesics in metric spaces

We introduce the concept of geodesic in metric spaces. For notations, definitions and results, we follow [Chodosh, 2011] and references therein.

Definition A.3. *A curve in a metric space (X, d) is a continuous function $\gamma : I \to X$, where $I \subset \mathbb{R}$ is a closed (not necessarily bounded) interval. Also*

(i) γ is said to pass through $z \in X$ if $\gamma(t) = z$, for some $t \in I$;

(ii) γ joins $x, y \in X$ if there exists $a, b \in I, a \leq b$, such that $\gamma(a) = x$ and $\gamma(b) = y$.

(iii) γ is rectifiable if its length $L(\gamma)$ is finite.

For convenience, without loss of generality, we consider I such that $[0, 1] \subset I$.

Definition A.4. *A metric space (X, d) is said to be geodesic if for every $x, y \in X$, there exists a rectifiable curve γ joining x and y, such that $d(x, y) = L(\gamma)$. Such minimum length curve γ is called a shortest path between x and y. A curve $\gamma : I \to X$ is a geodesic if for every $t \in I$, there exist $a, b \in I, a < b, a \leq t \leq b$ such that the restriction of γ to $[a, b]$ is a shortest path between $\gamma(a)$ and $\gamma(b)$.*

The following is a useful characterization of shortest path (See [Chodosh, 2011] for a proof).

Lemma A.1. *For any shortest path, there exists a continuous reparametrization γ on $[0, 1]$ such that*

$$d(\gamma(s), \gamma(t)) = |t - s| d(\gamma(0), \gamma(1)) \text{ for all } s, t \in [0, 1].$$

Lemma A.2. *Let H be a Hilbert space and $x, y \in H$. Then γ is a shortest path joining x and y if and only if $\gamma(t) = (1 - t)x + ty$, for all $t \in [0, 1]$, up to a continuous reparametrization.*

Proof. Denote the inner product and the induced norm in H by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ respectively. Let γ be a shortest path between x and y, and $t \in [0, 1]$. From Lemma A.1 we have $\|x - \gamma(t)\| = t\|x - y\|$ and $\|\gamma(t) - y\| = (1 - t)\|x - y\|$, then

$$\|x - \gamma(t)\| + \|\gamma(t) - y\| = \|x - y\|.$$

Squaring and simplifying the expression above, we get

$$\|x - \gamma(t)\| \|\gamma(t) - y\| = \langle x - \gamma(t), \gamma(t) - y \rangle.$$

Hence, by the Cauchy-Schwartz inequality, there exists $\lambda \geq 0$ such that $x - \gamma(t) = \lambda(\gamma(t) - y)$. Finally, taking norm we find $\lambda = \frac{t}{1-t}$ and the result follows. The other implication is direct. $\square$

From the previous lemma, we deduce that in Hilbert spaces, any geodesic is locally a segment, therefore geodesics correspond to straight lines. We state this in the following corollary.

Corollary A.1. *Let H be a Hilbert space and $\gamma : I \to H$ a curve such that $\gamma(0) = x \in H$ and $\gamma(1) = y \in H$. Then, γ is a geodesic if and only if $\gamma(t) = (1 - t)x + ty$, for all $t \in I$, up to a continuous reparametrization.*

Definition A.5. *Let (X, d) be a geodesic space and $Y \subset X$. We say that Y is geodesic if the induced metric space (Y, d) is geodesic. In other words, if for any $x, y \in Y$, there exists a shortest path joining x and y, totally contained in Y.*

Note that, as direct consequence of Lemma A.2, any Hilbert space H is geodesic and $C \subset H$ is geodesic if and only if C is convex.

A.3 K-convergence

In this section we provide definitions and results that we use for proving the existence of principal geodesics (see Section 2.3.2). In particular, we define an appropriate concept of convergence for sequences of convex sets in a metric space.

Definition A.6. *Let (X, d) be a metric space and $C, C_n \subset X, n \geq 1$. We say that the sequence (C_n) converges to C in the sense of Kuratowski, denoted by $K\text{-}\lim_{n\to\infty} C_n = C$, if*

(i) for every $x \in C$ there exists a sequence (x_n) converging to x such that $x_n \in C_n, n \geq 1$;

(ii) for any sequence (x_n) such that $x_n \in C_n, n \geq 1$, any accumulation point of (x_n) belongs to C.

Definition A.7. *Let (X, d) be a metric space and $A, B \subset X$. The deviation from $x \in X$ to B is defined by $d(x, B) := \inf_{x' \in B} d(x, x')$; the deviation from A to B is $d(A, B) := \sup_{x \in A} d(x, B)$ and the Hausdorff distance between the sets A and B is*

$$h(A, B) := \max\{d(A, B), d(B, A)\}. \tag{A.1}$$

It is well known (see [Price, 1940; Beer, 1985] and references therein) that convergence with respect to the Hausdorff distance is stronger than convergence in the sense of Kuratowski. Moreover, if X is compact, then both notions of convergence coincide.

Definition A.8. *Let (X, d) be a metric space. We define the metric space*

$$\mathrm{CL}(X) := \{C \subset X \mid C \neq \emptyset, \text{ closed, bounded}\},$$

endowed with the Hausdorff distance between sets.

Proposition A.2. *Let (X, d) be a compact metric space and $y \in X$, then the function $C \mapsto d(y, C)$ is continuous in $\mathrm{CL}(X)$.*

Proof. Let $y \in X$ and let (C_n) be a sequence in $\mathrm{CL}(X)$, converging to $C \in \mathrm{CL}(X)$. Observe that, by the compactness of X, there exists $x^* \in C$ and $x_n^* \in C_n, n \geq 1$, such that $d(y, C) = d(y, x^*)$ and $d(y, C_n) = d(y, x_n^*), n \geq 1$. From (i) in Definition A.6, there exists a sequence (x_n) converging to x^*, such that $x_n \in C_n$, for every $n \geq 1$. Then

$$\limsup_{n\to\infty} d(y, C_n) \leq \limsup_{n\to\infty} d(y, x_n) = \lim_{n\to\infty} d(y, x_n) = d(y, x^*) = d(y, C). \tag{A.2}$$

Now, let $\vartheta_n := d(y, C_n)$ and let (ϑ_{n_j}) be a subsequence of (ϑ_n) such that $\liminf_{n\to\infty} \vartheta_n = \lim_{j\to\infty} \vartheta_{n_j}$. By compactness, $(x_{n_j}^*)$ converges, up to a subsequence, to an element $x \in X$.

From (ii) in Definition A.6, x belongs to C, therefore

$$\liminf_{n\to\infty} d(y, C_n) = \lim_{j\to\infty} d(y, C_{n_j}) = \lim_{j\to\infty} d(y, x^*_{n_j}) = d(y, x) \geq d(y, C). \tag{A.3}$$

From (A.2) and (A.3) we obtain the result. □

If X is a subset of a Hilbert space, then, for an integer $k \geq 1$, we denote by $\mathrm{CC}_k(X)$ the collection of convex sets C in $\mathrm{CL}(X)$, such that $\dim(C) \leq k$, where $\dim(C)$ is the dimension of the smaller affine subspace containing C. We denote also by $\mathrm{CC}_{k,0}(X)$ the collection of sets C in $\mathrm{CC}_k(X)$ such that $0 \in C$.

Proposition A.3. *If X is a compact subset of a Hilbert space, then $\mathrm{CL}(X)$, $\mathrm{CC}_k(X)$ and $\mathrm{CC}_{k,0}(X)$ are compact.*

Proof. The compactness of $\mathrm{CL}(X)$ is proved in [Price, 1940] and [Hai and An, 2013]. Let us prove that $\mathrm{CC}_k(X)$ is closed. Take a sequence (C_n) in $\mathrm{CC}_k(X)$ that converges to $C \in \mathrm{CL}(X)$, for the Hausdorff distance. From Blaschke's selection theorem in Banach spaces (see [Price, 1940] and [Hai and An, 2013]), we have that C is convex.
Let us now check by contradiction that $\dim(C) \leq k$. If we assume that that $\dim(C) > k$, then there exists $x_1, \ldots, x_{k+1} \in C$ linearly independent or, equivalently, the Gram determinant $\det(GM)$ of $x_1, \ldots, x_{k+1}$ (the Gram matrix GM has elements $GM_{i,j} = \langle x_i, x_j \rangle$, $i, j = 1, \ldots, k+1$) is non-zero. By (i) in Definition A.6, there exist $x_{1,n}, \ldots, x_{k+1,n} \in C_n$, for every $n \geq 1$, such that $(x_{j,n})$ converges to x_j for $j = 1, \ldots, k+1$. But $\dim(C_n) \leq k$, therefore the Gram determinant $\det(GM_n)$ of $x_{1,n}, \ldots, x_{k+1,n}$ is zero. But, it is easy to see that $\det(GM_n) \to \det(GM)$ which implies that $\det(GM) = 0$ (a contradiction) and so $\dim(C) \leq k$. We have proved that $C \in \mathrm{CC}_k(X)$ and so $\mathrm{CC}_k(X)$ is closed, hence compact subset of the compact space $\mathrm{CL}(X)$.
On the other hand, observe that if $C_n \to C$ in $\mathrm{CL}(X)$ and $0 \in C_n$, for all $n \geq 1$, then $0 \in C$, by (ii) in Definition A.6. We conclude that $\mathrm{CC}_{k,0}(X)$ is also closed, thus compact. □

Lemma A.3. *Let (X, d) be a metric space and $B, C, B_n, C_n \subset X, n \geq 1$, such that K-$\lim_{n\to\infty} B_n = B$, K-$\lim_{n\to\infty} C_n = C$ and $B_n \subset C_n$, $n \geq 1$. Then $B \subset C$.*

Proof. Take $x \in B$ and note that there exists a sequence (x_n) converging to x such that $x_n \in B_n, n \geq 1$, by (i) in Definition A.6. As $x_n \in B_n \subset C_n$, $n \geq 1$, from (ii) in Definition A.6, we have that $x \in C$. □

A.4 Γ−convergence

The proof of Theorem 2.4 relies on the notion of Γ-convergence of functions [Attouch, 1984; Dal Maso, 1993].

Definition A.9. *Let (X, d) be a metric space and (F_n) a sequence of functions from X into $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty, -\infty\}$. We say that (F_n) Γ-converges to $F : X \to \overline{\mathbb{R}}$, denoted Γ-$\lim_{n\to\infty} F_n = F$, if for every $x \in X$, it holds that*

(i) for every sequence (x_n) converging to x, $F(x) \leq \liminf_{n\to\infty} F_n(x_n)$ and

(ii) there exists a sequence (x_n) converging to x such that $F(x) = \lim_{n\to\infty} F_n(x_n)$.

Definition A.10. *For $F : X \to \overline{\mathbb{R}}$, we denote $M(F) = \{x \in X : F(x) = \inf_{y\in X} F(y)\}$.*

The following result (see [Dal Maso, 1993], Theorems 7.8 and 7.23) shows that Γ-convergence together with compactness (or more generally equicoercivity) implies convergence of minimum values and minimizers.

Theorem A.1. *Let (X, d) be a compact metric space and (F_n) a sequence of functions from X to $\overline{\mathbb{R}}$, Γ-converging to $F : X \to \overline{\mathbb{R}}$. Then $M(F)$ is nonempty and*

$$\lim_{n\to\infty} \inf_{x\in X} F_n(x) = \min_{x\in X} F(x)$$

Moreover, if $x_n \in M(F_n), n \geq 1$, then the accumulation points of (x_n) belong to $M(F)$.

Definition A.11. *Let (X, d) be a metric space and $A \subset X$. The indicator of A is the function $\chi_A : X \to \mathbb{R} \cup \{+\infty\}$ defined by $\chi_A(x) = 0$, if $x \in A$, and $\chi_A(x) = +\infty$, if $x \notin A$.*

The following Proposition ([Attouch, 1984], Proposition 4.15.) shows the relation between K-convergence (see Definition A.6) and Γ-convergence.

Lemma A.4. *Let (X, d) be a metric space and $A, A_n \subset X$, $n \geq 1$. Then* K-$\lim_{n\to\infty} A_n = A$ *if and only if* Γ-$\lim_{n\to\infty} \chi_{A_n} = \chi_A$.

A.5 Minkowski functional

In this section we recall the notion of Minkowski functional of a convex set, introduce the concept of reciprocal Minkowski functional and state some properties.

Definition A.12. *Let X be a vector space and $C \subset X$. The Minkowski functional of C is the function $p : X \to \mathbb{R}_+ \cup \{+\infty\}$, given by $p(x) = \inf\{t > 0 : x \in tC\}, x \in X$, with the convention $\inf \emptyset = +\infty$. Also, the reciprocal Minkowski functional of C is the function $q : X \to \mathbb{R}_+ \cup \{+\infty\}$, given by $q(x) = \sup\{t > 0 : tx \in C\}, x \in X$, with the convention $\sup \emptyset = 0$.*

Lemma A.5. *Let X be a vector space and $C \subset X$ a closed and convex set containing 0. Let p and q be the Minkowski and reciprocal Minkowski functionals of C, respectively (see Definition A.12). We have the following properties:*

(i) For any $x \in X$,

$$q(x) = \begin{cases} \infty & \text{if } p(x) = 0, \\ \frac{1}{p(x)} & \text{if } 0 < p(x) < \infty, \\ 0 & \text{if } p(x) = \infty. \end{cases} \tag{A.4}$$

(ii) For any $x \in X$,

$$\operatorname{span}(x) \cap C = \begin{cases} \{tx : t \in \mathbb{R}\} & \text{if } q(-x) = \infty \text{ and } q(x) = \infty, \\ \{tx : t \leq q(x)\} & \text{if } q(-x) = \infty \text{ and } q(x) < \infty, \\ \{tx : t \geq -q(-x)\} & \text{if } q(-x) < \infty \text{ and } q(x) = \infty, \\ \{tx : -q(-x) \leq t \leq q(x)\} & \text{if } q(-x), q(x) < \infty. \end{cases} \tag{A.5}$$

(iii) For any $t > 0$, $\{x \in X : p(x) \leq t\} = tC$.

(iv) For any $t > 0$, $\{x \in X : q(x) \geq t\} = t^{-1}C$.

Proof. (i) For $x \in X$ define $T_x = \{t > 0 : x \in tC\}$ and observe that

$$p(x) = \inf T_x. \tag{A.6}$$

We claim that T_x is convex and, if $p(x) > 0$, then T_x is a closed subset of $\mathbb{R}_+$. Indeed, let $s, t \in T_x$ and $r = \lambda s + (1-\lambda)t$, with $0 \leq \lambda \leq 1$. As $x \in sC$ and $x \in tC$, we have $x = \lambda x + (1-\lambda)x \in \lambda sC + (1-\lambda)tC = rC$. We conclude that T_x is convex. Assume now that $p(x) > 0$ and take $t_n \in T_x, n \geq 1$, such that t_n converges to $t \geq 0$. From (A.6) and as $p(x) > 0$, we have $t > 0$. Moreover, the sequence $t_n x$ belongs to the closed set C, thus its limits tx also belongs to C. Hence $t \in T_x$ and we conclude that T_x is closed. It is clear that T_x in nonempty if and only if $p(x) < \infty$, therefore, from the previous claim,

$$T_x = \begin{cases} (0,\infty) & \text{if } p(x) = 0, \\ [p(x),\infty) & \text{if } 0 < p(x) < \infty, \\ \emptyset & \text{if } p(x) = \infty. \end{cases} \tag{A.7}$$

Now, define $S_x = \{t > 0 : tx \in C\}$ and observe that

$$q(x) = \sup S_x. \tag{A.8}$$

Note also that $t \in T_x$ if and only if $1/t \in S_x$. Hence, from (A.7),

$$S_x = \begin{cases} (0,\infty) & \text{if } p(x) = 0, \\ (0,1/p(x)] & \text{if } 0 < p(x) < \infty, \\ \emptyset & \text{if } p(x) = \infty. \end{cases} \tag{A.9}$$

From (A.8) and (A.9) we obtain (A.4).

(ii) Observe that $\operatorname{span}(x) \cap C = \{tx : t \in \mathbb{R}, tx \in C\}$, therefore

$$\operatorname{span}(x) \cap C = \{tx : t \in S_{-x} \cup \{0\} \cup S_x\}. \tag{A.10}$$

On the other hand, from (A.8) and (i),

$$S_x = \begin{cases} \emptyset & \text{if } q(x) = 0, \\ (0,q(x)] & \text{if } 0 < q(x) < \infty, \\ (0,\infty) & \text{if } q(x) = \infty. \end{cases} \tag{A.11}$$

Thus (A.5) follows from (A.10) and (A.11).

(iii) Take $t > 0$ and denote $D := \{x \in X : p(x) \leq t\}$. Let us first show that $tC \subset D$. If $x \in tC$, then $t \in T_x$ and $p(x) = \inf T_x \leq t$, thus $x \in D$. Now, let us show that $D \subset tC$, so take $x \in D$. If $0 < p(x) \leq t$, by (A.7) we have $T_x = [p(x),\infty)$, therefore $t \in T_x$ and $x \in tC$. Finally, if $p(x) = 0$, then $x \in t'C$, for all $t' > 0$, so, in particular, $x \in tC$.

(iv) It follows from (i) and (iii). $\square$

The following properties are well known, see for instance [Fabian et al., 2011], Lemma 2.11.

Lemma A.6. *Let p be the Minkowski functional of a convex subset C of a normed vector space X. Then*

(i) $p(\lambda x) = \lambda p(x)$, *for all* $\lambda \geq 0$, $x \in X$.

(ii) $p(x + y) \leq p(x) + p(y)$, *for all* $x, y \in X$.

(iii) If C is a neighborhood of 0, then p is continuous. Moreover $\{x \in X : p(x) < 1\} = \mathrm{interior}(C) \subset C \subset \mathrm{closure}(C) = \{x \in X : p(x) \leq 1\}$.

Definition A.13. *A subset C of a vector space X is said to be absorbing if for every $x \in X$, there exists $t > 0$ such that $x \in tC$.*

Lemma A.7. *Let p be the Minkowski functional of a convex subset C of a Banach space. If C is absorbing, then p is continuous.*

Proof. Recall that, by a version of Baire's Lemma, a convex absorbing set in a Banach space is a neighborhood of 0. Then, the result follows from Lemma A.6(iii) □

Lemma A.8. *Let H be a Hilbert space and $C \subset H$ a closed and convex set containing 0. Let p and q be the Minkowski and reciprocal Minkowski functionals of C, respectively (see Definition A.12). Then p and $-q$ are lower semicontinuous (lsc) with respect in the weak topology $\mathcal{T}$ of H.*

Proof. Recall that a function $g : X \to \mathbb{R}_+ \cup \{+\infty\}$ defined, on a topological space X, is lsc if and only if, for all $t \geq 0$, the set $\{x \in X : g(x) \leq t\}$ is closed. Let us show that p is $\mathcal{T}$-lsc, based on the previous characterization. For $t > 0$, by Lemma A.5-(iii), we have $\{x \in H : p(x) \leq t\} = tC$, which is a (strong) closed convex set, therefore $\mathcal{T}$-closed. Also, $\{x \in H : p(x) \leq 0\} = \cap_{t>0}\{x \in H : p(x) \leq t\}$, which is $\mathcal{T}$-closed. Finally, if $t < 0$ then $\{x \in H : p(x) \leq t\} = \emptyset$. We conclude that p is $\mathcal{T}$-lsc. By Lemma A.5-(iv) and using analogous arguments, it can be shown that $-q$ is $\mathcal{T}$-lsc as well. □

Appendix B

Appendix of Chapter 3

B.1 Preliminary technical results

In this section, we give a deviation inequality for $\sup_{\boldsymbol{v}\in\boldsymbol{\mathcal{V}}_\mu} |D^g(\boldsymbol{v}) - M^g(\boldsymbol{v},\boldsymbol{Y})|$, under appropriate assumptions on the deformation operators and the additive noise in model (3.13), where $M^g(\boldsymbol{v},\boldsymbol{Y})$ and $D^g(\boldsymbol{v})$ are defined in (3.4) and (3.17) respectively.

Lemma B.1. *Consider the model* (3.13)*, with μ-regular deformation operator φ. Let*

$$Q(\boldsymbol{v}) = \frac{\epsilon^2}{n}\sum_{i=1}^{n}\int_{\Omega} W_i^2(\varphi(v_i,x))dx,\ \boldsymbol{v}\in\boldsymbol{\mathcal{V}}_\mu. \tag{B.1}$$

Then, for any $s>0$,

$$\mathbb{P}\left(\sup_{\boldsymbol{v}\in\boldsymbol{\mathcal{V}}_\mu} Q(\boldsymbol{v}) \geq A_\mu h(s,n,\epsilon)\right) \leq \exp(-s),$$

where A_μ is given in Definition 3.5 (i) and $h(s,n,\epsilon) = \epsilon^2\left(1+2\frac{s}{n}+2\sqrt{\frac{s}{n}}\right)$.

Proof. From Definition 3.5 (i), we have, for $\boldsymbol{v}\in\boldsymbol{\mathcal{V}}_\mu$,

$$Q(\boldsymbol{v}) \leq A_\mu\frac{\epsilon^2}{n}\sum_{i=1}^{n}\int_{\Omega} W_i^2(x)dx. \tag{B.2}$$

Let $g\in L^2(\Omega)$ and $K^W g(x) = \int_\Omega k(x,y)g(y)dy$ be the covariance operator of the random process W, where $k(x,y) = \mathbb{E}W(x)W(y)$ for $x,y\in\Omega$. Then there exist orthonormal eigenfunctions $(\phi_k)_{k\in\mathbb{K}}$ in $L^2(\Omega)$ with strictly positive eigenvalues $(w_k)_{k\in\mathbb{K}}$, such that $K^W\phi_k = w_k\phi_k$, with $w_1\geq w_2\geq\ldots>0$ and $\mathbb{K}=\{1,2,\ldots\}$. For any $i=1,\ldots,n$, the Gaussian process W_i can thus be decomposed as

$$W_i = \sum_{k\in\mathbb{K}} w_k^{1/2}\xi_{i,k}\phi_k,$$

where $\xi_{i,k} = w_k^{-1/2}\langle W_i,\phi_k\rangle_2$, $i=1\ldots,n$, $k\geq 1$, are i.i.d. standard Gaussian random variables, and so $\|W_i\|_2^2 = \sum_{k=1}^{+\infty} w_k\xi_{i,k}^2$. We have, from the assumptions on W, $\mathbb{E}\|W_i\|_2^2 =$

$\sum_{k=1}^{+\infty} w_k = 1 < +\infty$, and one can thus consider the following centered random variable

$$Z = \sum_{i=1}^{n} \sum_{k=1}^{+\infty} w_k (\xi_{i,k}^2 - 1).$$

Since the generating function of a χ^2 random variable, with one degree of freedom, is $\mathbb{E}\left(e^{s\xi_{i,k}^2}\right) = (1-2s)^{-1/2}$, for $0 < s < \frac{1}{2}$, it follows that

$$\log\left(\mathbb{E}\left(e^{tZ}\right)\right) = -n \sum_{k=1}^{+\infty} \left(tw_k + \frac{1}{2}\log(1-2tw_k)\right). \tag{B.3}$$

Then, using the inequality $-s - \frac{1}{2}\log(1-2s) \leq \frac{s^2}{1-2s}$, which holds for all $0 < s < \frac{1}{2}$, from (B.3) we obtain

$$\log\left(\mathbb{E}\left(e^{tZ}\right)\right) \leq n \sum_{k=1}^{+\infty} \frac{t^2 w_k^2}{1-2tw_k} \leq \frac{t^2 n}{1-2tw_1}\left(\sum_{k=1}^{+\infty} w_k\right)^2 = \frac{t^2 n}{1-2tw_1} < \infty,\ 0 < t < (2w_1)^{-1}.$$

Applying the exponential Chebyshev's inequality $\mathbb{P}\left(Z > \varepsilon\right) \leq \exp(-t\varepsilon)\mathbb{E}(\exp(tZ))$, $t > 0$, $\varepsilon > 0$, we obtain

$$\mathbb{P}\left(Z > \varepsilon\right) \leq \exp(-h(\varepsilon)),\ \varepsilon > 0,$$

where

$$h(\varepsilon) := \sup_{0 < t < (2w_1)^{-1}} \left\{t\varepsilon - \frac{t^2 n}{1-2tw_1}\right\}.$$

The supremum is It can be checked that $h(2w_1 s + 2\sqrt{ns}) = s$, hence

$$\mathbb{P}\left(Z > 2w_1 s + 2\sqrt{ns}\right) \leq \exp(-s). \tag{B.4}$$

By (B.2), it follows that

$$\sup_{\boldsymbol{v} \in \boldsymbol{\mathcal{V}}_\mu} Q(\boldsymbol{v}) \leq A_\mu \frac{\epsilon^2}{n} \sum_{i=1}^{n} \int_\Omega W_i^2(x)dx = A_\mu \frac{\epsilon^2}{n} \sum_{i=1}^{n} \sum_{k=1}^{+\infty} w_k \xi_{i,k}^2 = A_\mu \frac{\epsilon^2}{n}(Z+n).$$

Hence, it follows from (B.4) that

$$\mathbb{P}\left(\sup_{\boldsymbol{v} \in \boldsymbol{\mathcal{V}}_\mu} Q(\boldsymbol{v}) > A_\mu \frac{\epsilon^2}{n}\left(n + 2w_1 s + 2\sqrt{ns}\right)\right) \leq \exp(-s),$$

for any $s > 0$. The conclusion follows noting that $w_1 \leq \mathbb{E}\|W\|_2^2 = 1$. $\square$

Lemma B.2. *Consider the model* (3.13)*, with μ-regular deformation operator φ. Let*

$$D^t(\boldsymbol{v}) := \frac{1}{n} \sum_{i=1}^{n} \int_\Omega \left(f_i^*(\varphi(v_i, x)) - f^*(x)\right)^2 dx,\ \boldsymbol{v} \in \boldsymbol{\mathcal{V}}, \tag{B.5}$$

with $f_i^*(x) := f^*(\varphi^{-1}(V_i, x)), x \in \Omega,\ i = 1, \ldots, n$. Then

$$\mathbb{P}\left(\sup_{\boldsymbol{v} \in \boldsymbol{\mathcal{V}}_\mu} |D^t(\boldsymbol{v}) - M^t(\boldsymbol{v}, \boldsymbol{Y}, f^*)| > C\left(h(s,n,\epsilon) + \sqrt{h(s,n,\epsilon)}\right)\right) \leq \exp(-s),\ s > 0,$$

where M^t is defined in (3.3)*, $C > 0$ is a constant depending only on f^* and μ, and $h(s,n,\epsilon)$ is defined in Lemma B.1.*

Proof. For $\boldsymbol{v} \in \boldsymbol{\mathcal{V}} = \mathcal{V}^n$, let

$$R(\boldsymbol{v}) = 2\epsilon \frac{1}{n} \sum_{i=1}^{n} \int_{\Omega} (f_i^*(\varphi(v_i, x)) - f^*(x)) W_i^*(\varphi(v_i, x)) dx.$$

For any $\boldsymbol{v} \in \boldsymbol{\mathcal{V}}_\mu$, we have the decomposition

$$M^t(\boldsymbol{v}, \boldsymbol{Y}, f^*) = D^t(\boldsymbol{v}) + Q(\boldsymbol{v}) + R(\boldsymbol{v}), \tag{B.6}$$

where Q is defined in (B.1).

By applying the Cauchy-Schwartz inequality in $L^2(\Omega)$ and in $\mathbb{R}^n$ we obtain $R(\boldsymbol{v}) \leq 2\sqrt{D^t(\boldsymbol{v})}\sqrt{Q(\boldsymbol{v})}$. Also, the μ-regularity of φ implies $D^t(\boldsymbol{v}) \leq 4A_\mu^2 \|f^*\|_2^2$, and therefore $R(\boldsymbol{v}) \leq 4A_\mu \|f^*\|_2 \sqrt{Q(\boldsymbol{v})}$. Now, using the decomposition (B.6), one obtains

$$\sup_{\boldsymbol{v} \in \boldsymbol{\mathcal{V}}_\mu} |D^t(\boldsymbol{v}) - M^t(\boldsymbol{v}, \boldsymbol{Y}, f^*)| \leq \max(1, 4A_\mu \|f^*\|_2) \left(\sup_{\boldsymbol{v} \in \boldsymbol{\mathcal{V}}_\mu} Q(\boldsymbol{v}) + \sup_{\boldsymbol{v} \in \boldsymbol{\mathcal{V}}_\mu} \sqrt{Q(\boldsymbol{v})} \right)$$

and the result follows from Lemma B.1. □

Lemma B.3. *Consider the model* (3.13)*, with μ-regular deformation operator φ. Then*

$$\mathbb{P}\left(\sup_{\boldsymbol{v} \in \boldsymbol{\mathcal{V}}_\mu} |D^g(\boldsymbol{v}, \boldsymbol{V}) - M^g(\boldsymbol{v}, \boldsymbol{Y})| > C\left(h(s,n,\epsilon) + \sqrt{h(s,n,\epsilon)} \right) \right) \leq \exp(-s), \ s > 0,$$

where M^g and D^g are defined in (3.4) *and* (3.17) *respectively; $C > 0$ is a constant, depending only on f^* and μ, and $h(s,n,\epsilon)$ is defined in Lemma B.1.*

Proof. Let

$$Q^g(\boldsymbol{v}) = \frac{\epsilon^2}{n} \sum_{i=1}^{n} \int_{\Omega} \left(W_i(\varphi(v_i, x)) - \frac{1}{n} \sum_{j=1}^{n} W_j(\varphi(v_j, x)) \right)^2 dx, \ \boldsymbol{v} \in \boldsymbol{\mathcal{V}}$$

and

$$\begin{aligned} R(\boldsymbol{v}) = 2\epsilon \frac{1}{n} \sum_{i=1}^{n} \int_{\Omega} & \left(\frac{1}{n} \sum_{j=1}^{n} f_j^*(\varphi(v_j, x)) - f_i^*(\varphi(v_i, x)) \right) \\ & \times \left(\frac{1}{n} \sum_{j=1}^{n} W_j^*(\varphi(v_j, x)) - W_i^*(\varphi(v_i, x)) \right) dx, \ \boldsymbol{v} \in \boldsymbol{\mathcal{V}}. \end{aligned}$$

Then, for any $\boldsymbol{v} \in \boldsymbol{\mathcal{V}}_\mu$, we have the decomposition

$$M^g(\boldsymbol{v}, \boldsymbol{Y}) = D^g(\boldsymbol{v}) + Q^g(\boldsymbol{v}) + R(\boldsymbol{v}). \tag{B.7}$$

From the Cauchy-Schwartz inequality in $L^2(\Omega)$ and in $\mathbb{R}^n$, we have $R(\boldsymbol{v}) \leq 2\sqrt{D^g(\boldsymbol{v})}\sqrt{Q^g(\boldsymbol{v})}$. Also, from the μ-regularity of φ, we obtain $D^g(\boldsymbol{v}) \leq \frac{1}{n}\sum_{i=1}^n \int_\Omega \left(f_i^*(\varphi^{-1}(v_i, x))\right)^2 dx \leq A_\mu^2 \|f^*\|_2^2$. So $R(\boldsymbol{v}) \leq 2A_\mu \|f^*\|_2 \sqrt{Q^g(\boldsymbol{v})}$. Now, using the decomposition (B.7), one obtains

$$\sup_{\boldsymbol{v} \in \boldsymbol{\mathcal{V}}_\mu} |D^g(\boldsymbol{v}) - M^g(\boldsymbol{v}, \boldsymbol{Y})| \leq \max(1, 2A_\mu \|f^*\|_2) \left(\sup_{\boldsymbol{v} \in \boldsymbol{\mathcal{V}}_\mu} Q^g(\boldsymbol{v}) + \sup_{\boldsymbol{v} \in \boldsymbol{\mathcal{V}}_\mu} \sqrt{Q^g(\boldsymbol{v})} \right)$$

and the result follows from the fact that $Q^g \leq Q$ (see (B.1)) and Lemma B.1. □

Remark B.1. *Observe that Lemma B.2 and Lemma B.3 imply*

$$\operatorname*{plim}_{\epsilon} \sup_{\boldsymbol{v}\in\boldsymbol{\mathcal{V}}_\mu} |D^t(\boldsymbol{v}) - M^t(\boldsymbol{v},\boldsymbol{Y},f^*)| = \operatorname*{plim}_{\epsilon} \sup_{\boldsymbol{v}\in\boldsymbol{\mathcal{V}}_\mu} |D^g(\boldsymbol{v},\boldsymbol{V}) - M^g(\boldsymbol{v},\boldsymbol{Y})| = 0, \tag{B.8}$$

and

$$\operatorname*{plim}_{n,\epsilon} \sup_{\boldsymbol{v}\in\boldsymbol{\mathcal{V}}_\mu} |D^t(\boldsymbol{v}) - M^t(\boldsymbol{v},\boldsymbol{Y},f^*)| = \operatorname*{plim}_{n,\epsilon} \sup_{\boldsymbol{v}\in\boldsymbol{\mathcal{V}}_\mu} |D^g(\boldsymbol{v},\boldsymbol{V}) - M^g(\boldsymbol{v},\boldsymbol{Y})| = 0. \tag{B.9}$$

The proofs of Theorems 3.1, 3.2 and 3.3 rely on the following two propositions that establish the consistency of the registration procedures described in Section 3.2.2.

Proposition B.1. *Let $\hat{\boldsymbol{V}} \in \arg\min_{\boldsymbol{v}\in\boldsymbol{\mathcal{V}}_\mu} M^t(\boldsymbol{v},\boldsymbol{Y},f^*)$ be the parameters obtained from template registration of $\boldsymbol{Y}$ on f^*. Then,*

1. *under the hypotheses of Theorem 3.1, $\operatorname{plim}_{n,\epsilon} \bar{d}^2(\hat{\boldsymbol{V}},\boldsymbol{V}) = 0$, and*
2. *under the hypotheses of Theorem 3.2,*

$$\mathbb{P}\left(\bar{d}^2(\hat{\boldsymbol{V}},\boldsymbol{V}) > C\left(h(s,n,\epsilon) + \sqrt{h(s,n,\epsilon)}\right)\right) \le \exp(-s),\ s > 0,$$

where $C > 0$ is a constant, depending only on f^*, μ and $h(s,n,\epsilon)$ is defined in Lemma B.1.

Proof. Observe that $D^t(\boldsymbol{V}) = 0$ and so

$$D^t(\boldsymbol{v}) = D^t(\boldsymbol{v}) - D^t(\boldsymbol{V}) \le 2 \sup_{\boldsymbol{v}\in\boldsymbol{\mathcal{V}}_\mu} |M^t(\boldsymbol{v},\boldsymbol{Y},f^*) - D^t(\boldsymbol{v})|,\ \boldsymbol{v} \in \boldsymbol{\mathcal{V}}_\mu, \tag{B.10}$$

where D^t and M^t are defined in (B.5) and (3.3) respectively. On the other hand, from Definition 3.5 (i),

$$\frac{1}{n}\sum_{i=1}^n \|\varphi^*(V_i) - \varphi^*(v_i)\|_2^2 \le A_\mu D^t(\boldsymbol{v}),\ \boldsymbol{v} \in \boldsymbol{\mathcal{V}}_\mu.$$

Hence,

$$\frac{1}{n}\sum_{i=1}^n \|\varphi^*(V_i) - \varphi^*(\hat{V}_i)\|_2^2 \le 2A_\mu \sup_{\boldsymbol{v}\in\boldsymbol{\mathcal{V}}_\mu} |M^t(\boldsymbol{v},\boldsymbol{Y},f^*) - D^t(\boldsymbol{v})|. \tag{B.11}$$

We proceed now to prove part (i). From (B.11) and (B.8) we have $\operatorname{plim}_\epsilon \|\varphi^*(V_i) - \varphi^*(\hat{V}_i)\|_2^2 = 0$, that is, $\operatorname{plim}_\epsilon \varphi^*(\hat{V}_i) = \varphi^*(V_i)$, for $i = 1,\ldots,n$. From the continuity of φ^{*-1} we have $\operatorname{plim}_\epsilon \hat{V}_i = V_i$, for $i = 1,\ldots,n$, therefore

$$\operatorname*{plim}_{n} \operatorname*{plim}_{\epsilon} \frac{1}{n}\sum_{i=1}^n \|V_i - \hat{V}_i\|^2 = 0.$$

Now, the fact that $\|V_1 - \hat{V}_1\|^2$ is bounded by 2μ and tends to 0 in probability, as $\epsilon \to 0$, implies that $\mathbb{E}\|V_1 - \hat{V}_1\|^2 \to 0$, as $\epsilon \to 0$. Noting that $(V_i - \hat{V}_i)_{i\ge1}$ are i.i.d. (see Remark 3.1), we conclude from the weak law of large number that

$$\operatorname*{plim}_{\epsilon} \operatorname*{plim}_{n} \frac{1}{n}\sum_{i=1}^n \|V_i - \hat{V}_i\|^2 = 0,$$

thus proving part (i).

For (ii), inequality (B.11) and the fact that φ^{*-1} is uniformly Lipschitz, with constant $L(f^*, \mu) > 0$, implies

$$\bar{d}^2(\boldsymbol{u}, \boldsymbol{v}) \leq 2A_\mu L(f^*, \mu) \sup_{\boldsymbol{v} \in \boldsymbol{\mathcal{V}}} |M^t(\boldsymbol{v}, \boldsymbol{Y}, f^*) - D^t(\boldsymbol{v})|$$

and the result follows from Lemma B.2. $\square$

Proposition B.2. *Let $\hat{\boldsymbol{V}} \in \arg\min_{\boldsymbol{v} \in \boldsymbol{\mathcal{U}}} M^g(\boldsymbol{v}, \boldsymbol{Y})$ be the parameters obtained from groupwise registration of $\boldsymbol{Y}$. Then, under the hypotheses of Theorem 3.3,* $\operatorname{plim}_{n,\epsilon} \bar{d}^2(\hat{\boldsymbol{V}}, \boldsymbol{V}) = 0$.

Proof. Let $\boldsymbol{u}^*(\boldsymbol{V})$ be the unique minimizer of $D^g(u, \boldsymbol{V})$ on $\boldsymbol{\mathcal{U}}$, which exists because the model is g-identifiable. Since (by definition) $\hat{\boldsymbol{V}} \in \arg\min_{\boldsymbol{v} \in \boldsymbol{\mathcal{U}}} M^g(\boldsymbol{v}, \boldsymbol{Y})$, one obtains that

$$\begin{aligned} D^g(\hat{\boldsymbol{V}}, \boldsymbol{V}) - D^g(\boldsymbol{u}^*, \boldsymbol{V}) &\leq 2 \sup_{\boldsymbol{u} \in \boldsymbol{\mathcal{U}}} |M^g(\boldsymbol{u}, \boldsymbol{Y}) - D^g(\boldsymbol{u}, \boldsymbol{V})| \\ &\leq 2 \sup_{\boldsymbol{u} \in \boldsymbol{\mathcal{V}}_\mu} |M^g(\boldsymbol{u}, \boldsymbol{Y}) - D^g(\boldsymbol{u}, \boldsymbol{V})|. \end{aligned}$$

Therefore, from (B.9) and the g-identifiability of the model, we have $\operatorname{plim}_{n,\epsilon} \bar{d}^2(\hat{\boldsymbol{V}}, \boldsymbol{u}^*) = 0$. Also, the g-identifiability implies that $\operatorname{plim}_{n,\epsilon} \bar{d}^2(\boldsymbol{u}^*, \boldsymbol{V}) = 0$. Finally, the conclusion follows from the inequality $\bar{d}^2(\hat{\boldsymbol{V}}, \boldsymbol{V}) \leq 2\bar{d}^2(\hat{\boldsymbol{V}}, \boldsymbol{u}^*) + 2\bar{d}^2(\boldsymbol{u}^*, \boldsymbol{V})$. $\square$

In what follows, $\|\cdot\|_{HS}$ denotes the Hilbert-Schmidt norm of operators on a Hilbert space $\mathcal{H}$. Recall that, given an orthonormal basis $\{e_j\}_{j\geq 1}$ of $\mathcal{H}$, the Hilbert-Schmidt norm of an operator K is defined as $\|K\|_{HS}^2 = \sum_{j,k} \langle K(e_j), e_k \rangle^2$.

Lemma B.4. *Let $\mathcal{H}$ be a separable Hilbert space, with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\|\cdot\|$. Let $\{u_i\}_{i=1}^n$, $\{v_i\}_{i=1}^n$ in $B_r = \{h \in \mathcal{H} : \|h\| \leq r\}$, for some $r > 0$. Define the covariance operators $K_u, K_v : \mathcal{H} \to \mathcal{H}$ by $K_u(h) = \frac{1}{n}\sum_{i=1}^n \langle u_i - \bar{u}, h \rangle (u_i - \bar{u})$ and $K_v(h) = \frac{1}{n}\sum_{i=1}^n \langle v_i - \bar{v}, h \rangle (v_i - \bar{v})$, where $\bar{u} = \frac{1}{n}\sum_{i=1}^n u_i$ and $\bar{v} = \frac{1}{n}\sum_{i=1}^n v_i$. Then*

$$\|K_v - K_u\|_{HS}^2 \leq (6r)^2 \frac{1}{n} \sum_{i=1}^n \|v_i - u_i\|^2.$$

Proof. Let us define $\epsilon_i = v_i - u_i$, so $v_i = u_i + \epsilon_i$. Let $h \in \mathcal{H}$ and write

$$K_v(h) = \frac{1}{n} \sum_{i=1}^n \langle u_i - \bar{u} + \epsilon_i - \bar{\epsilon}, h \rangle (u_i - \bar{u} + \epsilon_i - \bar{\epsilon}) = K_u(h) + L(h) + L^*(h) + S(h),$$

where $L(h) = \frac{1}{n}\sum_{i=1}^n \langle u_i - \bar{u}, h \rangle (\epsilon_i - \bar{\epsilon})$, L^* is the adjoint the operator of L and $S(h) = \frac{1}{n}\sum_{i=1}^n \langle \epsilon_i - \bar{\epsilon}, h \rangle (\epsilon_i - \bar{\epsilon})$. Then, after some simple calculations, we get

$$\|L\|_{HS}^2 = \sum_{j\geq 1} \sum_{k \geq 1} \left(\frac{1}{n} \sum_{i=1}^n \langle u_i - \bar{u}, e_j \rangle \langle \epsilon_i - \bar{\epsilon}, e_k \rangle \right)^2 \leq \frac{1}{n^2} \sum_{i=1}^n \sum_{i'=1}^n \|u_i - \bar{u}\| \|u_{i'} - \bar{u}\| \|\epsilon_i - \bar{\epsilon}\| \|\epsilon_{i'} - \bar{\epsilon}\|.$$

Hence, since $u_i \in B_r, i = 1, \dots, n$, one has

$$\|L\|_{HS}^2 \leq \left(\frac{1}{n} \sum_{i=1}^{n} \|u_i - \bar{u}\| \|\epsilon_i - \bar{\epsilon}\| \right)^2 \leq (2r)^2 \left(\frac{1}{n} \sum_{i=1}^{n} \|\epsilon_i - \bar{\epsilon}\| \right)^2 \leq (2r)^2 \frac{1}{n} \sum_{i=1}^{n} \|\epsilon_i\|^2.$$

Similarly, since $\|\epsilon_i\| \leq \|u_i\| + \|v_i\| \leq 2r$,

$$\|S\|_{HS}^2 \leq \left(\frac{1}{n} \sum_{i=1}^{n} \|\epsilon_i - \bar{\epsilon}\|^2 \right)^2 \leq \left(\frac{1}{n} \sum_{i=1}^{n} \|\epsilon_i\|^2 \right)^2 \leq (2r)^2 \frac{1}{n} \sum_{i=1}^{n} \|\epsilon_i\|^2.$$

Finally, $\|K_v - K_u\|_{HS} \leq 2\|L\|_{HS} + \|S\|_{HS} \leq 6r \left(\sum_{i=1}^{n} \|\epsilon_i\|^2\right)^{\frac{1}{2}}$, which completes the proof. $\square$

The following theorem follows from the theory developed in [Bhatia et al., 1983; Davis and Kahan, 1970].

Theorem B.1. *Let $\mathcal{H}$ be a separable Hilbert space endowed with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. Let $A, \hat{A} : \mathcal{H} \to \mathcal{H}$ be self-adjoint Hilbert-Schmidt operators on $\mathcal{H}$, with eigenvalues/eigenvectors pairs $(\lambda_k, \phi_k)_{k \geq 1}$ and $(\hat{\lambda}_k, \hat{\phi}_k)_{k \geq 1}$ respectively. Then,*

$$\sup_{k \geq 1} |\lambda_k - \hat{\lambda}_k| \leq \|A - \hat{A}\|_{HS}. \tag{B.12}$$

Moreover, if $\hat{\delta}_k = \min_{k' \in \mathbb{K} \setminus \{k\}} |\lambda_k - \hat{\lambda}_{k'}| > 0$, then

$$\sin(\phi_k, \hat{\phi}_k) \leq \hat{\delta}_k^{-1} \|A - \hat{A}\|_{HS}. \tag{B.13}$$

B.2 Proofs of main results

B.2.1 Proof of Theorem 3.1

Proof. Let $\tilde{K}_n$ be the sample covariance operator of $V_1, \dots, V_n$, that is $\tilde{K}_n v = \frac{1}{n} \sum_{i=1}^{n} \langle V_i - \bar{V}_n, v \rangle (V_i - \bar{V}_n)$, with $\bar{V}_n = \frac{1}{n} \sum_{i=1}^{n} V_i$. Note that

$$\|\hat{K}_n - K\|_{HS}^2 \leq 2\|\hat{K}_n - \tilde{K}_n\|_{HS}^2 + 2\|\tilde{K}_n - K\|_{HS}^2, \tag{B.14}$$

The first term in the right-hand side of inequality (B.14) can be controlled by using Lemma B.4 and noting that $\|V_i\|, \|\hat{V}_i\| \leq \mu$, $i = 1, \dots, n$, that is,

$$\|\hat{K}_n - \tilde{K}_n\|_{HS}^2 \leq (6\mu)^2 \frac{1}{n} \sum_{i=1}^{n} \|\hat{V}_i - V_i\|^2. \tag{B.15}$$

Let us now bound the second term in the right-hand side of (B.14). To do so, remark that $\|V\|, \|V_i\| \leq \mu$, $i = 1, \dots, n$, and, thanks to a Bernstein's inequality for Hilbert-Schmidt operators (see e.g. [Bosq, 1998], Chapter 3), it follows that

$$\mathbb{P}\left(\|\tilde{K}_n - K\|_{HS} > \eta \right) \leq 2 \exp\left(-\frac{n\eta^2}{\tilde{C}(\mu)(1+\eta)} \right), \ \eta > 0, \tag{B.16}$$

for some constant $\tilde{C}(\mu) > 0$ depending only on μ. Hence, we can combine (B.14), (B.15) and (B.16) with Proposition B.1 (i) to obtain that, for any $\eta > 0$, $\lim_{n,\epsilon} \mathbb{P}\left(\|\hat{K}_n - K\|_{HS}^2 > \eta\right) = 0$, that is,

$$\operatorname*{plim}_{n,\epsilon} \hat{K}_n = K. \tag{B.17}$$

Now, from (B.12) and (B.17) we obtain $\operatorname{plim}_{n,\epsilon} \hat{\lambda}_k = \lambda_k, k \in \mathbb{K}$. For $k \in \mathbb{K}$ define the k-th empirical eigen-gap as

$$\hat{\delta}_k = \min_{k' \in \mathbb{K} \setminus \{k\}} |\lambda_k - \hat{\lambda}_{k'}|.$$

From (B.12), it holds that $\delta_k \leq \hat{\delta}_k + \max_{k'} |\lambda_{k'} - \hat{\lambda}_{k'}| \leq \hat{\delta}_k + \|\hat{K}_n - K\|_{HS}$. By (B.17), it follows that

$$\lim_{n,\epsilon} \mathbb{P}\left(\hat{\delta}_k > \frac{\delta_k}{2}\right) = 1. \tag{B.18}$$

Recalling that, from the specification of model (3.13), we have $\delta_k > 0$, hence inequality (B.13) implies

$$\begin{aligned}
\mathbb{P}\left(\sin(\hat{\phi}_k, \phi_k) > \eta\right) &\leq \mathbb{P}\left(\|\hat{K}_n - K\|_{HS}/\hat{\delta}_k > \eta\right) \\
&= \mathbb{P}\left(\|\hat{K}_n - K\|_{HS}/\hat{\delta}_k > \eta, \hat{\delta}_k > \delta/2\right) + \mathbb{P}\left(\|\hat{K}_n - K\|_{HS}/\hat{\delta}_k > \eta, \hat{\delta}_k \leq \delta/2\right) \\
&\leq \mathbb{P}\left(\|\hat{K}_n - K\|_{HS} > (\delta_k \eta)/2\right) + \mathbb{P}\left(\hat{\delta}_k \leq \delta/2\right).
\end{aligned}$$

From the above inequality, combined with (B.17) and (B.18), we obtain $\operatorname{plim}_{n,\epsilon} \sin^2(\hat{\phi}_k, \phi_k) = 0$. ☐

B.2.2 Proof of Theorem 3.2

Proof. Combining (B.14), (B.15) and (B.16) with Proposition B.1 (ii), we obtain

$$\mathbb{P}\left(\|\hat{K}_n - K\|_{HS}^2 > C \max(h(s,n,\epsilon) + \sqrt{h(s,n,\epsilon)}; g(s,n))\right) \leq 2\exp(-s), \; s > 0,$$

where $C > 0$ is a constant depending only on f^* and μ and $g(s,n) = \left(\frac{s}{n} + \sqrt{\frac{s^2}{n^2} + \frac{s}{n}}\right)^2$. Hence, from (B.12) we obtain

$$\mathbb{P}\left(|\hat{\lambda}_k - \lambda_k|^2 > C \max(h(s,n,\epsilon) + \sqrt{h(s,n,\epsilon)}; g(s,n))\right) \leq 2\exp(-s), \; s > 0.$$

Take now $s^* > 0$ such that $C \max(h(s^*,n,\epsilon) + \sqrt{h(s^*,n,\epsilon)}; g(s^*,n)) < (\delta_k/2)^2$. Then, thanks to (B.18) and (B.13) we obtain, for any $0 < s \leq s^*$,

$$\begin{aligned}
1 - 2\exp(-s) &\leq \mathbb{P}\left(\|\hat{K}_n - K\|_{HS}^2 < C \max(h(s,n,\epsilon) + \sqrt{h(s,n,\epsilon)}; g(s,n))\right) \\
&= \mathbb{P}\left(\|\hat{K}_n - K\|_{HS}^2 < C \max(h(s,n,\epsilon) + \sqrt{h(s,n,\epsilon)}; g(s,n)), \hat{\delta}_k > \delta_k/2\right) \\
&\leq \mathbb{P}\left((1/\hat{\delta}_k)^2 \|\hat{K}_n - K\|_{HS}^2 < (2/\delta_k)^2 C \max(h(s,n,\epsilon) + \sqrt{h(s,n,\epsilon)}; g(s,n))\right) \\
&\leq \mathbb{P}\left(\sin^2(\hat{\phi}_k, \phi_k) < (2/\delta_k)^2 C \max(h(s,n,\epsilon) + \sqrt{h(s,n,\epsilon)}; g(s,n))\right).
\end{aligned}$$

☐

B.2.3 Proof of Theorem 3.3

Proof. We proceed similarly as in the proof of Theorem 3.1. In the case of groupwise registration, inequalities (B.14), (B.15) and (B.16) are still valid, and can be combined with Proposition B.2 to obtain $\mathrm{plim}_{n,\epsilon} \hat{K}_n = K$. The rest of proof is identical to that of Theorem 3.1. □

B.3 Technical results for translation operators

Lemma B.5. *Let φ be defined by* (3.1)*, then*

1. $d_\Omega(\varphi(v,x), \varphi(v,y)) = d_\Omega(x,y)$*, for all* $x, y \in \Omega$ *and* $v \in \mathcal{V}$*.*
2. $d_C(\varphi(u,\cdot), \varphi(v,\cdot)) \leq \sum_{k=1}^{d} |u_k - v_k|$*, for all* $x \in \Omega$ *and* $u, v \in \mathcal{V}$*.*

Proof. Remark that, for any $a \in \mathbb{R}$, there exists a unique $k(a) \in \mathbb{Z}$ such that $\mathrm{mod}(a,1) = a + k(a)$. Then

$$\mathrm{mod}(a,1) - \mathrm{mod}(b,1) = a - b + k(a) - k(b).$$

Take $a, b \in \mathbb{R}$ such that $|a-b| < 1$ and assume that $a \geq b$. Since $a - b \in [0,1)$ and $\mathrm{mod}(a,1) - \mathrm{mod}(b,1) \in [-1,1]$, we obtain that

$$k(a) - k(b) = \begin{cases} 0 & \text{if } \mathrm{mod}(a,1) \geq \mathrm{mod}(b,1), \\ -1 & \text{if } \mathrm{mod}(a,1) < \mathrm{mod}(b,1). \end{cases}$$

Then

$$|\mathrm{mod}(a,1) - \mathrm{mod}(b,1)| = \begin{cases} a - b & \text{if } \mathrm{mod}(a,1) \geq \mathrm{mod}(b,1), \\ 1 - (a-b) & \text{if } \mathrm{mod}(a,1) < \mathrm{mod}(b,1) \end{cases}$$

We conclude that, for $a \geq b$,

$$\min\{|\mathrm{mod}(a,1) - \mathrm{mod}(b,1)|, 1 - |\mathrm{mod}(a,1) - \mathrm{mod}(b,1)|\} = \min\{|b-a|, 1-|b-a|\}. \quad \text{(B.19)}$$

Because of the symmetry in the expression above, we conclude that (B.19) is valid for any $a, b \in \mathbb{R}$, such that $|a-b| < 1$.

For the sake of simplicity, let us prove the lemma in the one-dimensional case (i.e. $d = 1$), where $d_\Omega(x,y) := \min\{|x-y|, 1-|x-y|\}$. Take $x, y \in \Omega$ and $u, v \in \mathcal{V}$. Part (i) is directly implied by (B.19), taking $a := x+v$ and $b := y+v$. For part (ii), note that $d_\Omega(x,y) \leq \frac{1}{2}$, hence if $|u-v| \geq 1$, then $d_\Omega(\varphi(u,x), \varphi(v,x)|) \leq \frac{1}{2} \leq |u-v|$. On the other hand, if $|u-v| < 1$ we can use (B.19) with $a := x+v$ and $b := x+u$ to obtain $d_\Omega(\varphi(u,x), \varphi(v,x)|) \leq d_\Omega(u,v) \leq |u-v|$. Finally, $d_C(\varphi(u,\cdot), \varphi(v,\cdot)) \leq |u-v|$. □

In order to prove Lemma 3.1 and Proposition 3.3 , denote by $e_\ell(x) = e^{i2\pi \sum_{k=1}^{d} \ell_k x_k}$, for $x = (x_1, \ldots, x_d) \in \Omega = [0,1]^d$ and $\ell = (\ell_1, \ldots, \ell_d) \in \mathbb{Z}^d$, the Fourier basis of $L^2([0,1]^d)$. Let $\theta_\ell = \int_\Omega f(x)\overline{e_\ell(x)}dx$, $\ell \in \mathbb{Z}^d$, be the Fourier coefficients of f^*. For $1 \leq k \leq d$, denote by $\ell^{(k)} = (\ell_1^{(k)}, \ldots, \ell_d^{(k)})$ the vector of $\mathbb{Z}^d$ such that $\ell_{k'}^{(k)} = 0$, for $k' \neq k$ and $\ell_k^{(k)} = 1$. Remark that, with this notation, $\theta_k = \theta_{\ell^{(k)}}$, where θ_k is defined in (3.18).

B.3.1 Proof of Lemma 3.1

Proof. Recall that $\varphi^*(v) = f^* \circ \varphi^{-1}(v, \cdot)$, $v \in \mathcal{V}_\mu$. For $u, v \in [-\rho, \rho]^d$ with $0 < \rho < 1/2$, we have

$$\begin{aligned}
\|f^*(\varphi^{-1}(u,x)) - f^*(\varphi^{-1}(v,x))\|_2^2 &\geq \sum_{k=1}^d |\theta_{\ell(k)} e^{-i2\pi u_k} - \theta_{\ell(k)} e^{-i2\pi v_k}|^2 \\
&= \sum_{k=1}^d |\theta_{\ell(k)}|^2 |e^{-i2\pi u_k} - e^{-i2\pi v_k}|^2. \qquad \text{(B.20)}
\end{aligned}$$

Then, by the mean value theorem, we have $|e^{-i2\pi u_k} - e^{-i2\pi v_k}|^2 = |\cos(2\pi u_k) - \cos(2\pi v_k)|^2 + |\sin(2\pi u_k) - \sin(2\pi v_k)|^2 \geq (2\pi)^2 cos^2(\rho)|u_k - v_k|^2$, for any $0 \leq u_k, v_k \leq \rho$. Hence,

$$\|f^*(\varphi^{-1}(u,x)) - f^*(\varphi^{-1}(v,x))\|_2^2 \geq (2\pi)^2 cos^2(\rho) \min_{0 \leq k \leq d} |\theta_{\ell(k)}|^2 \sum_{k=1}^d |u_k - v_k|^2.$$

□

B.3.2 Proof of Proposition 3.3

Proof. Remark that D^g, defined in (3.17), has the following expression in the Fourier domain:

$$D^g(\boldsymbol{u}, \boldsymbol{V}) = \frac{1}{n} \sum_{m=1}^n \left(\sum_{\ell \in \mathbb{Z}^d} \left| \frac{1}{n} \sum_{j=1}^n \theta_\ell e^{-i2\pi \langle \ell, V_j - u_j \rangle} - \theta_\ell e^{-i2\pi \langle \ell, V_m - u_m \rangle} \right|^2 \right), \ \boldsymbol{u} \in \boldsymbol{\mathcal{V}}. \qquad \text{(B.21)}$$

For $\boldsymbol{u} \in \boldsymbol{\mathcal{U}}_0$ we have

$$\begin{aligned}
D^g(\boldsymbol{u}, \boldsymbol{V}) \ &\geq \ \frac{1}{n} \sum_{m=1}^n \left(\sum_{k=1}^d |\theta_{\ell(k)}|^2 \left| \frac{1}{n} \sum_{j=1}^n e^{-i2\pi \left(V_j^{(k)} - u_j^{(k)} \right)} - e^{-i2\pi \left(V_m^{(k)} - u_m^{(k)} \right)} \right|^2 \right) \\
&\geq \ \sum_{k=1}^d |\theta_{\ell(k)}|^2 \left(1 - \left| \frac{1}{n} \sum_{m=1}^n e^{i2\pi \left(u_m^{(k)} - V_m^{(k)} \right)} \right|^2 \right). \qquad \text{(B.22)}
\end{aligned}$$

Further, remark that

$$\left| \frac{1}{n} \sum_{m=1}^n e^{i2\pi \left(u_m^{(k)} - V_m^{(k)} \right)} \right|^2 = \frac{1}{n} + \frac{2}{n^2} \sum_{m=1}^{n-1} \sum_{m'=m+1}^n \cos \left(2\pi \left(\left(u_m^{(k)} - V_m^{(k)} \right) - \left(u_{m'}^{(k)} - V_{m'}^{(k)} \right) \right) \right).$$

Let $0 \leq \alpha < 1/4$. Using a second-order Taylor expansion and the mean value theorem, one has that $\cos(2\pi u) \leq 1 - C(\alpha)|u|^2$, for any real u such that $|u| \leq \alpha$, with $C(\alpha) = 2\pi^2 \cos(2\pi\alpha)$. From the hypotheses, one has that $\left| \left(u_m^{(k)} - V_m^{(k)} \right) - \left(u_{m'}^{(k)} - V_{m'}^{(k)} \right) \right| \leq 2(\mu + \rho) < 1/4$. Therefore, for $\alpha = 2(\mu + \rho)$, it follows that

$$\begin{aligned}
\left| \frac{1}{n} \sum_{m=1}^n e^{i2\pi \left(u_m^{(k)} - V_m^{(k)} \right)} \right|^2 &\leq \frac{1}{n} + \frac{2}{n^2} \sum_{m=1}^{n-1} \sum_{m'=m+1}^n 1 - C(\alpha) \left| \left(u_m^{(k)} - V_m^{(k)} \right) - \left(u_{m'}^{(k)} - V_{m'}^{(k)} \right) \right|^2 \\
&\leq 1 - \frac{2}{n^2} \sum_{m=1}^{n-1} \sum_{m'=m+1}^n C(\alpha) \left| \left(u_m^{(k)} - V_m^{(k)} \right) - \left(u_{m'}^{(k)} - V_{m'}^{(k)} \right) \right|^2.
\end{aligned}$$

Hence, using the lower bound (B.22), it follows that, for $\boldsymbol{u} \in \boldsymbol{\mathcal{U}}_0$,

$$D^g(\boldsymbol{u}, \boldsymbol{V}) \geq 2C(\alpha)\frac{1}{n^2}\sum_{m=1}^{n-1}\sum_{m'=m+1}^{n}\left(\sum_{k=1}^{d}|\theta_{\ell^{(k)}}|^2\left|\left(u_m^{(k)} - V_m^{(k)}\right) - \left(u_{m'}^{(k)} - V_{m'}^{(k)}\right)\right|^2\right). \quad \text{(B.23)}$$

The following identity is obtained from elementary algebraic manipulations and the fact that $\boldsymbol{u} \in \boldsymbol{\mathcal{U}}_0$ ($\sum_{m=1}^{n} u_m^{(k)} = 0$).

$$\frac{1}{n}\sum_{m=1}^{n-1}\sum_{m'=m+1}^{n}\left|\left(u_m^{(k)} - V_m^{(k)}\right) - \left(u_{m'}^{(k)} - V_{m'}^{(k)}\right)\right|^2 = \sum_{m=1}^{n}\left|u_m^{(k)} - (V_m^{(k)} - \bar{V}_n^{(k)})\right|^2,$$

where $\bar{V}_n^{(k)} = \frac{1}{n}\sum_{m=1}^{n} V_m^{(k)}$. Inserting the above equality in (B.23), we finally obtain

$$D^g(\boldsymbol{u}, \boldsymbol{V}) \geq C_0(f^*, \mu)\frac{1}{n}\sum_{m=1}^{n}\sum_{k=1}^{d}\left|u_m^{(k)} - \tilde{V}_m^{(k)}\right|^2, \quad \text{(B.24)}$$

with $C_0(f^*, \mu) = 2C(\alpha)\min_{1\leq k\leq d}\left\{|\theta_{\ell^{(k)}}|^2\right\}$ and $\tilde{V}_m^{(k)} = V_m^{(k)} - \bar{V}_n^{(k)}$. Thanks to the assumption $\theta_{\ell^{(k)}} \neq 0$, for all $1 \leq k \leq d$, it follows that $C_0(f^*, \mu) > 0$. The inequality $\mu \geq 2\rho$, implies that $|\tilde{V}_m^{(k)}| = |V_m^{(k)} - \bar{V}_n^{(k)}| \leq 2\rho \leq \mu$, for any $1 \leq k \leq d$ and all $1 \leq m \leq n$, therefore, $\boldsymbol{u} \in \boldsymbol{\mathcal{U}}_0$. Then, using inequality (B.24) and $D^g(\boldsymbol{u}^*, \boldsymbol{V}) = 0$, the proof is completed. $\square$

B.4 Technical results for non-rigid diffeomorphic operators

B.4.1 Proof of Proposition 3.4

As in Section 3.3.4, let $\Omega = [0, 1]$ and $\mathcal{V}$ be a Hilbert space of smooth functions, continuously embedded in $C_0^1(\Omega)$. Also, for $t \geq 0$ and $x \in \Omega$, consider the ODE

$$\phi_t = x + \int_0^t v(\phi_s)ds, \quad \text{(B.25)}$$

which is equivalent to (3.2), with initial condition $\phi_0 = x \in \Omega$. It is well known that there exists a unique solution $\phi_t^v(x)$ of (B.25), moreover $x \mapsto \phi_t^v(x)$ is a diffeomorphism on Ω and $t \to \phi_t^v(x)$ is continuous; see e.g. [Younes, 2010]. Observe that, $\varphi(v, \cdot) = \phi_1^v$.

Lemma B.6 (One-parameter semigroup)**.** *For $v \in \mathcal{V}$, the family $\phi_t^v, t \geq 0$, is a one-parameter semigroup over Ω. That is, $\phi_t^v \circ \phi_s^v = \phi_{t+s}^v$, for all $s, t \geq 0$ and $\phi_0^v = id$, where id is the identity map in Ω.*

Proof. From (B.25) it immediately follows that $\phi_0^v(x) = x$, for all $x \in \Omega$. Take $x' \in \Omega$ and observe that $t \to \phi_t^v \circ \phi_s^v(x')$ and $t \to \phi_{t+s}^v(x')$ are solutions of (B.25), with initial condition $x := \phi_s^v(x')$, hence the result follows from the uniqueness of the solution. $\square$

The proof of Proposition 3.4 is based on the following lemma.

Lemma B.7. *Let $u, v \in \mathcal{V}$ and $\bar{x} \in \Omega$.*

(i) If $u(\bar{x}) = 0$, then $\phi_t^u(\bar{x}) = \bar{x}$, for all $t \geq 0$.

(ii) If $u(\bar{x}) > 0$, then $\phi_t^u(\bar{x}) > \bar{x}$, for all $t > 0$.

(iii) If $u(x) \geq v(x)$ for all $x \in [\bar{x}, 1]$, then $\phi_t^u(\bar{x}) \geq \phi_t^v(\bar{x})$, for all $t \geq 0$.

(iv) If $u(\bar{x}) > 0$ and $u(x) \geq v(x)$, for all $x \in (\bar{x}, 1]$, then $\phi_t^u(\bar{x}) > \phi_t^v(\bar{x})$, for all $t > 0$.

Proof. (i) The function $\phi_t := \bar{x}, t \geq 0$ satisfies the O.D.E $\phi_t = \bar{x} + \int_0^t u(\phi_s)ds$, so the result follows from the uniqueness of the solution.
(ii) Consider the function $g(t) := \phi_t^u(\bar{x})$, $t \geq 0$ and observe that $g(0) = \bar{x}$ and $g'(0) = u(\phi_0^u(\bar{x})) = u(\bar{x}) > 0$. As g is continuous, there exists $s > 0$ such that

$$\phi_t^u(\bar{x}) > \bar{x}, \; t \in (0, s]. \tag{B.26}$$

From Lemma (B.6), (B.26) and since ϕ_{t-s}^u is strictly increasing for $t \geq s$, we obtain

$$\phi_t^u(\bar{x}) = \phi_{t-s}^u(\phi_s^u(\bar{x})) > \phi_{t-s}^u(\bar{x}) > \bar{x}, \; t \in (s, 2s].$$

Following the previous arguments we obtain by induction that $\phi_t^u(\bar{x}) > \bar{x}$, $t \in ((k-1)s, ks]$, for any $k \in \mathbb{N}$, which concludes the proof.
(iii) Consider the function $g(t) := \phi_t^u(\bar{x}) - \phi_t^v(\bar{x})$, $t \geq 0$, and suppose there exists $t_0 > 0$ such that $g(t_0) < 0$. Thanks to the continuity of g and from the fact that $g(0) = 0$, the set $S := \{t \in [0, t_0] : g(t) = 0\}$ is closed and nonempty, so we can define $s := \sup S$. Since $g(s) = 0$, it follows that $\phi_s^u(\bar{x}) = \phi_s^v(\bar{x})$ and thus

$$g'(s) = u(\phi_s^u(\bar{x})) - v(\phi_s^v(\bar{x})) = u(\phi_s^u(\bar{x})) - v(\phi_s^u(\bar{x})). \tag{B.27}$$

From (i) and (ii) we have $\phi_s^u(\bar{x}) \geq \bar{x}$ and from (B.27) we deduce that $g'(s) \geq 0$. But $g(t_0) < 0$, so there must exist $r \in (s, t_0)$ such that $g(r) = 0$, which is a contradiction by definition of s. Therefore, we conclude that $g(t) \geq 0$, for all $t \geq 0$, which proves the result.
(iv) Let again $g(t) := \phi_t^u(\bar{x}) - \phi_t^v(\bar{x})$, $t \geq 0$, and note that that $g(0) = 0$ and $g'(0) = u(\bar{x}) - v(\bar{x}) > 0$. Hence, there exists $s > 0$ such that $g(t) > 0, t \in (0, s]$ i.e,

$$\phi_t^u(\bar{x}) > \phi_t^v(\bar{x}), \; t \in (0, s]. \tag{B.28}$$

From the previous inequality and the fact that ϕ_{t-s}^v, for $t \geq s$, is strictly increasing, we obtain

$$\phi_{t-s}^v(\phi_s^u(\bar{x})) > \phi_{t-s}^v(\phi_s^v(\bar{x})), \; t \geq s. \tag{B.29}$$

Let $\tilde{x} := \phi_s^u(\bar{x})$ and note that $\tilde{x} > \bar{x}$, thanks to (ii). Hence, applying (iii) to the point $\tilde{x}$, we obtain

$$\phi_{t-s}^u(\phi_s^u(\bar{x})) \geq \phi_{t-s}^v(\phi_s^u(\bar{x})), \; t \geq s. \tag{B.30}$$

From Lemma B.6, (B.29) and (B.30),

$$\phi_t^u(\bar{x}) > \phi_t^v(\bar{x}), t \geq s. \tag{B.31}$$

Thus, the result follows from (B.28) and (B.31). □

Proof of Proposition 3.4. Let $u, v \in \mathcal{V}$ with $u \neq v$. Since $\varphi(u, \cdot)$ and $\varphi(v, \cdot)$ are continuous, we have to show that there exists $\bar{x} \in \Omega$, such that $\varphi(u, \bar{x}) \neq \varphi(v, \bar{x})$. As $w := u - v$ is non-oscillatory, according to Definition 3.7, there exists $0 \leq a < b \leq 1$ such that $w(a) = 0$, $w(x) \neq 0, x \in (a, b)$ and $w(x) = 0, x \in (b, 1]$. Hence, without loss of generality, we have that $u(a) = v(a)$, $u(x) > v(x), x \in (a, b)$ and $u(x) = v(x), x \in (b, 1]$, thanks to the continuity of u and v. Let us analyze the cases $u(a) \geq 0$, $u(a) < 0$ and $u(a) = 0$.
Case $u(a) > 0$: take $\bar{x} \in (a, b)$ such that $u(\bar{x}) > 0$ and note that $u(\bar{x}) > v(\bar{x})$ and $u(x) \geq v(x)$, $x \in (\bar{x}, 1]$. From Lemma B.7(iv), $\varphi(u, \bar{x}) > \varphi(v, \bar{x})$ and the proof is concluded in this case.
Case $u(a) < 0$: let now $\tilde{x} \in (a, b)$ such that $u(\tilde{x}) < 0$. With the notation $v^* = -v$ and $u^* = -u$, we have $v^*(\tilde{x}) > u^*(\tilde{x}) > 0$ and $v^*(x) \geq u^*(x)$, $x \in (\tilde{x}, 1]$. Hence, by Lemma B.7(iv), we obtain

$$\varphi(v^*, \tilde{x}) > \varphi(u^*, \tilde{x}). \tag{B.32}$$

Recall that $\varphi(v^*, \cdot) = (\varphi(v, \cdot))^{-1}$ and let $\bar{x} := \varphi(v^*, \tilde{x})$. So, from (B.32), we have

$$\varphi(u^*, \varphi(u, \bar{x})) = \bar{x} = \varphi(v^*, \tilde{x}) > \varphi(u^*, \tilde{x}) = \varphi(u^*, \varphi(v, \bar{x})). \tag{B.33}$$

As $\varphi(u, \cdot)$ is strictly increasing and $\varphi(u^*, \cdot) = (\varphi(u, \cdot))^{-1}$, from (B.33) we obtain $\varphi(u, \bar{x}) > \varphi(v, \bar{x})$, which concludes the proof for the case $u(a) < 0$.
Case $u(a) = 0$: in this case, the result can be obtained from the two previous cases. $\square$

B.4.2 Proof of Proposition 3.5

The proof of Proposition 3.5 make use of two lemmas that we state below. Let us consider a closed interval $\mathcal{D} = [a, b] \subset \mathbb{R}$, with $a < b$, and denote by $\mathrm{Diff}(\mathcal{D})_+$ the group of increasing C^1 diffeomorphisms of $\mathcal{D}$. Recall that a continuous function from $\mathcal{D}$ to $\mathcal{D}$ is one-to-one if and only if it is strictly monotone, hence we can characterize $\mathrm{Diff}(\mathcal{D})_+$ as

$$\mathrm{Diff}(\mathcal{D})_+ = \{\phi : \mathcal{D} \to \mathcal{D} : \phi(a) = a,\ \phi(b) = b, \phi \text{ is } C^1 \text{ and } \phi \text{ is strictly increasing}\}.$$

Lemma B.8. *Let $\phi \in \mathrm{Diff}(\mathcal{D})_+$ such that $\phi(x) > x$, for all $x \in (a, b)$. The only continuous functions $f : \mathcal{D} \to \mathbb{R}$ satisfying the equality $f \circ \phi = f$ are the constant functions.*

Proof. For an integer $n \geq 2$, let $\phi^n = \phi^{n-1} \circ \phi$, with the convention $\phi^1 = \phi$. Let us define the *$w - limit$* set of $x \in \mathcal{D}$ as the set of accumulations points of the sequence $(\phi^n(x))_{n \in \mathbb{Z}_+}$, that is

$$w(x) = \{y \in \mathcal{D} : \lim_{k \to \infty} \phi^{n_k}(x) \to y \text{ for some subsequence } n_k \to \infty\}.$$

Let us take $x \in (a, b)$ and remark that the sequence $(\phi^n(x))_{n \in \mathbb{Z}_+}$ is strictly increasing and bounded from above by b, hence $w(x)$ is a singleton. Moreover it is clear that $w(x)$ is totally invariant, that is, $\phi(w(x)) = w(x)$, hence $w(x) = \{b\}$. On the other hand, if $f \circ \phi = f$ then $f(\phi^n(x)) = f(x)$, for all $n \in \mathbb{Z}_+$. Hence, by letting $n \to \infty$ and using the continuity of f, we conclude that $f(x) = f(b)$. Therefore, we have proved that $f(x) = f(b)$, for any $x \in (a, b)$, which implies that $f(x)$ is constant over $\mathcal{D}$, and completes the proof. $\square$

Lemma B.9. *Assume that $f : \mathcal{D} \to \mathbb{R}$ is continuous and does not have flat regions (see Definition 3.8), then $I(f) = \{id\}$, where*

$$I(f) := \{\phi \in \mathrm{Diff}(\mathcal{D})_+ : f \circ \phi = f\}$$

is the set of deformations in $\mathrm{Diff}(\mathcal{D})_+$ *that act invariantly on* f *and* $id : \mathcal{D} \to \mathcal{D}$ *denotes the identity function, namely* $id(x) = x$*, for all* $x \in \mathcal{D}$*.*

Proof. It is clear that $\{id\} \subset I(f)$, so let us prove the other inclusion. Let $\phi \in I(f)$ and suppose there exists $\tilde{x} \in \mathcal{D}$ such that $\phi(\tilde{x}) \neq \tilde{x}$. Without loss of generality we will assume that $\phi(\tilde{x}) > \tilde{x}$. As ϕ is continuous, the sets $K_1 = \{x \in [0, \tilde{x}) \mid \phi(x) = x\}$ and $K_2 = \{x \in (\tilde{x}, 1] \mid \phi(x) = x\}$ are compact. Therefore, there exists $\tilde{a} = \max\{x : x \in K_1\}$, $\tilde{b} = \min\{x : x \in K_2\}$ satisfying $a \leq \tilde{a} < \tilde{x} < \tilde{b} \leq b$, $\phi(\tilde{a}) = \tilde{a}$ and $\phi(\tilde{b}) = \tilde{b}$. Let $\tilde{f}$ and $\tilde{\phi}$ be the restriction of f and ϕ to $[\tilde{a}, \tilde{b}]$, respectively. By the continuity of ϕ it is clear that $\tilde{\phi}(x) > x$, for all $x \in (\tilde{a}, \tilde{b})$ and $\tilde{\phi}([\tilde{a}, \tilde{b}]) = [\tilde{a}, \tilde{b}]$. Since $\tilde{f} \circ \tilde{\phi} = \tilde{f}$ and $\tilde{f}$ is a continuous function, Lemma B.8 implies that $\tilde{f}$ is constant over $[\tilde{a}, \tilde{b}]$, which contradicts the assumption that f does not have flat regions. Therefore, we conclude that $\phi(x) = x$, for all $x \in \mathcal{D}$, which completes the proof. $\square$

Proof of Proposition 3.5. If $\varphi^*(u) = \varphi^*(v)$, then $f = f \circ \varphi(u, \cdot) \circ (\varphi(v, \cdot))^{-1}$. As $\varphi(u, \cdot) \circ (\varphi(v, \cdot))^{-1} \in \mathrm{Diff}(\Omega)_+$, then, by Lemma B.9, we obtain that $\varphi(u, \cdot) = \varphi(v, \cdot)$. $\square$