Procrustes Metrics and Optimal Transport for Covariance Operators

Thèse N° 9418

Présentée le 17 mai 2019

à la Faculté des sciences de base
Chaire de statistique mathématique
Programme doctoral en mathématiques

pour l'obtention du grade de Docteur ès Sciences

par

Valentina MASAROTTO

Acceptée sur proposition du jury

Prof. S. Morgenthaler, président du jury
Prof. V. Panaretos, directeur de thèse
Prof. I. Dryden, rapporteur
Dr D. Kraus, rapporteur
Prof. A. Davison, rapporteur

2019

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

As you set out on the way to Ithaca
hope that the road is a long one,
filled with adventures, filled with understanding.
The Laestrygonians and the Cyclopes,
Poseidon in his anger: do not fear them,
you'll never come across them on your way
as long as your mind stays aloft, and a choice
emotion touches your spirit and your body.
The Laestrygonians and the Cyclopes,
savage Poseidon; you'll not encounter them
unless you carry them within your soul,
unless your soul sets them up before you.

Hope that the road is a long one.
Many may the summer mornings be
when—with what pleasure, with what joy—
you first put in to harbors new to your eyes;
may you stop at Phoenician trading posts
and there acquire fine goods:
mother-of-pearl and coral, amber and ebony,
and heady perfumes of every kind:
as many heady perfumes as you can.
To many Egyptian cities may you go
so you may learn, and go on learning, from their sages.

Always keep Ithaca in your mind;
to reach her is your destiny.
But do not rush your journey in the least.
Better that it last for many years;
that you drop anchor at the island an old man,
rich with all you've gotten on the way,
not expecting Ithaca to make you rich.

Ithaca gave to you the beautiful journey;
without her you'd not have set upon the road.
But she has nothing left to give you any more.

And if you find her poor, Ithaca did not deceive you.
As wise as you'll have become, with so much experience,
you'll have understood, by then, what these Ithacas mean.

(Ithaca, C.P. Cavafy, translated by D. Mendelsohn)

Acknowledgements

I would like to start by expressing my gratitude to my supervisor, Professor Victor Panaretos. His patience and his availability in explaining and answering my many questions were essential to the completion of this thesis. As there were his feedback on this manuscript and his encouragements in the (admittedly numerous) moments when I felt nothing was moving.

I would like to thank in advance Professors Anthony C. Davison, Ian Dryden and David Kraus for having kindly accepted to be part of my thesis committee, and Professor Stephan Morgenthaler for having accepted to be the president of the jury.

Administrative help from Anna, Maroussia, Jocelyne and Nadia was throughout important, especially when sleep-deprivation took over the organisational skills.

Many thanks go to my colleagues, current and past, for making my EPFL days enjoyable: Shahin, Mikael, Pavol, Kate, Guillaume, Anirvan, Andrea, Matthieu, Laya, Kartik, Tomas R., Tomas M. and Neda (thanks for stepping in for me at the most crucial time!). A special mention goes to Yoav, for his limitless kindness in facing each and every one of my doubts, and for the exchanges over the struggles and joys (but mostly the struggles) of parenthood.
The months I have spent in the same office as Susan and Marie were the most entertaining of my Ph.D. life. Such a good office synergy will be hard to match, no matter whom I will share the next office with.
Talking of offices, I carry very fond memories of my time in the LTS workspace with Vassilis and Natanael. I hope to meet you both soon in another “lunch for four” or on some marathon dance floor.

Combining research and motherhood meant I mostly felt at the wrong place at all times. Sharing my struggle with somebody who could really feel it, was soothing. For this I must thank Elisabet.
Giulia was the perfect companion during fakely-fit lunch breaks and evenings full of tales and gossip and discussions on every topic: from sewing courses to the love for the Netherlands, passing through statistics. I missed you dearly when you left UNIL, and even more when you left Lausanne.
The "Picnic" group and the restricted "trust nights" one are a truly wonderful group of people. Engaged, interesting, fun in ways that really speak to me. I wish I had more time to share with

them, but no matter how sparse, every occasion has been nurturing and deeply enjoyable. Among them, Paolo and Damiano stand out. Paolo for his quiet support and the inspiration he gives. I can hardly think of somebody more diversely meaningful. And Damiano because there has hardly be any daily life, at any stage of University, that did not involve him. The companionship of our daily meetings was very precious.

Having lived in different countries means that I do not have the luxury to have all the people I love close to me. But regardless of the distance, some people have been really invaluable to me throughout these years, and even more throughout the difficulties of the last few months. Marta, Gianfranco, Eleonora and Clara have always welcomed me and loved me as if I was really family, and I will always be grateful. Phone calls with Piera made me laugh no matter what crisis my world was going through, but I especially thank her for the occasional shrewd perspective on things, including motherhood. Vera is a jewel. No matter the problem: from life lessons to papers to submit, from personal struggles to mud baths, she was there to listen, bringing lightness and wisdom and friendship and giggles. As for Alice. I think I have exhausted the words to convey the richness she brought to my life over the last 17 years. I am so grateful for your colourful, family-filled presence and continuous display of affection.
And thanks to Pantelis. Your generosity and support were a gift, as it was the (one of many) lesson learnt: ask. You are empowering, and forever surprising.

A dedicated mention goes to Luigi's family, because they welcomed me as one of them. Their kindness and the food-filled seaside holidays in winter and summer were genuinely recharging. Finally, my family: my brother, and my mum and dad. I do not think it has always been easy on them to stand by my side, but they truly and daily showed me the meaning of unconditional love, and unwavering support. I would not be writing this today if it was not for them, and this is literal in so many ways. Grazie mamma for endlessly giving yourself, e grazie papà for being my personal superhero, especially over the last year.
But family is also, or above all, what I have build on my own. So I want to thank Luigi for building these last 5 years with me and for sharing this sometimes bumpy piece of road. We came a long way together.
Lastly, thanks to Lorenzo. For giving meaning and perspective and joy and light to everything. Your imagination, your resilience and your level of understanding will never cease to impress me. Being your mother is truly a privilege.

Lausanne, 18.01.2019 V. M.

Abstract

Covariance operators play a fundamental role in functional data analysis, providing the canonical means to analyse functional variation via the celebrated Karhunen-Loéve expansion. These operators may themselves be subject to variation, for instance in contexts where multiple functional populations are to be compared. Statistical techniques to analyse such variation are intimately linked with the choice of metric on the space of such operators, as well as with their intrinsic infinite-dimensionality.

We will show that we can identify the space of infinite-dimensional covariance operators equipped with the Procrustes size-and-shape metric from shape theory, with that of centred Gaussian processes, equipped with the Wasserstein metric of optimal transportation. We then describe key geometrical and topological aspects of the space of covariance operators endowed with the Procrustes metric. Through the notion of multicoupling of Gaussian measures, we establish existence, uniqueness and stability for the Fréchet mean of covariance operators with respect to the Procrustes metric. Furthermore, we provide generative models that are canonical for such metrics.

We then turn to the problem of comparing several samples of stochastic processes with respect to their second-order structure, and we subsequently describe the main modes of variation in this second order structure. These two tasks are carried out via an Analysis of Variance (ANOVA) and a Principal Component Analysis (PCA) of covariance operators respectively. In order to perform ANOVA, we introduce a novel approach based on optimal (multi)transport and identify each covariance with an optimal transport map. These maps are then contrasted with the identity with respect to a norm-induced distance. The resulting test statistic, calibrated by permutation, outperforms the state-of-the-art in the functional case. If the null hypothesis postulating equality of the operators is rejected, thanks to a geometric interpretation of the transport maps we can construct a PCA on the tangent space with the aim of understanding sample variability. Finally, we provide a further example of use of the optimal transport framework, by applying it to the problem of clustering of operators. Two different clustering algorithms are presented, one of which is innovative. The transportation ANOVA, PCA and clustering are validated both on simulated scenarios and real dataset.

Keywords: Fréchet mean, functional ANOVA, functional clustering, functional data analysis, functional PCA, optimal transportation, phase variation, Procrustes distance, Wasserstein distance.

Résumé

Les opérateurs de covariance jouent un rôle fondamental dans l'analyse de données fonctionnelles, car ils fournissent le moyen canonique d'analyser la variation fonctionnelle via la célèbre expansion de Karhunen-Loéve. Ces opérateurs peuvent eux-mêmes être sujets à des variations, par exemple dans des contextes où plusieurs populations fonctionnelles doivent être comparées. Les techniques statistiques permettant d'analyser une telle variation sont intimement liées au choix de la métrique sur l'espace de ces opérateurs, ainsi qu'à leur dimensionnalité infinie.
Nous montrons que nous pouvons identifier l'espace des opérateurs de covariance à dimension infinie équipé avec la métrique procustéenne de "size-and-shape", avec celle des processus gaussiens centrés, équipé de la metrique de Wasserstein. Nous décrivons ensuite les principaux aspects géométriques et topologiques de l'espace de opérateurs de covariance dotés de la métrique procustéenne. À travers la notion du couplage-multiple des mesures gaussiennes, nous établissons l'existence, unicité et stabilité pour le moyenne de Fr échet des opérateurs de covariance par rapport à la distance de Procrustes. De plus, nous fournirons des modèles génératifs canoniques pour une telle métrique.

Nous passons ensuite au problème de la comparaison de plusieurs échantillons de processus stochastiques via leur structure de second ordre puis de la description des principaux modes de variation de cette structure de second ordre. Ces deux tâches sont effectuées via une Analyse de variance (ANOVA) et analyse en composantes principales (PCA) des opérateurs de covariance, respectivement. Afin de réaliser une ANOVA, nous introduisons une nouvelle approche basée sur un (multi) transport optimal suivi d'une identification de chaque covariance avec une carte de transport optimale. Ces cartes sont ensuite contrastées avec l'identité par rapport à une dis- tance induite par une norme. Le test statistique résultant, calibré par permutation, surpasse la l'état-de-l'art dans le cas fonctionnel. Si l'hypothèse nulle postulant l'égalité des opérateurs est rejetée, grâce à une interprétation géométrique du transport optimal, nous pouvons construire une PCA sur l'espace tangent dans le but de comprendre la variabilité de l'échantillon. Enfin, nous fournissons un autre exemple d'utilisation du transport optimal, en l'appliquant au problème de partionnement (clustering) des opérateurs de covariance. Deux algorithmes différents sont présentés, dont l'un est innovant. Ces méthodes d'ANOVA, de PCA et de partionnement via transport optimal sont validés sur des scénarios simulés et sur des données réelles.

Mots clefs : moyenne de Fréchet, Analyse de données fonctionnelles, transport optimal, va-

riation de phase, distance de Wasserstein, distance de Procrustes, ANOVA functionelle, PCA functionelle, clustering functionell

Contents

Acknowledgements v

Abstract vii

Abstract (English/Français) vii

Résumé ix

List of figures xii

List of tables xv

Introduction 1

1 Overview 5

1.1 Functional data analysis 5
  1.1.1 Operators on Hilbert Spaces 5
  1.1.2 Random elements in Hilbert spaces and their covariance operators . . . 8
  1.1.3 Estimation of functional data 10
  1.1.4 The problem of registration 10
1.2 Geometry of covariances 12
  1.2.1 Covariance matrices as a non linear space 13
  1.2.2 Non-linear distances for covariance matrices 14
  1.2.3 Statistical shape analysis and Procrustes size-and-shape distance 15
  1.2.4 Procrustes size-and-shape distance for covariance matrices 18
  1.2.5 Procrustes distance for operators 20
1.3 Optimal Transport and Wasserstein Spaces 22
  1.3.1 Wasserstein distance 25
1.4 Procrustes–Wasserstein distance and optimal transportation of Gaussian processes 25
  1.4.1 Optimal transportation of Gaussian processes 25
  1.4.2 The tangent bundle and geodesics in Wasserstein space 27
  1.4.3 The tangent bundle 27
1.5 Fréchet means in general metric spaces 29
  1.5.1 Existence, uniqueness and the Gaussian case 30

2 Optimal transport of Gaussian processes **31**
2.1 Equivalence of measures 31
2.1.1 Topological properties 33
2.1.2 Finite-rank approximations 34
2.1.3 Geometry of the Gauss–Wasserstein space 35
2.2 Fréchet means of covariance operators 36
2.2.1 Fréchet mean of Gaussian measures 39
2.2.2 Stability of the Fréchet mean 41
2.3 Computing the Fréchet mean through gradient descent 44
2.4 Phase variation and generative models for deformations 45

3 Applications **51**
3.1 Transportation-based functional ANOVA of Covariances 51
3.1.1 Simulation scenarios 55
3.1.2 Numerical simulations 61
3.1.3 Comparison on generative model 67
3.1.4 Data Analysis 72
3.2 Transportation-based functional PCA of covariances 75
3.2.1 PCA under the tangent space inner product 77
3.2.2 Tangent space PCA and multicoupling 79
3.2.3 Data Analysis 80
3.2.4 Numerical simulation 91
3.3 Clustering of covariance operators 92
3.3.1 Tangent space K-means clustering. 95
3.3.2 Soft clustering 99

4 Conclusions and outlook **107**

Bibliography **117**

Curriculum Vitae **119**

List of Figures

1.1 Growth velocity curves of four girls from the Berkeley growth study dataset . . . 11

3.1 Original and smoothed log-periodograms for each phoneme 58

3.2 Mean Log-periodograms of the phonemes. Colours are as follows: black for “sh”, red for “iy”, green for “dcl”, blue for “aa”, cyan for “ao”. 59

3.3 Estimated covariances of the log-periodograms in the five phoneme groups. . . 60

3.4 Comparison of our method against the Pairwise test of Cabassi et al. [2017] and the Concentration-based test of Kashlak et al. [2016] with Gaussian data and geodesic perturbations. Dotted horizontal line gives the nominal level under H_0. 62

3.5 Comparison of our method against the Pairwise test of Cabassi et al. [2017] and the Concentration-based test of Kashlak et al. [2016] with Gaussian data and additive perturbations. Dotted horizontal line gives the nominal level under H_0. 63

3.6 Comparison of our method against the Pairwise test of Cabassi et al. [2017] and the Concentration-based test of Kashlak et al. [2016] with t-Student data and geodesic perturbations. Dotted horizontal line gives the nominal level under H_0. 64

3.7 Comparison of our method against the Pairwise test of Cabassi et al. [2017] and the Concentration-based test of Kashlak et al. [2016] with t-Student data and additive perturbations. Dotted horizontal line gives the nominal level under H_0. 65

3.8 Empirical power comparison, in Berkeley growth data scenario, for the transport functional ANOVA test, in the case where the maps are contrasted using the tangent space inner product rather than the Hilbert–Schmidt one. The black line corresponds to the transport-based ANOVA, the red line is the pairwise test from Cabassi et al. [2017] and blue line is the transport based ANOVA with respect to the tangent space metric. Power is estimated via permutations. 66

3.9 Comparison of empirical powers computed with respect to the the three different norms (Nuclear, Hilbert–Schmidt and operator) with geodesic perturbations using the Berkeley growth data. Dotted horizontal line gives the nominal level under H_0. 68

3.10 Comparison of empirical powers computed with respect to the the three different norms (Nuclear, Hilbert–Schmidt and operator) with additive perturbations using the Berkeley growth data. Dotted horizontal line gives the nominal level under H_0. 69

3.11 Empirical power of our and Cabassi et al. [2017]'s tests as a function of the dispersion parameter σ of the von Mises distribution in the generative model of Equation (3.8). Our test is performed using three different norms: trace-class, Hilbert–Schmidt and operator norm. 71

3.12 Empirical power of our and Cabassi et al. [2017]'s tests as a function of the dispersion parameter σ of the Von Mises distribution in the generative model of Equation (3.8) (3.8). Our test is performed under three different norms: trace-class, Hilbert–Schmidt and operator norm. The origin for the generative model is taken to be a Wishart distribution with 70 degrees of freedom. 73

3.13 PCA scores, as computed from the phoneme dataset. The colours are as follows: "sh" black, "iy" red, "dcl "green", "aa" blue, "ao" cyan 82

3.14 Screeplot of eigenvalues, phoneme dataset 82

3.15 Variation along the first principal geodesic. Contour plot of the difference $\bar{\Sigma}_{aa,ao,iy} - \bar{\Sigma}_{dcl,sh}$ between the barycenters of the covariances of "aa,ao and iy" and "dcl and sh" on the left-hand side, and the arithmetic difference between two operators computed by moving in opposite directions along $\gamma_t^{(1)}$ on the right. 83

3.16 Variation along the second principal geodesic. Contour plot of the difference $\Sigma_{dcl} - \Sigma_{sh}$ between the covariances of "dcl" and "sh" on the left-hand side, and the arithmetic difference between two operators computed by moving in opposite directions along $\gamma_t^{(2)}$ on the right. 84

3.17 Variation along the third principal geodesic. Contour plot of the difference $\Sigma_{iy} - \bar{\Sigma}_{aa,ap}$ between the covariances of "iy" and the barycenter of "aa and ao" on the left, and the arithmetic difference between two operators computed by moving in opposite directions along $\gamma_t^{(3)}$ on the right. 84

3.18 Variation along the fifth principal geodesic. Variation along the third principal geodesic. Contour plot of the difference $\Sigma_{aa} - \Sigma_{ao}$ between the covariances of "aa" and "ao" on the left, and the arithmetic difference between two operators computed by moving in opposite directions along $\gamma_t^{(5)}$ on the right. 85

3.19 PCA scores of the expanded phoneme dataset. The colours are as follows: "sh" black, "iy" red, "dcl "green", "aa" blue, "ao" cyan 86

3.20 PCA scores of the expanded phoneme dataset in a 3D scatterplot. The colours are as follows: "sh" black, "iy" red, "dcl "green", "aa" blue, "ao" cyan 87

3.21 Screeplot of eigenvalues, expanded phoneme dataset 87

3.22 PCA scores of the 60 covariance operators based on the Hilbert–Schmidt distance. 88

3.23 Variation along the first principal geodesic. Contour plot of the difference $\bar{\Sigma}_{aa,ao,iy} - \bar{\Sigma}_{dcl,sh}$ between the barycenters of the covariances of "aa,ao and iy" and "dcl and sh" on the left-hand side, and the arithmetic difference between two operators computed by moving in opposite directions along $\gamma_t^{(1)}$ on the right. 89

3.24 Variation along the second principal geodesic. Contour plot of the difference $\Sigma_{dcl} - \Sigma_{sh}$ between the covariances of "dcl" and "sh" on the left-hand side, and the arithmetic difference between two operators computed by moving in opposite directions along $\gamma_t^{(2)}$ on the right. 90

3.25 Variation along the third principal geodesic. Contour plot of the difference $\bar{\Sigma}_{aa,ao} - \Sigma_{iy}$ between the barycenters of the covariances of “aa and ao” and the covariance of “iy” on the left-hand side, and the arithmetic difference between two operators computed by moving in opposite directions along $\gamma_t^{(3)}$ on the right. 90
3.26 Variation along the fifth principal geodesic. Contour plot of the difference $\Sigma_{aa} - \Sigma_{ao}$ between the covariances of “aa” and “ao” on the left-hand side, difference between two operators computed by moving in opposite directions along $\gamma_t^{(5)}$ on the right. 91
3.27 PCA scores, first experiment with the generative model. Colours correspond to the three maps generated from the model. 93
3.28 Eigenvalues screeplot, first experiment with the generative model 93
3.29 PCA scores, second experiment with the generative model. The colours are as follows: black corresponds to $\sigma = 0.1$, green to $\sigma = 1$ and red to $\sigma = 5$ 94
3.30 Screeplot of eigenvalues, second experiment with the generative model. 94
3.31 Misclassification as a function of intra-cluster distance. 98
3.32 AIC_c criterion according to the number of clusters 98
3.33 Classifications into clusters based on a subsample of 25. The picture show the medians of the column of the partition matrix P for the rows corresponding to the various phonemes. 103
3.34 Classifications into clusters based on a subsample of 50. The picture show the medians of the column of the partition matrix P for the rows corresponding to the various phonemes. 104
3.35 Tangent space PCA of estimated prototypes (experiment n. 3 with sample size equal to 25). 105

List of Tables

3.1 Comparison of the three different norms, geodesic perturbations, Berkeley growth data. 67

3.2 Comparison of the three different norms, additive perturbations, Berkeley growth data. 70

3.3 Comparison of the empirical power of the three different testing methods on the `phoneme` dataset, when applied on the phonemes "aa", "ao" and "iy". 74

3.4 Comparison of the empirical power of the three different testing methods on the `phoneme` dataset, when applied on the phonemes 74

3.5 Importance of each PC for the phoneme dataset 81

3.6 Importance of each PCs in the expanded phoneme dataset 85

3.7 Importance of each PC, first experiment with the generative model. 92

3.8 Importance of each PC, second experiment with the generative model. 92

3.9 Mean of the percentage of wrong classification for different values of the von Mises dispersion σ. 97

3.10 Percentage of cases with perfect classification for different values of the von Mises dispersion σ. 97

3.11 Distribution of the true phonemes in the cluster identified by the tangent space K-means algorithm. 97

3.12 Uncertainty matrix P^*P (experiment n. 3 with sample size equal to 25). Observe, from Figure 3.33, that the classes identified by the algorithm correspond to: 1 ↔ “dcl”; 2 ↔ “sh”; 3 ↔ “aa” (or “ao”); 4 ↔ “ao” (or “aa”); 5 ↔ “iy”. 102

Introduction

Thanks to advances in science and technology, more and more datasets over the last decades are being sampled with increasingly high precision and are recorded in increasingly complex forms. Examples of such complexity comprise data that are sampled so finely as to be assumed to be smooth curves, or surfaces. These data are called functional and from a mathematical perspective they are taken to be random elements of an infinite-dimensional space. Examples of datasets include growth and temperature curves, electricity consumption curves, density functions, speech recordings, satellite images, brain images, or DNA mini-circles vibrating in solution [Ramsay and Silverman, 2005a, Pigoli et al., 2014b, Tavakoli and Panaretos, 2016].

Functional data analysis (FDA) is a branch of statistics dealing with the analysis of these complex (functional) data objects. Most recently, research has been carried out on inferential procedures concerning not only the functional curves but also their covariance operators [Panaretos et al., 2010a, Fremdt et al., 2013, Pigoli et al., 2014a]. Covariances are key elements in FDA. Their spectrum provides a singular system that allows to separate stochastic and functional fluctuations of random elements, via the renowned Karhunen–Loève expansion. Such singular systems also allow to write optimal finite-dimensional approximations of functional data. Through these, we obtain means to carry out inferential procedures like functional Principal Component Analysis (PCA), as well as regression and testing, both of which would be otherwise ill-posed in infinite dimensions [Panaretos et al., 2010a, Tavakoli and Panaretos, 2016].

One may conceive of statistical applications where covariance operators may exhibit variation of their own, and thus be taken to be the main object of statistical inference. Situations displaying this kind of variability comprise cases when data curves are supposed to stem from different functional populations (Pigoli et al. [2014a], Tavakoli and Panaretos [2016], see also Chapter 3 of this thesis).
Research in this direction aims to quantify and understand their level of variability. *First-order variation* is a classical problem in FDA, and concerns the variability across populations of their mean structure. A further type of variation is manifested when the populations differ in their smoothness and fluctuation properties. We call this *second-order variation*, and it arises when the covariance operators of the various populations are distinct. Early studies concerning this second-order variation were motivated through financial and biophysical applications

[Benko et al., 2009, Panaretos et al., 2010a] and were followed by many more contributions with a diverse span of applications (e.g. Horváth et al. [2013], Gabrys et al. [2010], Fremdt et al. [2013], Kraus [2014]).
Most of these works, though different in the techniques they propose, share a common assumption: they imbed covariance operators in Hilbert–Schmidt space, and carry out the statistical inference with respect to the corresponding metric. The issue with employing the Hilbert–Schmidt metric is that it implicitly assumes a linear structure, while covariances can be seen as "squares" of Hilbert–Schmidt operators, and as such they are not closed under linear operations. It is therefore desirable to employ statistical methods respecting this non-linear geometry of the space.

In finite dimensions and in the statistical analysis of covariance matrices, the curved nature of the space is well-documented, especially due to its connection with shape theory and the problem of diffusion tensor imaging [Dryden et al., 2009, Schwartzman, 2006].
In infinite dimensions, the first steps in the direction of a non-linear analysis of covariances were taken by Pigoli et al. [2014a]. With the goal of analysing phonetic variations across Romance languages, they described a 2-sample testing procedure which is respectful of the intrinsic geometry of covariances. To this purpose, they defined a functional extension of the so-called Procrustes size-and-shape metric (abbreviated to just Procrustes metric hereafter) and derived some of its properties: they showed that it is well-behaved with respect to finite-dimensional projections and that is computationally valid in applications. They went on to discuss Fréchet means – that is, the extension of linear averages to general metric spaces – with respect to the Procrustes metric, arguing that they can be successfully computed via a version of the Generalised Procrustes Algorithm [Gower, 1975]. In summary, their contribution inaugurated the non-Euclidean statistical analysis of covariance operators. While doing so, it produced many further questions about the Procrustes metric. For example, one might wonder whether it is possible to deduce theoretical properties of the Fréchet mean, such as existence and uniqueness, or whether the inference can be expanded into more general procedures beyond 2-sample testing, such as functional Principal Component Analysis (fPCA). Perhaps an even more relevant issue concerns the geometrical and statistical interpretation of the Procrustes metric: the Procrustes size-and-shape distance for covariance matrices stems from shape theory, and in view of that it comes with a well-rooted geometrical interpretation. Can we establish an equivalent connection in the infinite-dimensional case, and gain a similar understanding of the geometry of covariances under the Procrustes metric? This thesis sets out to address these questions. In particular we will show that some of these problems can be read through the lens of optimal transport and can be answered thanks to the intimate connection between the Procrustes distance and the Wasserstein metric between Gaussian processes.
The detailed structure of the thesis and a summary of its main contributions are given in the next paragraph. The innovative results are mostly collected in Chapter 2 and Chapter 3 (with the exception of Section 2.1.2 in Chapter 1). Chapter 2 is largely based on Masarotto et al. [2018], while Chapter 3 is based on Masarotto et al. [2019], and it extends the application

results found there.

Detailed structure of the thesis.

Chapter 1. Following the review of widely-known notions of operators on Hilbert spaces, we proceed towards reviewing basic concepts of (inference for) functional data (Section 1.1). We pay special attention to the covariance structure of these data, which is highlighted as a statistically interesting quantity in itself. Paragraph 1.1.4 addresses the problem of registration of curves, which will be revisited in Chapters 2 and 3 due to its connection with tangent space PCA.

Subsequently we describe the geometry of the space of covariances, both finite- and infinite- dimensional (Section 1.2). This will require some concepts of statistical shape theory (Section 1.2.3), of Wasserstein spaces and Optimal Transport (Section 1.3 with a special focus on Gaussian processes 1.4). We conclude by treating the problem of computing means in general metric spaces in Section 1.5.

Chapter 2. This chapter collects most of the theoretical contributions of the thesis. As mentioned in the introduction, Pigoli et al. [2014a] initiated the study of second-order variation across populations of functional data in a non-linear manner. Some of the research questions (implicitly or explicitly) generated by their work were addressed in Masarotto et al. [2018]. The chapter begins by making the connection between the Procrustes and Wasserstein metrics explicit (Section 2.1). This will allow us to benefit from the rich theory of Optimal Transport, which in turn will lead us to establish new results related to existence, uniqueness and stability of the Fréchet mean of Gaussian measures on a Hilbert space (Section 2.2). Section 2.3 gives details on the convergence and the practical implementations of a gradient descent algorithm to compute the Fréchet mean, while in Section 2.4 we present a generative statistical model compatible with the Procrustes metric and linking it with the problem of registration (Paragraph 1.1.4).

Chapter 3. This chapter contains methods and results regarding the applied analysis on a dataset of populations of covariances.

We begin by considering the problem of testing the hypothesis of equality of covariances across the different populations. We view the testing problem through the lens of the optimal multicoupling of Gaussian processes. Specifically, we adopt a novel transportation perspective to introduce a new ANOVA test, translating the task of testing the hypothesis of equality of covariance operators into that of testing whether the optimal multicoupling between the corresponding centered Gaussian measures is “trivial”.

The 2-sample testing procedure of Pigoli et al. [2014a] has been generalised into K-sample testing by Cabassi et al. [2017]. They present simulations illustrating state-of-the-art performance of their method. We will show that the (multi)transport perspective allows us to construct a 2- and K-sample test that is more powerful than other approaches when applied to functional data. Hypothesis testing is reported in Section 3.1.

If the null hypothesis is rejected, Principal Component Analysis (PCA) offers a useful tool to understand the differences within the data and describe the main mode(s) of variation. The understanding of the Wasserstein geometry allows us to perform PCA on the tangent space. To the best of our knowledge, this is the first instance of a functional PCA on covariance operators that respects their intrinsic geometric features as trace-class positive operators. We describe tangent space PCA in Section 3.2.

Clustering of operators is treated in Section 3.3. Two different clustering methods for covariances are presented, one of which, coined *soft* clustering, is innovative.

Each Section in this Chapter contains a description of the methodology and data analysis. Simulations are performed in a variety of scenarios. For convenience, we collected them in Section 3.1.1.

1 Overview

Functional data analysis is the field of statistics that treats the cases where the single data atoms are continuous curves. In this chapter we review the basic literature regarding operators in Hilbert spaces and (inference for) functional data (Section 1.1) with a special focus on their covariance structure, which is highlighted as a statistically interesting quantity in itself. We will then move to the geometrical structure of the space of covariances, both finite- and infinite-dimensional (Section 1.2). This will require some concepts of Statistical Shape Theory (Section 1.2.3) and of Wasserstein spaces and Optimal Transport (Sections 1.3 and 1.4). We conclude by treating the problem of computing means in general metric spaces in Section 1.5.

1.1 Functional data analysis

A data set is called functional when its individual elements are of infinite dimension. Hence these are distinct from high-dimensional data, whose dimension is larger than the sample size, yet still finite. Functional data are thus taken to be random elements of an infinite-dimensional space. Such space can be (and will often be, as we will see) non-linear. Another identifying characteristic that separates functional from high-dimensional data, is that they are assumed to vary smoothly in their domain. It makes sense therefore to consider notions such as derivatives and continuous transformations of the domain (such as deformations) which otherwise make no sense in the high-dimensional setting. The most common mathematical setting for functional data consists of having a collection of independent realisations of a random element X taking values on a separable Hilbert space $\mathscr{H}$, most usually assumed to be $L^2[0,1]$ or some reproducing kernel Hilbert subspace thereof. Before moving forward in the characterisation of X, we recall some basic notions of operators on Hilbert spaces.

1.1.1 Operators on Hilbert Spaces

We follow Hsing and Eubank [2015] to recollect some basic facts about operators on Hilbert spaces.

Let $\mathscr{H}$ be a real separable Hilbert space with inner product $\langle\cdot,\cdot\rangle_{\mathscr{H}}:\mathscr{H}\times\mathscr{H}\to\mathbb{R}$, and induced norm $\|\cdot\|_{\mathscr{H}}:\mathscr{H}\to[0,\infty)$. A linear map $A:\mathscr{H}\to\mathscr{H}$ is said to be bounded, or equivalently continuous, if its operator norm is finite [1], i.e.

$$\|A\|_\infty := \sup\{\|Ax\|_{\mathscr{H}} \ : \ \|x\|_{\mathscr{H}}\le 1\} < \infty.$$

We might denote the operator norm simply by $\|\cdot\|$ if this does not give rise to misunderstanding.

The space of bounded operators on $\mathscr{H}$ equipped with the operator norm forms a Banach space. A bounded operator A is called

- *self-adjoint* if $A=A^*$, where A^* is the unique operator such that $\langle Af,g\rangle_{\mathscr{H}}=\langle f,A^*g\rangle_{\mathscr{H}}$, for all $f,g\in\mathscr{H}$;
- *compact* if for any bounded sequence $\{f_n\}_{n\in\mathbb{N}}\in\mathscr{H}$ there exists a convergent subsequence of $\{Af_n\}_{n\in\mathbb{N}}$.

A *non-negative* operator is a self-adjoint, possibly unbounded operator A, such that $\langle Au,u\rangle_{\mathscr{H}}\ge 0$ for all u in the domain of A. If in addition A is compact, then there exists a unique *square root*, that is a non-negative operator whose square equals A. This will be denoted by either $A^{1/2}$ or $\sqrt{A}$. For any bounded operator A, A^*A is non-negative.

Associated with a linear map $A:\mathscr{H}\to\mathscr{H}$ there are the spaces:

- $\mathrm{Dom}(A)$ = the subset of $\mathscr{H}$ on which A is defined;
- $\mathrm{Im}(A)=\{Ax \ : \ x\in\mathrm{Dom}(A)\}$;
- $\ker(A)=\{x\in\mathrm{Dom}(A) \ : \ Ax=0\}$;

called the *domain*, *image* and *kernel* of A respectively. We will also use the notation $\mathrm{range}(A)$ for $\mathrm{Im}(A)$ and call it alternatively the *range* of A. Unless otherwise stated, we will take $\mathrm{Dom}(A)$ to be the entire $\mathscr{H}$ space.

The *rank* of A is defined to be

$$\mathrm{rank}(A)=\dim(\mathrm{Im}(A))$$

and it can be infinite. The closure of a set E will be denoted as $\overline{E}$.

For a pair $f,g\in\mathscr{H}$, the tensor product $f\otimes g:\mathscr{H}\to\mathscr{H}$ is the linear operator defined by

$$(f\otimes g)u=\langle g,u\rangle_{\mathscr{H}}f, \qquad u\in\mathscr{H}.$$

---

[1] In general the domains of definitions of A can be different, $A:\mathscr{H}\to\mathscr{Y}$, but we restrict ourselves to a single space here for clarity and relevance to the purposes of this thesis.

In the following, to lighten the notation we will drop the subscript in $\langle\cdot,\cdot\rangle_{\mathscr{H}}$ unless there is a risk of misinterpretation.

Similarly to what happens in finite dimensions, a compact, non-negative definite operator is "diagonalizable", in the sense of the next proposition, known as the Spectral Theorem:

Proposition 1. *Let A be a compact, self-adjoint and non-negative operator. Then it admits a spectral decomposition given by*

$$A = \sum_{i=1}^{\infty} \lambda_i (\varphi_i \otimes \varphi_i) = \sum_{i=1}^{\infty} \lambda_i \langle \varphi_i, \cdot \rangle \varphi_i,$$

with $\{\lambda_i\}_{i=1}^{\infty}$ and $\{\varphi_i\}_{i=1}^{\infty}$ called the eigenvalues and the (orthonormal) eigenfunctions of A respectively. Moreover, the set of non-zero eigenvalues is either finite or constitutes a sequence decreasing to zero, and the associated set of eigenfunctions forms a basis for $\overline{\mathrm{Im}(A)}$.

Note that λ_i and φ_i behave as in finite dimensions, in the sense that $A\varphi_i = \lambda_i \varphi_i$, for any i.

A linear mapping A is *one-to-one* if $\ker(A) = \{0\}$, and *surjective* if $\mathrm{range}(A) = \mathscr{H}$. When A is both one-to-one and surjective it is said to be bijective. Bijective linear mappings are invertible, in the sense that there exists a linear map $A^{-1} : \mathscr{H} \to \mathscr{H}$ such that $A^{-1}A = \mathfrak{I} = A^{-1}A$, $\mathfrak{I}$ denoting the identity operator in $\mathscr{H}$. In general A^{-1} is only defined on a subspace (often dense) of $\mathscr{H}$.

A bounded and compact operator is *Hilbert–Schmidt* (HS) if its Hilbert–Schmidt norm $\|\cdot\|_2$ is finite, where

$$\|A\|_2 = \sqrt{\mathrm{tr}\,(\mathrm{A}^*\mathrm{A})}.$$

The space of Hilbert–Schmidt operators equipped with the inner product

$$\langle A_1, A_2 \rangle_{HS} = \sum_{i=1}^{\infty} \langle A_1 e_i, A_2 e_i \rangle_{\mathscr{H}}$$

is itself a separable Hilbert space. Here $\{e_i\}_{i=1}^{\infty}$ can be taken to be any orthonormal basis of $\mathscr{H}$. A Hilbert–Schmidt operator is *trace-class* or *nuclear* if it has finite trace (nuclear) norm, i.e. if

$$\|A\|_1 = \mathrm{tr}\left(\sqrt{\mathrm{A}^*\mathrm{A}}\right)$$

is finite.

It is well-known that

$$\|A\|_\infty \leq \|A\|_2 \leq \|A\|_1$$

for any bounded linear operator A.

1.1.2 Random elements in Hilbert spaces and their covariance operators

For a random element X in a separable Hilbert space $\mathscr{H}$, we define its *mean* or expectation as the unique element $\mathbb{E}X$ such that for all continuous linear functionals f in $\mathscr{H}$, $f(\mathbb{E}X) = \mathbb{E}f(X)$, if $\mathbb{E}\|X\| < \infty$.

The *covariance operator* of X is defined as

$$R = \mathbb{E}\left[(X-\mu) \otimes [(X-\mu)\right]$$

if $\mathbb{E}\|X\|^2 < \infty$.
The covariance operator is a trace-class non-negative operator, and since the trace is continuous and linear we have the following relation between the total variance and the nuclear norm of X,

$$\mathrm{trR} = \mathbb{E}\left[\mathrm{tr}(\mathrm{X}-\mu) \otimes (\mathrm{X}-\mu)\right] = \mathbb{E}\|\mathrm{X}-\mu\|^2.$$

A typical setup to perform inference on functional data is based on an independent sample $X_1, \ldots, X_n \sim X$, where X is a random function on the Hilbert space $\mathscr{H}$ such that $\mathbb{E}\|X\|^2 < \infty$. To clarify the concept of random variable in a general Hilbert space, $\mathscr{H}$ can be thought to be the space of real-valued measurable and square integrable functions $L^2([0,1],\mathbb{R})$ and X can be thought of as a collection of random variables $\{X(t) : t \in [0,1]\}$.

We point out that for the evaluation $X(t)$ on a specific point $t \in [0,1]$ to be well-defined, one of the following assumptions is required: either a Reproducing Kernel Hilbert Space (RKHS) structure on $\mathscr{H}$ (see Berlinet and Thomas-Agnan [2011]) or, following Hsing and Eubank [2015], the interpretation of X as a (mean-square continuous) stochastic process. In the rest of the thesis we adopt this second point of view. The notation when X takes value in $L^2[(0,1)]$ is fixed below.

Let $X : \Omega \to L^2[(0,1),\mathbb{R}]$ be a continuous stochastic process on the probability space $(\Omega, \mathscr{B}, \mathbb{P})$, and assume that $\mathbb{E}\|X\|^2 < \infty$. The mean function of $\mu \in L^2[(0,1),\mathbb{R}]$ of X is defined by

$$\mu(t) = \mathbb{E}[X(t)], \quad t \in [0,1],$$

while its covariance kernel is $r \in L^2[(0,1),\mathbb{R}]$, defined by

$$r(t,s) = \mathbb{E}\left[(X(t)-\mu(t))(X(s)-\mu(s))\right], \quad t,s \in [0,1].$$

The operator

$$Rh(t) = \int_0^1 r(t,s)h(s)ds, \quad h \in L^2[(0,1),\mathbb{R}]; t \in [0,1],$$

is a well-defined, non-negative, self-adjoint and trace class operator in $L^2[(0,1),\mathbb{R}]$ called the covariance operator of X. If X also respects

$$\lim_{n\to\infty} \mathbb{E}\left[(X(t_n) - X(t))^2\right] = 0, \quad t \in [0,1],$$

for any sequence $\{t_n\}_{n\in\mathbb{N}}$ converging to t in $[0,1]$, then X is said to be mean-square continuous. A mean-square continuous stochastic process with continuous sample paths $t \to X_\omega(t)$, for all $\omega \in \Omega$, is a random element of $L^2[(0,1),\mathbb{R}]$ ([Hsing and Eubank, 2015, Theorem 7.4.1]), and we can therefore establish the desired equivalence between random objects in Hilbert spaces and stochastic processes. Moreover X is mean square continuous if and only if its mean and covariance functions are continuous [Hsing and Eubank, 2015, Theorem 7.3.2]. As a consequence, if the mean is continuous, then the covariance function is continuous at all (s,t) if and only if it is continuous at all "diagonal points" (t,t).

We now turn our attention to covariance operators. Covariance operators are key elements in functional data analysis because they are a canonical means to study the variation of random functions. We already know that they admit the spectral decomposition as in Proposition 1. Furthermore, their spectrum provides a way to separate the stochastic and the functional fluctuations of a random function, through the Karhunen–Loève expansion (see, e.g., Karhunen [1947]).

Theorem 2 (Karhunen–Loève expansion)**.** *Let X be a mean square continuous process on $L^2[(0,1),\mathbb{R}]$ with $\mathbb{E}\|X\|^2 < \infty$ and with covariance operator R. Let $R = \sum_{n=1}^{\infty} \lambda_n \varphi_n \otimes \varphi_n$ be the singular value decomposition of R, where λ_n denotes the n-th largest eigenvalue (i.e. $\lambda_1 \geq \lambda_2 \geq \ldots 0$) and φ_n its corresponding eigenfunction. The random function X admits the decomposition*

$$X = \mu + \sum_{n=1}^{\infty} \xi_n \varphi_n, \tag{1.1}$$

where $\xi_n = \langle \varphi_n, X - \mu \rangle$, $\mathbb{E}\xi_n = 0$ and $\mathbb{E}[\xi_n \xi_m] = \lambda_n \delta_{n,m}$ with $\delta_{n,m} = 1$ if $n = m$ and 0 otherwise. It holds that

$$\sup_{t\in[0,1]} \mathbb{E}\left[(X - \mu - \sum_{n=1}^{K} \xi_n \varphi_n)^2\right] \xrightarrow[K\to\infty]{} 0.$$

The Karhunen–Loève expansion allows the separation of X into a sum of random variables $(\xi_n)_{n=1}^{\infty}$ times orthogonal deterministic functions $(\varphi_n)_{n=1}^{\infty}$. The random variables $(\xi_n)_{n=1}^{\infty}$ in (1.1) are uncorrelated and, if X is Gaussian, independent.

Furthermore, it holds the *best basis property*, that is, the truncation of the series (1.1) at any finite level K yields the best $K-$dimensional linear approximation of X. For this reason the Karhunen–Loève expansion plays a crucial role in FDA, in particular in dealing with functional Principal Component Analysis (fPCA, see e.g. Grenander [1950b], Dauxois et al. [1982]) or as natural means of regularisation for inference problems such as regression and testing, which are ill-posed in infinite dimensions (Panaretos and Tavakoli [2013], Wang et al. [2016]).

1.1.3 Estimation of functional data

In practice, when performing inference on functional data, one has to deal not with the fully observed curves, but with discrete measurements, arising for example when the full curves are observed on a discrete grid of points. Often, such measurements are corrupted by noise, and are subjected to an extra level of smoothing.

To fix the notation, we assume that we have observed a sample of discrete realisations

$$Y_k(j) = X_k(t_{ij}) + \varepsilon_{ij},$$

of a smooth continuous process X_k, with $t_{i1}, \ldots, t_{in_i} \in [0,1]$ and ε_{ij} being mean zero and covarying according to some Σ.

The two main approaches to deal with estimation of the X_k's are:

- (first smooth then estimate) popularised by Ramsay and Silverman [2005b]. This approach consists of initially reconstructing the functional smooth version of the data $\tilde{X}_k$ and subsequently constructing a smoothed version of the covariance operator $\tilde{R}$. From $\tilde{R}$ it is then possible to approximate the functions X_k by their Karhunen–Loève expansion.
  Common methods to perform the smoothing step are kernel smoothing, localised basis or polynomial expansions (see e.g. Wand and Jones [1995], Fan et al. [1996], Efromovich [2008]) or again via a least squares estimation of the functions X_i through a finite basis expansion. This latter method is the most popular, and the most commonly chosen basis are the Fourier basis, the B-spline basis and more recently a wavelet basis (Yao and Lee [2006], Pigoli and Sangalli [2012] and references therein).
  A high number of basis functions will possibly capture local features of the curves X_k, and to prevent this, an additional penalty is often imposed on the roughness of the eigenfunctions ([Ramsay and Silverman, 2005b, Silverman, 1996]).

- (first estimate then smooth) estimate multivariate versions of the mean and covariance operators of X_k's and successively smooth these quantities into functions. We refer to the work of Staniswalis and Lee [1998], Yao et al. [2003] among others and to Descary [2017] and especially the references in Section 1.2.2.

1.1.4 The problem of registration

In addition to dealing with discrete measurements, often we have to deal with perturbed data as well. Intuitively speaking, such a perturbation implies that the “peaks” and the “valleys” of given functions are not aligned properly. Figures 1.1 shows a subset of four growth velocity curves from the Berkeley growth study data set [Jones and Bayley, 1941]. It is visible from the picture that the curves show similar behaviour (a growth spurt between the ages of 10 and 15) but with a difference in the magnitude of the peaks as well as a misalignment with respect to the x-axes. These horizontal perturbations might come from an uncertainty in the data sampling process or they can represent inherent variability of the process itself that needs to be separated from the variability along the y-axes.

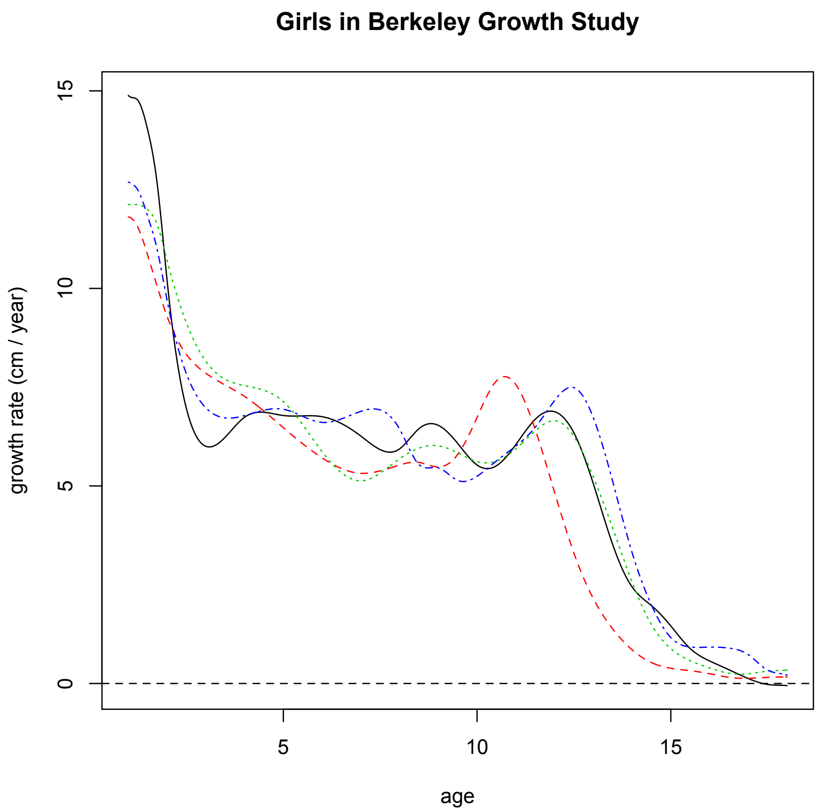


Figure 1.1: Growth velocity curves of four girls from the Berkeley growth study dataset

Formally, we talk of *amplitude* and *phase* variation while referring to the variation along the y- and x-axes respectively. Amplitude variation captures stochastic deviations of the random vector $X_1, \ldots, X_N$ from its mean, while phase variation implies that rather than a random sample $X_1, \ldots, X_N$ from X, we observe $\widetilde{X}_1, \ldots, \widetilde{X}_N$ from a perturbed random element $\widetilde{X} = X(T(\cdot))$, where T is a random invertible function taking values in an often non-linear space and often called a *warping* function. Mathematically they can be more precisely described as follows:

- Amplitude variation: realisations of a process X with Karhunen–Loève expansion

$$X = \sum_{n=1}^{\infty} \lambda_n^{1/2} \xi_n \varphi_n$$

fluctuating around fixed (deterministic) modes φ_n. Here $\{\lambda_n, \varphi_n\}$ are eigenvalues and eigenfunctions of the covariance operator R, and $\{\xi_n\}_n$ is a sequence of random variables as in Theorem 2.

- Phase variation: the realisations from X are warped into realisations from $\widetilde{X}$ via bounded non-negative operator T (usually uncorrelated with X),

$$\widetilde{X} = TX = \sum_{n=1}^{\infty} \lambda_n^{1/2} \xi_n T\varphi_n$$

  with T such that $\widetilde{X}$ has finite variance.

Now if each of the observed $\widetilde{X}_1, \ldots, \widetilde{X}_N$ is warped through the functions $T_1, \ldots, T_N$, the goal of registration is to separate the functions $X_1, \ldots, X_N$ from the warping functions $T_1, \ldots, T_N$. Whether or not phase variation is a problem for the statistical inference depends on the specific application. In general, the Karhunen–Loève expansion of $\widetilde{X}$ is very different from that of X and an unaccounted-for warping can introduce errors and artefacts into the statistical analysis. There is no canonical way to solve the curve registration problem. Kneip and Gasser [1992], Liu and Müller [2004], Tang and Müller [2008], Panaretos and Zemel [2016] and others offer a variety of different registration methods. We will see later when talking of optimal transport that the Wasserstein distance offers a natural way to deal with the non-linearity of the phase variation (Section 1.3 and Section 2.4) and that principal component analysis can offer a way to recover the warping map T (Section 3.2).

1.2 Geometry of covariances

Performing inference on random objects in a general metric space is deeply connected with the metric structure of the space itself. Even the most basic operations, like normalising or computing linear averages, implicitly assume a way to compare objects in terms of their proximity, and consequently a way to compute distances. Covariances themselves are a way to encapsulate variation, and are computed with respect to an inner product. In classical multivariate analysis, the ambient inner product is taken to be the standard Euclidean one. We can imagine situations, however, where the single data atom is a covariance matrix itself (or, in infinite dimension, a covariance operator), and we wish to perform inference on a collection of covariances. This happens for example in longitudinal data analysis [Daniels and Pourahmadi, 2002] or diffusion tensor imaging [Schwartzman, 2006, Alexander et al., 2007]. Diffusion Tensor Imaging (DTI) is a form of Magnetic Resonance Imaging (MRI) that measures the diffusion of water molecules in tissue. In DTI, the data contained at every 3D pixel is not a scalar, but rather it is identified through a 3×3 positive definite matrix.

Inference for symmetric positive definite matrices has been investigated under a variety of possible metrics. The purpose of this section is to describe the intrinsic non-linearity of the Riemannian manifold structure of the space of symmetric positive definite matrices, and therefore the need to introduce a metric structure beyond Euclidean. We will follow Schwartzman [2006] and Bhattacharya and Patrangenaru [2003, 2005]. We then describe some of the metrics employed when studying variations of symmetric positive definite matrices, devoting special attention to the Procrustes size-and-shape distance. The latter distance, inspired by statistical shape theory, unlike many other metrics generalises easily and naturally to infinite dimensions, and plays a crucial role in the rest of this thesis.

1.2.1 Covariance matrices as a non linear space

Denote as $Sym^+ n$ the set of symmetric positive definite matrices in $\mathbb{R}^{n\times n}$. $Sym^+ n$ forms a differentiable manifold of dimension $n(n+1)/2$. The tangent space to such a manifold at the identity $\mathcal{I}_{n\times n}$ has dimension $n(n+1)/2$ and can be identified with a copy of the space of symmetric matrices [Schwartzman, 2006, Prop. 2.2.1].

A Riemannian manifold is a differential manifold endowed with an inner product on the tangent space that varies smoothly from point to point. The most straightforward way to turn $Sym^+ n$ into a Riemannian manifold would be to consider it as a subset of the Euclidean space of symmetric matrices (not constrained to be positive) endowed with the Frobenius (Hilbert–Schmidt) inner product. This generates a “flat” manifold, where the *geodesics*, that is, the shortest paths between two points of the manifold, are straight lines.

For illustration, consider the set of 2×2 symmetric positive definite matrices of the form

$$X = \begin{pmatrix} a & c \\ c & b \end{pmatrix}.$$

The set of triples $\{(a,b,c) \in \mathbb{R}^3 \ : a>0, b>0, ab-c^2>0\}$ giving rise to $Sym^+(2)$ is an open subset of $\mathbb{R}^3$ shaped as a cone. According to the Euclidean metric the geodesic between two matrices X_1 and X_2 would result in a straight line cutting through the cone. Therefore extrapolation along a geodesic might return a matrix which is not lying on the cone surface and might not be positive definite. This happens because positivity is a not-linear constraint, and imposing a linear geometry on a non linear space might distort the statistical analysis. In Section 3.2.3 we provide an example of what can happen when the non-linearity of the space is ignored in favour of the Euclidean distance.

A final important point about Sym^+ is that its elements are related to each other by an action of the general linear group $GL(n)$. This action is manifested through the transformation $\phi : GL(n) \times Sym^+(n) \to Sym^+(n)$, $\phi_G(X) = GXG^*$, where G^T denotes the transpose of G. Moreover, it is transitive (i.e., for every $X, Y \in Sym^+$ there exists a non-unique G that allows one to travel from one to the other) and it translates into a similar group action between the tangent spaces at X and $\phi_G(X)$. Intuitively, the group action allows one to travel through the space of positive definite matrices and can be thought of as a data-generating mechanism: any set of covariances $X_1, \dots, X_N$ can be seen as a perturbation of a generating matrix X via a set of linear invertible transformations $G_1, \dots, G_N$. We will see in Section 2.4 how the Procrustes distance allows us to write a similar generating mechanism for covariance operators.

The rest of this section is devoted to the introduction of different metrics which respects the "curved" nature of the Riemannian manifold of covariances. Here curved means that geodesics starting at a point p may meet for a second time in the cut locus of p, in contraposition to flat manifolds, where geodesics, being straight lines, meet at most once. Geodesics defined according to this curved geometry will in return be fully contained in the manifold itself.

1.2.2 Non-linear distances for covariance matrices

We report the definition of several non-linear metrics which have been used to perform inference on covariance matrices. We mostly follow Dryden et al. [2009].

Given a sample $X_1, \dots, X_N$ of observations and a distance d we can extend the notion of linear average to a more general concept of mean $\overline{X}$ in the following way,

$$\overline{X} = \mathrm{arginf}_X \sum_{i=1}^{N} d(X_i, X)^2. \tag{1.2}$$

Such generalised means are called Fréchet means and will be described in detail in Section 1.5. However, some distances d allow us to write an estimator for (1.2) explicitly, and when possible we will report the relevant expressions here.

- *Log-Euclidean distance.* The logarithm and the exponential of the symmetric positive definite matrix X with SVD $X = U\Lambda U^T$ are $\log(X) = U\log(\Lambda)U^T$ and $\exp(X) = U\exp(\Lambda)U^T$ respectively, where $\log(\Lambda)$ and $\exp(\Lambda)$ are diagonal matrices with logarithm and exponential of the elements of Λ on the diagonal. The log-diagonal distance [Arsigny et al., 2007] between the covariance matrices X_1 and X_2 is defined as,

  $$d_L(X_1, X_2) = \| \log(X)_1 - \log(X)_2 \|. \tag{1.3}$$

  An estimation of the mean covariance matrix according to d_L is

  $$\overline{X}_L = \exp\left\{ \frac{1}{N} \sum_{i=1}^{N} \log(X_i) \right\}.$$

  Geodesics computed with respect to d_L are fully contained in the manifold. As a drawback, d_L cannot be computed for matrices which are not full rank.

- *Riemannian distance* [Schwartzman, 2006]. Also known as *trace metric* [Lawson and Lim, 2013]. It is a log-based distance as d_L and is defined as

  $$d_R(X_1, X_2) = \| \log(X_1^{-1/2} X_2 X_1^{-1/2}) \|. \tag{1.4}$$

Again, extrapolations along geodesics with respect to d_R are fully contained in the manifold. d_R does not admit a closed-form expression for the Fréchet mean and one has to rely on numerical estimation. However, the sample Fréchet mean computed with respect to d_R is unique, a fact that is in general not true for generic Fréchet means.

- *Cholesky distance and square root distance.* Both of these distances are based on the decomposition of the matrices X_i as $X_i = L_i L_i^\star$. For the Cholesky distance [Wang et al., 2004], L_i is the lower triangular matrix with positive diagonal yielding the Cholesky decomposition of X_i. We denote it as L_i^{chol}. In the square root distance [Dryden et al., 2009], L_i represents the positive square root of X_i. The distances are respectively given by

  $$d_C(X_1, X_2) = \|L_1^{chol} - L_2^{chol}\|, \tag{1.5}$$

  and

  $$d_C(X_1, X_2) = \|X_1^{1/2} - X_2^{1/2}\|. \tag{1.6}$$

  They both admit a least squares estimator for the Fréchet mean,

  $$\overline{X}_{C/H} = \left(\frac{1}{N}\sum_{i=1}^{N} L_i\right)\left(\frac{1}{N}\sum_{i=1}^{N} L_i\right)^T,$$

  with L_i being L_i^{chol} or $X_i^{1/2}$ depending on the distance considered.

The most relevant metric for the purposes of this work is the Procrustes-size-and-shape distance, and we will dedicate the next sections to its description. The Procrustes-size-and-shape distance is also based on a decomposition of the kind $X_i = L_i L_i^\star$, just in this case the L_i are optimised over rotations and reflections. In order to introduce it, we first need to lay out the background and talk about the field of Statistical Shape Analysis.

1.2.3 Statistical shape analysis and Procrustes size-and-shape distance

Statistical shape analysis uses statistical methods to study the geometrical properties of some given set of shapes. It was pioneered by the work of Kendall [1989]. Other relevant works were done by Bookstein et al. [1986] and more recently, and more relevantly to our context, by Dryden and Mardia [1998] and Dryden and Mardia [2016]. This section reviews some relevant results from Dryden and Mardia [1998]. No extended knowledge of Riemannian geometry is required. However, the reader who wishes to have further insights can find them in Barbosa and Carmo [1976] and Lang [2012].

The definition of shape in statistical shape analysis is the intuitive one. From Kendall [1989], the *shape* of an object is all the geometrical information that remains when location, scale and rotational effects are filtered out. We talk of *size-and-shape* when there is an interest in retaining scale information, as well as shape of the object. Two objects have the same size-and-shape when one of the two is a rigid body transformation of the other.

The way a shape is described mathematically is through a set of landmarks: salient points on each shape outline that match across and within populations. A landmark can be either anatomical, mathematical or a pseudo-landmark. Anatomical landmarks are biologically meaningful points, normally assigned by experts. Mathematical ones are location points on the object that retain some mathematical or geometrical meaning. Pseudo-landmarks are constructed points on an object either on the outline or between landmarks. Landmarks are encoded into $k \times m$ matrices $X \in \mathbb{R}^{k\times m}$ called configuration matrices, where k gives the number of landmarks in m dimensions.

To obtain a shape representation of an object according to Kendall [1989]'s definition, location, scale and rotations should be filtered out. This is done through establishing a *coordinate reference*. A common tool to obtain such a coordinate reference is turning shapes into *shape spaces*, that is, the sets of all possible shapes of the objects in question. More formally, denote by Z the pre-shape of a configuration matrix X, that is, the geometrical information about X which stays when location and scale are removed [Dryden and Mardia, 1998, Definition 4.4]. Mathematically we can write Z as

$$Z = \frac{HX}{\|HX\|},$$

where H is the Helmert sub-matrix H whose $j-$th row is given by

$$(\underbrace{h_j, \dots, h_j}_{j \text{ times}}, -jh_j, \underbrace{0, \dots, 0}_{k-j \text{ times}}), \quad h_j = -[j(j+1)]^{1/2}, \qquad \text{for} \quad j = 1, \dots, k-1. \tag{1.7}$$

The shape of X can be represented as

$$[X] = \{ZG : G \in SO(m)\},$$

with $SO(m)$ being the special orthogonal group of rotations. The shape of X is therefore an equivalence class under the action of $SO(m)$ and can be visualised by picking out a representative from the class (called an *icon*). The shape space $S_m^k \equiv \mathbb{R}^{k\times m}/SO(m)$ is the set of all orbits $[X]$ of the k-point set configurations in $\mathbb{R}^m$ under the action of the (Euclidean) similarity transformations [Kendall, 1989, Dryden and Mardia, 1998].
The set of shapes forms a Riemannian manifold containing the class object in question, and as such, is inherently not Euclidean (see Dryden and Mardia [1998, Chapters 1-4] for more details).

We talk of *size-and-shape* (or form) of a configuration matrix X to indicate all the geometrical information about X which is invariant under location and rotation (but not scale). The size-and-shape of X can be represented by the equivalence class

$$[X]_S = \{HXG \,:\, G \in SO(m)\}.$$

Quotienting out the size will return the shape of X. If moreover we remove reflections, we obtain the so called *reflection size-and-shape* of a configuration matrix X, represented as

$$[X]_{RS} = \{HXR \,:\, R \in O(m)\},$$

where $O(m)$ is the set of orthogonal transformations in m-dimensions.

Establishing a relationship between distances in shape space and the Euclidean distance in the original space will yield a *shape metric*. A commonly used shape metric is the so called *Procrustes (size and shape) distance*, which we will as well abbreviate Procrustes distance for convenience.

By an abuse of language, call X_1, X_2 two $k-$points configurations pre-multiplied by the Helmert submatrix of equation (1.7). This pre-multiplication will remove location information. The (squared) Procrustes size-and-shape distance between the size-and-shapes of X_1 and X_2 is found by minimising the Euclidean distance over rotations as in the following definition

$$\begin{aligned} d_P^2(X_1, X_2) &= \inf_{G \in SO(m)} \|X_1 - X_2 G\|^2 && (1.8) \\ &= \mathrm{tr}\,(\mathrm{X}_1^{\mathrm{T}}\mathrm{X}_1) + \mathrm{tr}\,(\mathrm{X}_2^{\mathrm{T}}\mathrm{X}_2) - 2 \sup_{\mathrm{G} \in \mathrm{SO(m)}} (\mathrm{X}_1^{\mathrm{T}}\mathrm{X}_2\mathrm{G}). && (1.9) \end{aligned}$$

d_P is a Riemannian distance in the size-and-shape space, which is the space of all size-and-shapes of k-points configurations in m dimensions.

Defining a metric allows us to define the concept of the *average shape* of shapes $X_1, \ldots, X_N$ as

$$\arg\min_{Y \in S_m^k} \sum_{i=1}^{N} d_P^2(Y, X_i).$$

The average shape is a Fréchet mean (Section 1.5) of $X_1, \ldots, X_N$. We will see in Section 1.5 that Fréchet means are computable through a generalised Procrustes algorithm, in this case called *generalized Procrustes analysis*. The algorithm has been shown to converge quickly to a solution [Dryden and Mardia, 1998, Gower, 1975] and it consists of an iterative procedure that involves translating and rotating the configurations relative to each other as to superimpose them "one on top of the other" and minimize the total sum of squares of their respective Euclidean distances [Dryden and Mardia, 1998, Chapter 5].

When G takes values in $O(m)$ rather than $SO(m)$ in (1.8), we talk of Procrustes *reflection size-and-shape distance*. Dryden et al. [2009] extended the notion of Procrustes reflection size-and-shape distance from reflection size-and-shapes to covariance matrices, and this is the topic of the next section.

1.2.4 Procrustes size-and-shape distance for covariance matrices

Covariance matrices play a fundamental role in many statistical applications. For example in Diffusion Tensor Imaging (DTI) [Schwartzman, 2006, Alexander et al., 2007] every 3D pixel contains a data atom identifiable with a 3×3 positive definite matrix and inference is performed on a dataset constituted of covariance matrices. In longitudinal data analysis [Daniels and Pourahmadi, 2002] we find similar datasets of covariances.

The positivity of the matrices sets a crucial constraint in the inferential procedure as it implies non-linearity. Consider, for example, a basic operation like computing the mean. A very common approach to estimate the mean covariance matrix is to assume that the data are sampled according to a scaled Wishart. In this case, the mean covariance can be found as a least squares (Euclidean) estimator, and coincides with the maximum likelihood estimator of the population covariance matrix. However, minimising the objective function in order to compute the least square estimator implies a choice of a metric, which in this case is the Euclidean one, while non-linearity should be taken into consideration for a better statistical analysis.

Dryden et al. [2009] compared several metrics on the space of covariance matrices (Section 1.2.2), favouring especially the Procrustes size-and-shape metric inspired by shape theory. Dryden et al. [2009] defined the Procrustes distance between covariances $S_1, S_2 \in \mathbb{R}^{k\times k}$ as

$$\Pi(S_1, S_2) = \inf_{R:R^TR=I} \|{S_1}^{1/2} - {S_2}^{1/2}R\|_2. \tag{1.10}$$

The unique non-negative matrix roots $S_i^{1/2}$ in (1.10) can be replaced by any matrices Y_i such that $S_i = Y_i Y_i^T$. For example, the Y_i can be chosen via the Cholesky decomposition of S_i.

The optimal matching in (1.10) is attained at

$$\hat{R} = V_R V_L^T, \tag{1.11}$$

where $Y_1^T Y_2 = V_L \Lambda V_R^T$ is the SVD of $Y_1^T Y_2$ with $V_L, V_R \in O(k)$ orthonormal matrices and Λ a diagonal matrix of positive singular values.

To make the connection with shape theory and Section 1.2.3, Π is the reflection size-and-shape distance between the configurations $H^T S_1^{1/2}$ and $H^T S_2^{1/2}$, where H is the Helmert sub-matrix in (1.7).

The connection between shapes and covariances is better understood through the concept of Gram matrices: the Gram matrix of a set of vectors $X = \{x_1, \dots, x_k\}$ (stored by column) is $S = X^T X$, i.e. the matrix encoding all possible inner products between the X_i's. S can be thought of as a way to parametrize the shapes X_1, X_2 of two configurations of k points in $\mathbb{R}^m$. From (1.8) we know that the Procrustes matching between two configurations HX_1, HX_2 depends only on the Gram matrices $S_1 = X_1 X_1^T$ and $S_2 = X_2 X_2^T$ of X_1 and X_2 and the mixed term $X_1^T X_2$. Since Gram matrices are symmetric and non-negative definite, it is possible to deduce a metric on the space of covariance matrices S_i by taking inspiration from distances between shapes X in $\mathbb{R}^{k\times k}$.

Dryden et al. [2009] also describe the geometry induced by the Procrustes size-and-shape distance. The tangent space provides a linearized approximation to a manifold in the neighbourhood of a particular point. Exploiting the tangent space linearity, one can apply Euclidean metric on tangent space coordinates and use it as an approximation to a non-linear metric onto the manifold (or, in this case, the shape space).
In the notation $S_i = Y_i Y_i^T$, $i = 1,2$, the horizontal tangent coordinates T of Y_2 with pole Y_1 are given by [Kendall et al., 2009]

$$T = Y_2\hat{R} - Y_1, \qquad \hat{R} = \inf_{R\in O(k)} \|Y_1 - Y_2 R\|,$$

and satisfy $Y_1 T^T = T Y_1^T$. Here $\hat{R}$ is the matrix giving the optimal matching as in (1.11).

As the tangent space provides only a local linear approximation to the manifold, the tangent space coordinates depend on the specific point where the tangent space is computed. For a sample $S_1,\ldots,S_N$ a natural choice of pole is given by the Fréchet mean of equation (1.2) (see also Section 1.5). If $\bar{\Sigma} = \bar{Y}\bar{Y}^T$ is the decomposition of the Fréchet mean, then the tangent space coordinates of S_i with pole $\bar{\Sigma}$ are expressed as

$$T_i = \bar{Y} - Y_i\hat{R}_i,$$

where $\bar{R}_i$ gives the rotation optimally matching Y_i onto $\bar{Y}$, $i = 1,\ldots,N$ [Dryden et al., 2009].

The minimal geodesic starting from Y_1 and ending at Y_2 is the minimal length path between the two points which is fully contained in the manifold, and can be expressed as

$$t_1 Y_1 + t_2 Y_2 \hat{R},$$

where $t_1 + t_2 = 1$, $t_1, t_2 \geq 0$.

In order to analyse linguistic data, Pigoli et al. [2014a] pointed out how one can be interested in the variation that can be found in speech frequency intensity within different languages. Such variations are better captured by covariance operators, and therefore it is required to have a suitable metric to compare such operators. To this purpose, Pigoli et al. [2014a] showed that the Procrustes distance (1.10) can be generalised to the infinite-dimensional space of covariances on $L^2(0,1)$. In the next sections, we will describe the extension of the Procrustes distance to the functional case and we will explore its connection to the field of optimal transportation and the Wasserstein metric between Gaussian distributions.

1.2.5 Procrustes distance for operators

Inference on the analysis of the variation of functional data in their covariance structure is well-documented, e.g. in Gabrys et al. [2010], Horváth et al. [2013], Paparoditis and Sapatinas [2014], Kraus [2014]. What is common among these works is that they exploit the immersion of the space of covariances into the wider Hilbert–Schmidt space, implicitly ignoring the intrinsic non-linearity of these operators.

Section 1.2.2 gave a summary of various distances which can be employed while doing inference on covariance matrices. When moving to the operator case, the trace-class structure of the covariances entails that not all matrix-based distances admit a well-defined infinite dimensional correspondent. The difference of logarithms entering into the log-Euclidean distance for example, is not readily extendable to operators whose eigenvalues decay to zero. A similar argument holds for the Riemannian distance, as it includes square root inverses, which are generally unbounded. Pigoli et al. [2014a] showed that the Procrustes distance favoured by Dryden et al. [2009] admits a well-defined functional generalisation, and we devote this section to its description.

From the definition of the nuclear and HS norm in Section 1.1.1, we can observe that trace class operators can be seen as "squares" of Hilbert–Schmidt operators. This establishes a parallelism between a covariance operator Σ together with its root $\Sigma^{1/2}$ and the pair $S, S^{1/2}$ entering in (1.10). Pigoli et al. [2014a] showed that (1.10) is well-defined when extended to covariance operators in $L^2[(0,1),\mathbb{R}]$. The functional extension of d_P does actually apply to any separable Hilbert space $\mathcal{H}$ [Masarotto et al., 2018] and we report here this more general definition.

Definition 3 (Procrustes Metric on Covariance Operators)**.** *For any pair of nuclear and non-negative linear operators $\Sigma_1, \Sigma_2 : \mathcal{H} \times \mathcal{H} \to \mathcal{H}$ on the separable Hilbert space $\mathcal{H}$, we define the Procrustes metric as*

$$\Pi(\Sigma_1, \Sigma_2) = \inf_{U:\, U^* U = \mathfrak{I}} \|\Sigma_1^{1/2} - \Sigma_2^{1/2} U\|_2, \tag{1.12}$$

where $\{U : U^* U = \mathfrak{I}\}$ is the set of unitary operators on $\mathcal{H}$.

As already mentioned, the work of Pigoli et al. [2014a] was sparked by the need to analyse linguistic data. More specifically, their dataset is a series of recordings of people pronouncing a collection of words in different Romance languages. The covariances on which the test is applied are estimated from different recorded frequency intensities in the log-spectrogram. Covariance operators have been shown to well characterise the phonetic structure of a language [Aston et al., 2010], in the sense that difference between the operators are linked to phonetic differences between languages. The investigation of similarities and relationships across languages typically relies on written documentation, so their methods allow to complement the textual analysis with a phonetic one.

They propose a 2-sample testing procedure on covariances which respects the curved geometry of the space. Since in practice they observed a finite-dimensional representation of the operators, they also consider the behaviour of Π under finite-dimensional projection, showing that the distance between two finite-dimensional projections converges to that between the two infinite-dimensional operators.
Moreover Pigoli et al. [2014a] show that the Fréchet mean with respect to the Procrustes distance can be computed in practice via a version of the generalised Procrustes algorithm Gower [1975], which we describe in the next paragraph.

Generalized Procrustes algorithm on operators

Pigoli et al. [2014a] proposed the following adaptation of the Generalized Procrustes Algorithm. It alternates registration and averaging steps, and a high level description is as follows.

- Initialise the algorithm at L^0, taken to be the average of $L_i^0 = L_i = \Sigma_i^{1/2}$.
- At step k, compute, for each i, the unitary operator R_i that minimises $\|L^{k-1} - L_i^{k-1} R_i\|_2$, and set $L_i^k = L_i^{k-1} R_i$.
- Define L^k as the average of $\{L_1^k, \dots, L_N^k\}$ and repeat until convergence.

Pigoli et al. report that in practice, if suitably initialised at $N^{-1} \sum_{i=1}^N \Sigma_i^{1/2}$, this algorithm converges to a local minimum, which corresponds to an estimate of the Fréchet mean.

The advantage of this procedure is that it only involves successively matching pairs of operators while minimising $\|L^{k-1} - L_i^{k-1} R_i\|_2$. This pairwise matching admits an explicit solution depending on the product of the square roots of the operators in question and their singular value decomposition. However, for operators, there is no theoretical guarantee of the convergence of the algorithm, and it is not clear whether it converges when $\{\Sigma_1, \dots, \Sigma_N\}$ do not commute.

The work of Pigoli et al. [2014a] has the significant value of pioneering a non-Euclidean analysis of covariance operators. They give the infinite-dimensional formulation of the Procrustes distance, showing that it is computationally valid in applications and it behaves well in terms of finite-dimensional approximation, in the sense that the distance between two finite-dimensional projections converges to the distance between their functional counterparts.
Their work spawned many further questions about this metric. As we just mentioned, the theoretical convergence of the Procrustes algorithm is one of them. However, one might also wonder whether we can derive further properties of the Procrustes Fréchet mean (such as uniqueness) or whether other inferential procedures (e.g., PCA) can be implemented under the Procrustes framework. Maybe the most relevant question concerns the geometrical and statistical interpretation of the Procrustes metric: the size-and-shape distance for covariance matrices admitted a dual interpretation in terms of shape theory, and we wish to establish

a similar connection for the functional case and this way gain a better understanding of the geometry of the space. These topics will be addressed in the next few sections and in the next chapter. In particular we will show that some of these problems can be better understood through the lens of optimal transport, and can be answered thanks to the intimate connection between the Procrustes distance and the Wasserstein metric between Gaussian processes. In order to do this we need to introduce the relevant notions from Optimal transport and Wasserstein spaces, and this is done in the next sections.

1.3 Optimal Transport and Wasserstein Spaces

Optimal transport aims at investigating how to transfer one mass distribution into another at a minimal cost. This problem was formally treated for the first time by Monge [1781], who heuristically formulated it in terms of a mass of sand and a pit, asking what would be the optimal way to move such sand into such a pit. The question can be translated into mathematical terms as follows. Given two probability maps μ and ν on some spaces $\mathscr{H}$ and $\mathscr{Y}$ and a cost function $c : \mathscr{H} \times \mathscr{Y} \to \mathbb{R}$, minimize the transportation cost

$$C(T) = \int_{\mathscr{H}} c(x, T(x))\, d\mu(x)$$

among all *transport maps* T, i.e., Borel maps such that $\mu(T^{-1}(E)) = \nu(E)$ for all Borel subsets E of $\mathscr{Y}$. In terms of Monge's sandpit problem, the above conditions state that the amount of sand to go into a hole of volume $\nu(E)$ has to have exactly the mass to fit that volume, that is, mass cannot be compressed or inflated. Mathematically we say that T pushes μ forward to ν and write that $\nu = T\#\mu$.
The push-forward is characterised by the fact that

$$\int_{\mathscr{Y}} f\, d(T\#\mu) = \int_{\mathscr{H}} f \circ T\, d\mu,$$

for every Borel function $f : \mathscr{Y} \to \mathbb{R} \cap \{\pm\infty\}$[2].

Monge's problem also admits a probabilistic interpretation, more relevant to this thesis. Before giving the details, we need to introduce some basic definitions of random elements in general metric spaces.

We say that X is a *random element* in a separable Hilbert space $\mathscr{H}$ (in any topological space actually), if X is a measurable function from a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ into $\mathscr{H}$. The *probability distribution*, or probability law, of X, is the Borel measure μ_X such that $\mu_X(E) = \mathbb{P}(X \in E)$ for all Borel sets E, that is, $\mu_X = X\#\mathbb{P}$. Given two random elements X and Y in $\mathscr{H}$ and $\mathscr{Y}$ and a cost function $c : \mathscr{H} \times \mathscr{Y} \to \mathbb{R}$, Monge's problem is the minimisation of the expected value of c over all measure preserving measurable functions T, that is, all functions such that $T(X)$ and

---

[2] The above identity implicitly assumes that one of the integral exists, possibly attaining an infinite value, if and only if the other exists, and in this case the values are equal.

Y have the same distribution. In mathematical terms, this becomes the minimisation of

$$C(T) = \int_{\mathscr{H}} c(x, T(x)) d\mu(x) = \mathbb{E} c(X, T(X)).$$

Monge's problem is mathematically very difficult, starting from the fact that the existence of a way to optimally transport a mass is not in general guaranteed. For example, for a linear cost function $c(x, y) = |x - y|$ and $\mu = \delta_0$, $\nu = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$ on the interval $[-1, 1]$, no transport map exists. Indeed, the set $E = \{T(0)\}$ satisfies $\mu(T^{-1}(E)) > \nu(E)$, so there cannot be such a T. The key breakthrough in the resolution of Monge's optimal transport problem was made by Kantorovich [1942], who proposed a reformulation of Monge's problem which allows the mass to be split at each single point $x \in \mathscr{H}$ according to some probability measure μ_x (rather than being sent entirely to $T(x)$).
This work earned Kantorovich the Nobel prize for Economics in 1975 and as a consequence, the optimal transport problem is now largely known as the Monge–Kantorovich problem.

We now present a more formal description of Kantorovich's version of the problem. Consider a measure π on the product space $\mathscr{H} \times \mathscr{Y}$ and let $A \subseteq \mathscr{H}$ and $B \subseteq \mathscr{Y}$. The total mass sent to A and B is $\pi(A \times \mathscr{Y})$ and $\pi(\mathscr{H} \times B)$ respectively. If these quantities equal exactly the masses of A and B, that is if

$$\pi(A \times \mathscr{Y}) = \nu(A), \tag{1.13}$$

$$\pi(\mathscr{H} \times B) = \nu(B), \tag{1.14}$$

for every $A \subseteq \mathscr{H}$ and $B \subseteq \mathscr{Y}$ measurable sets, then we say that π is a *transport plan*, or a *coupling* of μ and ν. In the latter case, μ and ν are called *marginals* of the coupling π.
If ν is the push-forward of μ by T, we can define a transport plan $\pi = (id \times T)\#\mu$ via

$$\pi(A \times B) = \mu(\{x \in A \, : \, T(x) \in B\}) = \mu\left(A \bigcap T^{-1}(B)\right).$$

In other words, transport plans can be seen as joint measures on the product space with given marginals.
The Kantorovich problem then translates into minimising the total cost

$$C(\pi) = \int_{\mathscr{H} \times \mathscr{Y}} c(x, y) d\pi(x, y)$$

over the class of transport plans

$$\Pi(\mu, \nu) = \left\{\pi \in P(\mathscr{H} \times \mathscr{Y} \, : \, \pi(A \times Y) = \mu(A) \text{ and } \pi(X \times B) = \nu(B)\right\},$$

for every $A \subseteq \mathscr{H}$ and $B \subseteq \mathscr{Y}$ measurable sets.

The probabilistic interpretation of the Kantorovich problem can be written as the minimisation of

$$C(\pi) = \int_{\mathscr{X}\times\mathscr{Y}} c(x,y)\,d\pi(x,y) = \mathbb{E}_\pi c(X,Y)$$

over all couplings $\pi = (X,Y)\#\mathbb{P}$ with marginals X, Y.

To see the connection between a transport plan and a transport map, think that transport plans can be seen as "multivalued" transport maps: indeed π can be represented as a collection of probability measures $\pi_x \in P(\{x\}\times Y)$ such that $\pi = \int \pi_x d\mu(x)$ (see [Dudley, 2018, 31, Section 10.2]).

The following relationship holds between Monge's and Kantorovich's formulation:

$$C(\pi) \leq C(T).$$

A transport map always induces a transport plan of the same cost: if $T\#\mu = \nu$ then $\pi = (Id \times T)\#\mu$ is a transport plan, making it clear that the set of solutions of the Kantorovich problem is at least as large as the set of solution of the Monge one. Furthermore the set of transport plans is never empty, because it always contains the product measure $\mu\otimes\nu$ defined by $[\mu\otimes\nu](A) = \mu(A)\nu(B)$ for measurable $A \subseteq \mathscr{X}$, $B \subseteq \mathscr{Y}$, and the objective function is also linear in π, thus overcoming one of the difficulties in Monge's problem. Moreover, the set of transport plans is convex and compact with respect to the narrow topology in $P(\mathscr{X}\times\mathscr{Y})$ [Evans, 1992], and if C is lower semicontinuous and bounded from below, then there always exists a minimiser of the Kantorovich problem [Zemel, 2017, Villani, 2008]. We call the minimiser of $J(\pi)$ an *optimal trasport plan.*

When an optimal transport map $T : \mathscr{X} \to \mathscr{Y}$ exists, the optimal transport plan and the optimal transport map are related through

$$\int_{\mathscr{X}\times\mathscr{Y}} c(x,y)\,d\pi(x,y) = \int_{\mathscr{X}} c(x,T(x))\,d\mu(x).$$

The Monge–Kantorovich problem is particularly relevant also for its many connections with other fields of mathematics, which have emerged in the last couple of decades. Seminal work in this is due to Benamou and Brenier [2000] on fluids and, earlier, McCann [1997] on gas mechanics (see also Sections 1.3.1 and 1.4.2 respectively), but applications expands to shape optimisation, non-linear diffusion, partial differential equations and Riemannian geometry, to name a few (Benamou and Brenier [2000], McCann [1997], Gangbo and Święch [1998], von Renesse and Sturm [2009], Rüschendorf and Uckelmann [2002], Villani [2003]).

1.3.1 Wasserstein distance

The $p-$Wasserstein space on the separable Hilbert space $\mathscr{H}$ is defined by

$$\mathscr{W}_p = \mathscr{W}_p(\mathscr{H}) = \left\{ \mu \in P(\mathscr{H}) : \int_{\mathscr{H}} ||x||^p d\mu(x) \infty \right\}, \quad p \geq 1,$$

where $P(\mathscr{H})$ is the space of probability measure on $\mathscr{H}$. The p-Wasserstein space is closely related to the Monge–Kantorovich problem of Section 1.3, since the minimum value in the latter gives rise to a metric on the space of probability measures of $\mathscr{H}$ called the *p-Wasserstein distance*. More precisely, the p-Wasserstein distance between μ, $\nu \in P(\mathscr{H})$ is defined as

$$\mathscr{W}_p(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} J_p(\pi) \right)^{1/p} = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathscr{H} \times \mathscr{H}} ||x_1 - x_2||^p d\pi(x_1, x_2) \right)^{1/p}; \tag{1.15}$$

that is, is the minimum of the Kantorovich problem for the cost function $c_p(x, y) = \|x - y\|^p$. For the purpose of this thesis, unless differently stated, we always assume that $1 \leq p < \infty$, and in fact we will mostly treat the case $p = 2$. Extension to $0 < p < 1$ is possibile by removing the power $1/p$ in the definition of $\mathscr{W}$.
We refer to Villani [2003] for a proof that $\mathscr{W}_p$ is indeed a metric. However from (1.15) it follows immediately that $\mathscr{W}_p$ is non-negative, symmetric and that $W_p(\mu, \nu) = 0$ only along the "diagonal", that is, if and only if $\mu = \nu$. Finally, the Wasserstein distance between $\mu, \nu \in \mathscr{W}_p(\mathscr{H})$ is finite, since

$$\|x_1 - x_2\|^p \leq 2^p \|x_1\|^p + 2^p \|x_2\|^p.$$

1.4 Procrustes–Wasserstein distance and optimal transportation of Gaussian processes

Once the existence of the optimal transport plan is established and attained at some coupling π for any marginal pair of measures $\mu, \nu \in \mathscr{W}(\mathscr{H})$, a natural question is whether this optimal minimiser is also unique. In general the answer to this question is negative, unless some regularity conditions on the measures are imposed.
If, moreover, the two measures μ and ν are Gaussian on $\mathscr{H}$, then not only the solution is unique, but under some regularity conditions, both the Wasserstein distance and the optimal transport map admit a closed-form expression.

1.4.1 Optimal transportation of Gaussian processes

A random element X in a separable Hilbert space $(\mathscr{H}, \langle \cdot, \cdot \rangle)$ is *Gaussian* with mean m and covariance $\Sigma : \mathscr{H} \times \mathscr{H} \to \mathscr{H}$ if

$$\langle X, h \rangle \sim \mathscr{N}(\langle m, h \rangle, \langle h, \Sigma h \rangle)$$

for all $h \in \mathcal{H}$. A *Gaussian measure* is the law of a Gaussian random element.

We say that a Gaussian measure is *regular* if and only if its covariance operator is injective (for a more general definition of regularity see Ambrosio et al. [2008, Definition 6.2.2]). Moreover we say that a couple is deterministic if it is manifested as the joint distribution of $(X, T(X))$ for some optimal deterministic map $T : \mathcal{H} \to \mathcal{H}$.

Now assume that μ is regular. In this case the optimal coupling is unique and given by a deterministic coupling $\pi = (\mathcal{I}, \mathbf{t}_{\mu}^{\nu})\#\mu$ [Villani, 2003], $\mathbf{t}_{\mu}^{\nu}$ being the optimal plan pushing μ onto ν. In addition *Brenier's theorem* establishes that the optimal map $\mathbf{t}_{\mu}^{\nu}$ can be recovered as a gradient of a convex function. See Brenier [1991] for a reference of Brenier's theorem, although the result was proved independently by other authors (Cuesta-Albertos and Matrán [1989]; Knott and Smith [1984]; Rüschendorf and Rachev [1990]). We underline that as a consequence of Brenier's theorem, any (finite) non-negative linear combination of optimal transport maps remains optimal, as convexity is preserved under such combinations.

Regardless of the useful characterisation of Brenier's theorem, the optimal transportation map $\mathbf{t}_{\mu}^{\nu}$ for a regular μ and the corresponding Wasserstein distance $W(\mu, \nu) = \sqrt{\int_{\mathcal{H}} \|x - \mathbf{t}_{\mu}^{\nu}(x)\|^2 \, \mathrm{d}\mu(x))}$ rarely admit closed-form expressions. Gaussian processes are a notable exception to this.

Let $\mu \sim \mathcal{N}(m_1, \Sigma_1)$ and $\nu \sim \mathcal{N}(m_2, \Sigma_2)$. Then the Wasserstein distance between μ and ν is

$$W^2(\mu, \nu) = \|m_1 - m_2\|^2 + \operatorname{tr}(\Sigma_1) + \operatorname{tr}(\Sigma_2) - 2\operatorname{tr}\sqrt{\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}}. \tag{1.16}$$

In finite dimension, this result was obtained independently by Dowson and Landau [1982] and Olkin and Pukelsheim [1982]. For a proof in infinite dimensions, the reader is referred to Cuesta-Albertos et al. [1996].

We turn now out attention to optimal maps. To lighten the notation, we assume throughout the rest of the Section that μ and ν are centered, i.e., $m_1 = m_2 = 0$. We also adopt the notation $\mathbf{t}_{\Sigma_1}^{\Sigma_2}$ in place of $\mathbf{t}_{\mu}^{\nu}$. In finite dimension, invertibility of Σ_1 guarantees the existence and uniqueness of a deterministic optimal coupling of $\mu \sim N(0, \Sigma_1)$ and $\nu \sim N(0, \Sigma_2)$, induced by the linear transport map

$$\mathbf{t}_{\Sigma_1}^{\Sigma_2} := \Sigma_1^{-1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\Sigma_1^{-1/2}.$$

Cuesta-Albertos et al. [1996, Proposition 2.2] proved that once some regularity is assumed, the previous formula also holds in infinite-dimensional Hilbert spaces.

Proposition 4. *Let $\mu \sim N(0, \Sigma_1)$ and $\nu \sim N(0, \Sigma_2)$ be centred Gaussian measures in $\mathcal{H}$ and suppose that* $\ker(\Sigma_1) \subseteq \ker(\Sigma_2)$ *(equivalently,* $\overline{\operatorname{range}(\Sigma_1)} \supseteq \overline{\operatorname{range}(\Sigma_2)}$*). Then there exists a linear subspace of $\mathcal{H}$ with μ-measure 1, on which the optimal map is well-defined and is given by the linear operator*

$$\mathbf{t}_{\Sigma_1}^{\Sigma_2} = \Sigma_1^{-1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\Sigma_1^{-1/2}.$$

Section 2.1 will make the key observation on how the Procrustes and the Wasserstein distance coincide on the space of centered Gaussian process. We will therefore focus our attention to the Gauss–Wasserstein space, i.e., the space of Gaussian measures equipped with the 2-Wasserstein distance.

1.4.2 The tangent bundle and geodesics in Wasserstein space

In differential geometry, a geodesic is a generalisation to what a straight line is in Euclidean spaces, that is, it is connected to the idea of the shortest path between two points. We say that a curve $\gamma : [0,1] \to X$ is a *constant speed geodesic* if $d(\gamma_t, \gamma_s) = |t-s| d(\gamma_0, \gamma_1)$, $\forall t, s \in [0,1]$. In Hilbert spaces, for every two points x, y there is only one constant speed geodesic connecting them, which is the curve $t \mapsto (1-t)x + ty$ (e.g. Villani [2008]). The latter result holds as well true for Gaussian measures in Wasserstein space.

Let $\mu, \nu \in \mathcal{W}(\mathcal{H})$ be such that the optimal map from μ to ν, $\mathbf{t}_\mu^\nu$ exists. Recall that when $\mu \sim \mathcal{N}(0, \Sigma_1)$, $\nu \sim \mathcal{N}(0, \Sigma_2)$ a sufficient condition is that $\ker(\Sigma_1) \subseteq \ker(\Sigma_2)$. Moreover we know that in the Gaussian case, the Monge–Kantorovich problem admits a unique solution, namely the transport map $\mathbf{t}_\mu^\nu$. Let $\mathfrak{I} : \mathcal{H} \to \mathcal{H}$ be the identity map on $\mathcal{H}$. We can define a curve

$$\mu_t = \left[\mathfrak{I} + t(\mathbf{t}_\mu^\nu - \mathfrak{I})\right] \# \mu, \qquad t \in [0,1], \tag{1.17}$$

known as McCann's interpolation (McCann [1997, Equation (7)]) and constructed by interpolating the optimal transport map and the identity.

Equation (1.17) defines a constant speed geodesic, since clearly $\mu_0 = \mu$, $\mu_1 = \nu$ and it respects

$$W(\mu_t, \mu_s) = (t-s) W(\mu, \nu), \qquad 0 \le s \le t \le 1.$$

To better understand the last equality, notice that for any two Borel maps $t_1, t_2 : \mathcal{H} \to \mathcal{Y}$, $(t_1, t_2)\#\mu$ is a valid transport plan for the measures $t_1\#\mu$ and $t_2\#\mu$ with cost $\int W_{\mathcal{Y}}^2(f(x), g(x)) d\mu(x)$, and since W_2^2 is minimised by the optimal transport plan we obtain the inequality

$$W_2(t_1\#\mu, t_2\#\mu) \le \left(\int_{\mathcal{H}} \| t_1(x) - t_2(x) \|^2 d\mu(x)\right)^{1/2}. \tag{1.18}$$

Now, applying (1.18) to both $W_2(\mu_t, \mu)$ and $W_2(\mu_t, \nu)$, we obtain the geodesic condition (see also Zemel [2017, Subsection 3.3.1]).

1.4.3 The tangent bundle

In this subsection, we describe the tangent bundle of $\mathcal{W}(\mathcal{H})$ and we characterise optimal transport maps as tangent space elements. We follow von Renesse and Sturm [2009] and Ambrosio et al. [2008].

We start by introducing the L_2-like space and norm of Borel functions $f : \mathcal{H} \to \mathcal{H}$ by

$$\|f\|_{\mathscr{L}_2(\mu)} = \left(\int_{\mathcal{H}} \|f(x)\|^2 \,\mathrm{d}\mu(x)\right)^{1/2}, \qquad \mathscr{L}_2(\mu) = \{f : \|f\|_{\mathscr{L}_2(\mu)} < \infty\}.$$

If we now fix a reference measure μ, any other measure ν such that the optimal map $\mathbf{t}_\mu^\nu$ exists uniquely can be identified with $\mathbf{t}_\mu^\nu$ itself. Subtracting the identity $\mathfrak{I}$ from $\mathbf{t}_\mu^\nu$ creates a correspondence between a subset of the Wasserstein space with a subset of the linear space $\mathscr{L}_2(\mu)$, in such a way that μ itself corresponds to $0 \in \mathscr{L}_2(\mu)$.
This motivates the following definition of the tangent space of $\mathcal{W}(\mathcal{H})$ at μ (Ambrosio et al. [2008, Definition 8.5.1]):

$$\mathrm{Tan}_\mu = \overline{\{t(\mathbf{t} - \mathfrak{I}) : \mathbf{t} \in \mathscr{L}_2(\mu); \mathbf{t} \text{ optimal between } \mu \text{ and } \mathbf{t}\#\mu; t > 0\}}^{\mathscr{L}_2(\mu)}.$$

Tan_μ inherits the inner product of $\mathscr{L}_2(\mu)$,

$$\langle \mathbf{s}, \mathbf{r} \rangle_\mu = \int_{\mathcal{H}} \langle \mathbf{s}(x), \mathbf{r}(x) \rangle \,\mathrm{d}\mu(x), \qquad \mathbf{s}, \mathbf{r} \in \mathscr{L}_2(\mu).$$

Notice that Tan_μ is a strict subset of $\mathscr{L}_2(\mu)$. For example, a bounded linear operator A is always an element of $\mathscr{L}_2(\mu)$, but it belongs to the tangent space if and only if A is self-adjoint. Moreover, Tan_μ is a linear space. Linearity comes from an equivalent definition [Ambrosio et al., 2008, Definition 8.4.1] of the tangent space in terms of cylindrical functions.

In differential geometry, points on the manifold can be lifted to the tangent space via the Riemannian logarithm map. Viceversa, tangent space vectors can be retracted onto the manifold via the Riemannian exponential map.

The exponential map $\exp_\mu : \mathrm{Tan}_\mu \to \mathcal{W}(\mathcal{H})$ at μ is the restriction of the transformation that sends $\mathbf{r} \in \mathscr{L}_2(\mu)$ to $(\mathbf{r} + \mathfrak{I})\#\mu \in \mathcal{W}(\mathcal{H})$. Specifically,

$$\exp_\mu(t(\mathbf{t} - \mathfrak{I})) = [t(\mathbf{t} - \mathfrak{I}) + \mathfrak{I})\#\mu = [t\mathbf{t} + (1 - t)\mathfrak{I}]\#\mu, \quad t \in \mathbb{R}.$$

When μ is regular, the log map $\log_\mu : \mathcal{W}(\mathcal{H}) \to \mathrm{Tan}_\mu$, is well-defined throughout $\mathcal{W}(\mathcal{H})$, and given by

$$\log_\mu(\nu) = \mathbf{t}_\mu^\nu - \mathfrak{I}.$$

It is the (surjective) right inverse of the exponential map:

$$\exp_\mu(\log_\mu(\nu)) = \nu, \quad \nu \in \mathcal{W}, \qquad \log_\mu(\exp_\mu(t(\mathbf{t} - \mathfrak{I}))) = t(\mathbf{t} - \mathfrak{I}) \quad t \in [0, 1],$$

because convex combinations of optimal maps are optimal maps as well (Masarotto et al. [2018, Subsection 2.2]), and so the log map is bijectively mapping McCann's interpolant $\left[\mathfrak{I} + t(\mathbf{t}_\mu^\nu - \mathfrak{I})\right]\#\mu$ to the line $t(\mathbf{t}_\mu^\nu - \mathfrak{I}) \in \mathrm{Tan}_\mu$. In Wasserstein space, the tangent space and the exponential and log maps can be thought of as Riemannian ones, since optimal maps arise as minimal tangent vectors to absolutely continuous curves in Wasserstein space [Ambrosio et al., 2008, Sections 8.4–8.5].

1.5 Fréchet means in general metric spaces

Fréchet means are the extensions to a general metric space of what arithmetic means are in Euclidean space. Let $x_1, \dots, x_n$ be elements of a normed vector space or more generally an Hilbert space $\mathscr{H}$. The empirical mean $\overline{x}$ of $x_1, \dots, x_n$ is the unique minimiser of the sum of squared distances from the $x_i's$, that is

$$\overline{x} = \text{argmin}_y F(y) = \arg\min_y \sum_{i=1}^{n} \|y - x_i\|^2, \qquad y \in \mathscr{H}.$$

This definition can be generalized beyond vector spaces. Such generalisation is credited to Fréchet [1948] and, in the context of Wasserstein spaces, was dealt with for the first time by Agueh and Carlier [2011], who considered empirical Fréchet means in $\mathscr{W}_2(\mathbb{R}^d)$. A formal definition is the following.

Definition 5 (Empirical Fréchet mean and empirical Fréchet functional)**.** . *Let* $\mu_1, \dots, \mu_n \in \mathscr{W}_2(\mathscr{H})$*. A Fréchet mean of* $(\mu, \dots, \mu_n)$ *is a minimiser in* $\mathscr{W}_2(\mathscr{H})$ *of the associated Fréchet functional* $F : \mathscr{W}_2(\mathscr{H}) \to \mathbb{R}$

$$F(\gamma) = \frac{1}{2N} \sum_{i=1}^{N} W_2^2(\gamma, \mu_i), \qquad \gamma \in \mathscr{W}_2,$$

if such a minimiser exists.

Fréchet means are also known as of *barycenters*, and we might use either term in the sequel.

When it comes to covariance operators $\Sigma_1, \dots, \Sigma_n$, we can define their Fréchet mean with respect to the Procrustes metric as the Wasserstein Fréchet mean of the corresponding centered Gaussian processes, that is, as the minimiser of the Fréchet functional

$$F(\Sigma) = \frac{1}{2N} \sum_{i=1}^{n} \Pi^2(\Sigma, \Sigma_i) = \frac{1}{2N} \sum_{i=1}^{n} W^2(\mathscr{N}(0,\Sigma), \mathscr{N}(0,\Sigma_i)).$$

Empirical Fréchet means extend to their corresponding population version by replacing summations by expectation:

Definition 6 (Population Fréchet mean)**.** . *Let* μ *be a random measure in* $\mathscr{W}$*. The Fréchet mean of* μ *is the minimiser of the Fréchet functional* $F : \mathscr{W} \to \mathscr{W}$

$$F(\gamma) = \frac{1}{2} \mathbb{E} W_2^2(\gamma, \mu), \qquad \mu \in \mathscr{W}$$

if such a minimiser exists uniquely.

For covariance operators, considering the population mean corresponds to consider the Fréchet mean of a random operator $\mathscr{A}$, and is given by the minimiser of the functional

$$F(\Sigma) = \frac{1}{2}\mathbb{E}\Pi^2(\Sigma, \mathscr{A}).$$

When $\mathscr{A}$ is uniformly distributed over the finite set $\{\Sigma_1, \dots, \Sigma_n\}$ then the population Fréchet mean reduces to the empirical one.

1.5.1 Existence, uniqueness and the Gaussian case

Existence and uniqueness of a Fréchet mean is in general not guaranteed. Bhattacharya and Patrangenaru [2003, 2005] and Karcher [1977] give conditions for the existence and uniqueness of Fréchet means in a Riemannian manifold. In particular, Bhattacharya and Patrangenaru [2003] establish the existence of a unique Fréchet mean of a probability measure Q on a completely metric space (M, ρ), when M has non-positive curvature and Q is sufficiently concentrated with bounded Fréchet functional $F(p) = \int_M \rho^2(p,x)Q(dx)$, $p \in M$. In general, existence proofs are easier, but even if a Fréchet mean exists, there is no guarantee of its uniqueness. The reader is referred to Bhattacharya and Patrangenaru [2003, 2005], Karcher [1977] for general Fréchet means, to Zemel [2017] for Fréchet mean in $\mathscr{W}^2$ and to Le Gouic and Loubes [2016] for $\mathscr{W}^p$.

Wasserstein spaces represent a notable exception, since relatively weak assumptions made it possible to establish existence and uniqueness. When $\mathscr{H} = \mathbb{R}^d$, Agueh and Carlier [2011] provide necessary and sufficient conditions for γ to be the unique Fréchet mean of absolutely continuous measures $\mu_1, \dots, \mu_N \in \mathscr{W}_{\in}(\mathscr{H})$ in terms of the convex potentials of $\mathbf{t}_{\mu}^{\gamma}$. Zemel [2017, Theorem 4.2.4] gives conditions on the population version.

If $\mu_1, \dots, \mu_N$ are Gaussian on $\mathbb{R}^d$, Knott and Smith [1984] show that a Fréchet mean for their respective covariances $\Sigma_1, \dots, \Sigma_n$ is a positive definite solution Σ to the equation

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}(\Sigma^{1/2}\Sigma_i\Sigma^{1/2})^{1/2}.$$

Existence of a solution of this equation was also proved in Rüschendorf and Uckelmann [2002]. Agueh and Carlier [2011] extended this result, proving that (2.10) has as unique invertible solution the Fréchet mean of $\Sigma_1, \dots, \Sigma_n$, which turns out to be Gaussian as well. In Masarotto et al. [2018] we proved that part of their results extend easily to infinite dimensions. The infinite-dimensional extension is presented in Proposition 16 in the next chapter.

2 Optimal transport of Gaussian processes

Functional datasets are becoming more and more common, and an increasing amount of literature focusses on dealing with such data [Ramsay and Silverman, 2005b, Hsing and Eubank, 2015, Horvath and Kokoszka, 2012]. Many of these contributions have dealt with the problem of the estimation of, and inference for, a mean function. In this Chapter, we present a collection of results concerning especially the covariance structure of such data. This Chapter contains most of the theoretical contributions of the thesis.

We begin by establishing the key connection between the Procrustes metric for operators and the Wasserstein metric for centered Gaussian processes (Section 2.1). We then show how this will lead to new results related to existence, uniqueness and stability of Fréchet mean, which are presented in Section 2.2. Section 2.3 gives details on the practical implementations of an algorithm to compute Fréchet means, while Section 2.4 describes a “canonical” generative model for functional data giving rise to the Wasserstein–Procrustes metric.

2.1 Equivalence of measures

We set off by establishing the equivalence between the Procrustes and the Wasserstein metric. This connection will allows us to describe the geometry and derive key properties of the space of covariance operators.

Let $\mu \equiv N(m_1, \Sigma_1)$ and $\nu \equiv N(m_2, \Sigma_2)$ be Gaussian measures (in the sense of Section 1.4.1). Following Pigoli et al. [2014a] we can write the Procrustes distance between Σ_1 and Σ_2 of Equation (1.12) as

$$\begin{aligned}\Pi^2(\Sigma_1, \Sigma_2) = & \inf_{R\,:\,R^*R=\mathscr{I}} \mathrm{tr}[(\sqrt{\Sigma_1} - \sqrt{\Sigma_2}\mathrm{R})^*(\sqrt{\Sigma_1} - \sqrt{\Sigma_2}\mathrm{R})] \\ = & \mathrm{tr}\Sigma_1 + \mathrm{tr}\Sigma_2 - 2 \sup_{\mathrm{R}\,:\,\mathrm{R}^*\mathrm{R}=\mathscr{I}} \mathrm{tr}(\mathrm{R}^*\Sigma_2^{1/2}\Sigma_1^{1/2}). \end{aligned} \tag{2.1}$$

Now consider the non-negative product

$$C = [\Sigma_2^{1/2}\Sigma_1^{1/2}]^*\Sigma_2^{1/2}\Sigma_1^{1/2} = \Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}$$

and write the singular value decomposition of $\Sigma_2^{1/2}\Sigma_1^{1/2}$ as $\Sigma_2^{1/2}\Sigma_1^{1/2} = UC^{1/2}V$, for unitary U and V. The expression for the Procrustes distance in (2.1) becomes

$$\Pi^2(\Sigma_1,\Sigma_2) = \mathrm{tr}\Sigma_1 + \mathrm{tr}\Sigma_2 - 2 \sup_{\mathrm{R}\,:\,\mathrm{R}^*\mathrm{R}=\mathfrak{I}} \mathrm{tr}(\mathrm{VR}^*\mathrm{UC}^{1/2}). \tag{2.2}$$

By observing that $\{VR^*U : R^*R = \mathfrak{I}\}$ is just an alternative definition of the set of unitary operators, it follows that the Procrustes distance in (2.2) is maximised when VR^*U is the identity, and (2.2) becomes

$$\Pi^2(\Sigma_1,\Sigma_2) = \mathrm{tr}(\Sigma_1) + \mathrm{tr}(\Sigma_2) - 2\mathrm{tr}\left[\sqrt{\Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2}}\right],$$

which is exactly the expression of the Wasserstein distance between μ and ν given in Equation (1.16).

We have just given a proof of the following theorem, which connects the theory in Sections 1.3.1 and 1.2.5.

Theorem 7. *The Procrustes distance between two trace-class covariance operators Σ_1 and Σ_2 on $\mathscr{H}$ coincides with the Wasserstein distance between two second-order Gaussian processes $N(0,\Sigma_1)$ and $N(0,\Sigma_2)$ on $\mathscr{H}$,*

$$\begin{aligned}\Pi(\Sigma_1,\Sigma_2) &= \inf_{R:R^*R=\mathfrak{I}} \|\Sigma_1^{1/2} - \Sigma_2^{1/2}R\|_2 \\ &= \sqrt{\mathrm{tr}(\Sigma_1) + \mathrm{tr}(\Sigma_2) - 2\mathrm{tr}\sqrt{\Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2}}} = W(N(0,\Sigma_1),N(0,\Sigma_2)).\end{aligned}$$

It is worth noticing that in finite dimensions, Bhatia et al. [2018] obtain the same conclusion of Theorem 7 as a variational principle. Furthermore, Bhatia et al. [2018] make a connection to quantum information, where a related quantity to the Wasserstein distance is known as the *Bures* distance [Bures, 1969]. For Gaussian measures, the 2-Wasserstein distance coincides with the Bures distance.
In view of Theorem 7, we will sometimes call the Procrustes metric the Procrustes–Wasserstein metric.

Commuting operators

A noteworthy simplification for the expression that the Wasserstein distance takes place when the operator Σ_1 and Σ_2 commute, i.e., $\Sigma_1\Sigma_2 = \Sigma_2\Sigma_1$.

Let $\mathbf{t}_{\Sigma_1}^{\Sigma_2}$ be the optimal transport map between Σ_1 and Σ_2. Assume that $\ker\Sigma_1 \subseteq \ker\Sigma_2$, to guarantee that $\mathbf{t}_{\Sigma_1}^{\Sigma_2}$ stays well defined. In the case of commuting operators, $\mathbf{t}_{\Sigma_1}^{\Sigma_2}$ simplifies to $\Sigma_2^{1/2}\Sigma_1^{-1/2}$ and the Wasserstein distance between μ and ν reduces to the Hilbert–Schmidt distance between the square roots of the covariance operators. To see this, note first that in this case $\Sigma_1^{1/2}\Sigma_2^{1/2}$ is self-adjoint. Then

$$\begin{aligned}
W^2(N(0,\Sigma_1),N(0,\Sigma_2)) &= \mathrm{tr}(\Sigma_1)+\mathrm{tr}(\Sigma_2)-2\mathrm{tr}\sqrt{\Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2}}\\
&= \mathrm{tr}(\Sigma_1)+\mathrm{tr}(\Sigma_2)-2\mathrm{tr}\sqrt{(\Sigma_1^{1/2}\Sigma_2^{1/2})^*(\Sigma_1^{1/2}\Sigma_2^{1/2})}\\
&= \mathrm{tr}(\Sigma_1)+\mathrm{tr}(\Sigma_2)-2\mathrm{tr}(\Sigma_1^{1/2}\Sigma_2^{1/2})\\
&= \|\Sigma_1^{1/2}\|_2^2+\|\Sigma_2^{1/2}\|_2^2-2\langle\Sigma_1^{1/2},\Sigma_2^{1/2}\rangle_{HS} = \|\Sigma_1^{1/2}-\Sigma_2^{1/2}\|_2^2.
\end{aligned}$$

We can now move and characterise the topology and the geometry of the Gauss-Wasserstein space and the behaviour of the Procrustes–Wasserstein under finite-rank projection.

2.1.1 Topological properties

The topology of Gaussian measures under the Wasserstein metric can be deduced by means of the fact that Gaussian measures converge weakly to a Gaussian measure if and only if their moments also converge. The next paragraphs report the topological properties of the Wasserstein distance between Gaussian measures. For a proof see [Zemel, 2017] and [Panaretos and Zemel, To appear].

In order to state the result, we need to recall some basic notions of analysis in Polish spaces. Recall that a Polish space is a separable completely metrisable topological space.

Definition 8. *A sequence $\mu_n \subset P(\mathscr{H})$ converges* narrowly *(or* in distribution*) to a measure μ if*

$$\int f d\mu_n \mapsto \int f d\mu,$$

for all bounded and continuous functions $f : \mathscr{H} \to \mathbb{R}$.

We refer to Villani [2008] for a proof that the topology of narrow convergence is metrizable.

Definition 9. *A set $K \subset P(\mathscr{H})$ is called* tight *provided that for every $\varepsilon > 0$ there exists a compact set $K_\varepsilon \subset \mathscr{H}$ such that*

$$\mu(\mathscr{H}\backslash K_\varepsilon) \le \varepsilon, \quad \forall \mu \in K.$$

Now we are ready to characterise the topology induced by the Wasserstein distance. Let $\{\Sigma_n\}_{n=1}^{\infty}$, Σ be covariance operators on $\mathscr{H}$. Then the following statements are equivalent:

1. $N(0,\Sigma_n) \overset{n\to\infty}{\longrightarrow} N(0,\Sigma)$ in distribution.
2. $\Pi(\Sigma_n,\Sigma) \overset{n\to\infty}{\longrightarrow} 0$.

3. $\|\sqrt{\Sigma_n} - \sqrt{\Sigma}\|_2 \overset{n\to\infty}{\longrightarrow} 0$.
4. $\|\Sigma_n - \Sigma\|_1 \overset{n\to\infty}{\longrightarrow} 0$.

More is known about the topology of Wasserstein space, for instance, that $\mathscr{W}(\mathscr{H})$ is homeomorphic to an infinite-dimensional convex subset of a Hilbert space $\mathscr{L}_2(\mu)$ (for any regular measure μ) and has therefore a manifold-like structure inherited from $L_2(\mu)$. The reader is referred to [Zemel, 2017, Lemmas 3.4.4 and 3.4.5] and to [Panaretos and Zemel, To appear] for a more in-depth treatment of this subject.

2.1.2 Finite-rank approximations

Any real-life statistical inference deals with finite-dimensional approximations to their functional counterpart. Therefore, in order to use the Procrustes metric in practice, we need some sort of stability of the distances across finer finite-dimensional approximations. In the following Lemma we prove a result on stability of the distance once the operators are acted upon by any sequence of projections (non-necessarily finite dimensional) converging strongly to the identity in $\mathscr{H}$. We say that $\mathcal{P}$ is a projection operator if $\mathcal{P}^* = \mathcal{P} = \mathcal{P}^2$, while we say that a sequence of operators T_n converges to T *strongly* if $T_n x \to Tx$ for all $x \in \mathscr{H}$ (Stein and Shakarchi [2009, p. 198])[1]. We can now state:

Lemma 10. *Let* Σ *a covariance operator and* $\mathcal{P}_n$ *be a sequence of projections that converges strongly to the identity. Then* $\Pi(\Sigma, \mathcal{P}_n\Sigma\mathcal{P}_n) \to 0$.

The Lemma was given in Masarotto et al. [2018]. Stability of the distance between finite-dimensional projections of two infinite operators was already considered in Pigoli et al. [2014a], but their result is a special case of Lemma 10: in the notation of Lemma, simply take $\mathcal{P}_n = \sum_{j=1}^n e_j \otimes e_j$ as the projection onto the span of $\{e_1, \dots, e_n\}$, for $\{e_j\}_{j\geq 1}$ an orthonormal basis of $\mathscr{H}$.

In order to give a proof of Lemma 10 we need the following result (Masarotto et al. [2018, Lemma 19])

Lemma 11. *An operator* $A \in \mathscr{L}_2(N(0,\Sigma))$*, possibly unbounded, is optimal from* $N(0,\Sigma)$ *to* $A\#N(0,\Sigma)$ *if and only if* A *is non-negative, and then* $A\#N(0,\Sigma) = N(0, A\Sigma A)$.

Proof of Lemma 10. Let $\mu \in \mathscr{W}(\mathscr{H})$ Gaussian with covariance Σ and $\mathcal{P}$ be a projection operator. By Lemma 11, $\mathcal{P}$ is an optimal map from μ to $\mathcal{P}\#\mu$, and if $\mathcal{P}\Sigma\mathcal{P}$ is the projection of Σ onto the range of $\mathcal{P}$, we have

$$\Pi^2(\Sigma, \mathcal{P}\Sigma\mathcal{P}) = W^2(\mu, \mathcal{P}\#\mu) = \int_{\mathscr{H}} \|x - \mathcal{P}x\|^2 \, \mathrm{d}\mu(x) = \operatorname{tr}\left\{(\mathcal{I}-\mathcal{P})\Sigma(\mathcal{I}-\mathcal{P})\right\} = \operatorname{tr}\left\{(\mathcal{I}-\mathcal{P})\Sigma\right\}.$$

---

[1] This is much weaker than convergence in operator norm, but stronger than requiring that $\langle T_n x, y\rangle \to \langle Tx, y\rangle$ for all $x, y \in H$, which is called *weak* convergence of T_n to T.

Now since $\mathcal{P}$ converges strongly to the identity, $\mathcal{P}_n x \to x$ for all x and $\|\mathcal{P}_n x\| \leq \|x\|$, so the dominated convergence theorem yields that $\mathcal{P}_n \# \mu \to \mu$ in $\mathcal{W}(\mathcal{H})$, completing the proof. □

Lemma 10 can be actually stated for every $\mu \in \mathcal{W}(\mathcal{H})$, in which case $\mathcal{P}_n \# \mu \to \mu$ in $\mathcal{W}(\mathcal{H})$. When the projections are finite-dimensional, the result of Lemma 10 can be strengthened, as in this case convergence of the finite-dimensional distance will converge uniformly over compacta to its infinite-dimensional counterpart (Masarotto et al. [2018, Proposition 6]). We report it here for completeness.

Proposition 12. *Let $\{e_j\}_{j\geq 1}$ be an orthonormal basis of $\mathcal{H}$ and $\mathcal{P}_n = \sum_{j=1}^n e_j \otimes e_j$ be the projection on the span of $\{e_1, \dots, e_n\}$. Let $\mathcal{B}$ be a collection of non-negative bounded operators satisfying*

$$\sup_{\Sigma \in \mathcal{B}} \sum_{j=n+1}^{\infty} \left\langle \Sigma e_j, e_j \right\rangle \to 0, \qquad n \to \infty. \tag{2.3}$$

Then,

$$\sup_{\Sigma_1, \Sigma_2 \in \mathcal{B}} |\Pi(\mathcal{P}_n \Sigma_1 \mathcal{P}_n, \mathcal{P}_n \Sigma_2 \mathcal{P}_n) - \Pi(\Sigma_1, \Sigma_2)| \to 0, \qquad n \to \infty.$$

Proof. Let $\mathcal{K} \subset \mathcal{W}(\mathcal{H})$ be a collection of measures with $\Sigma(\mu) \in \mathcal{B}$ for all $\mu \in \mathcal{K}$. It suffices to show that $W(\mu, \mathcal{P}_n \# \mu) \to 0$ uniformly and indeed

$$W^2(\mu, \mathcal{P}_n \# \mu) = \mathrm{tr}(\mathcal{I} - \mathcal{P}_\mathrm{n})\Sigma(\mu) = \sum_{\mathrm{j=n+1}}^{\infty} \left\langle \Sigma(\mu)\mathrm{e_j}, \mathrm{e_j} \right\rangle$$

vanishes uniformly as $n \to \infty$. □

2.1.3 Geometry of the Gauss–Wasserstein space

Thanks to theorem 7, the geometrical characterisation of the Wasserstein space given in Section 1.4.2 applies readily to the space of covariance operators (identified with the space of centered Gaussian processes) equipped with the Procrustes–Wasserstein metric Π. We translate the results for the space of operators in this section.

Thanks to Lemma 11, the optimal map $\mathbf{t}$ is the unique non-negative, possibly unbounded operator such that $\mathbf{t}\Sigma\mathbf{t}$ is trace-class. Recalling that trace-class operators can be seen as "squares" of Hilbert-Schmidt operators, we have that $\Sigma^{1/2}\mathbf{t}$ is Hilbert–Schmidt and thus so is $\Sigma^{1/2}(\mathbf{t} - \mathcal{I})$.
Using the simplified notation Tan_Σ for $\mathrm{Tan}_{N(0,\Sigma)}$, the inner product on Tan_Σ is defined as

$$\langle A, B \rangle_{\mathrm{Tan}_\Sigma} = \int_{\mathcal{H}} \langle Ax, Bx \rangle \, \mathrm{d}\mu(x) = \mathrm{tr(A\Sigma B)} = \mathbb{E}[\langle \mathrm{AX, BX} \rangle], \qquad \mathrm{X} \sim \mu \equiv \mathrm{N}(0, \Sigma). \tag{2.4}$$

We can now write the following description for the tangent space Tan_Σ,

$$\mathrm{Tan}_\Sigma = \overline{\left\{t(S-\mathscr{I}) : t > 0,\ S \geq 0,\ \|\Sigma^{1/2}(S-\mathscr{I})\|_2 < \infty\right\}}, \tag{2.5}$$

where the closure is taken with respect to $\langle,\rangle_{\mathrm{Tan}_\Sigma}$. Now, observe that if Q is a bounded self adjoint operator, and $t > \|Q\|_\infty$, then $S = \mathscr{I} + Q/t$ is non-negative. In this we can approximate unbounded Q's. A self-adjoint operator $Q \in \mathrm{Tan}_\Sigma$ has zero norm if and only if $Q\Sigma Q = 0$ or, equivalently, $Q\Sigma = 0$. Thus Tan_Σ can be turned into an Hilbert space of equivalent classes of operators Q modulo the relation $Q \sim Q' \iff (Q-Q')\Sigma = 0$ [2].
Equation (2.5) can be rewritten as

$$\mathrm{Tan}_\Sigma = \overline{\left\{Q : Q = Q^*,\ \|\Sigma^{1/2}Q\|_2 < \infty\right\}}.$$

We can also give expressions for the exponential and the log maps on Tan_Σ, the first being

$$\exp_\Sigma(A) = (A+\mathscr{I})\Sigma(A+\mathscr{I}).$$

For the latter to be well defined, we need one of the following regularity conditions to hold: $\ker(\Sigma_0) \subseteq \ker(\Sigma_1)$ or $\overline{\mathrm{range}(\Sigma_0)} \supseteq \overline{\mathrm{range}(\Sigma_1)}$. If $\mu \equiv N(0,\Sigma_0)$ and $\nu \equiv N(0,\Sigma_1)$ and using Proposition 4, fulfilment of the above conditions will yield sufficiency for the existence of

1. the log map of Σ_1 at Σ_0,

   $$\log_{\Sigma_0}\Sigma_1 = \mathbf{t}_0^1 - \mathscr{I} = \Sigma_0^{-1/2}(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2})^{1/2}\Sigma_0^{-1/2} - \mathscr{I},$$

   and defined $N(0,\Sigma)$-almost surely;

2. a unique (unit speed) geodesic from Σ_0 to Σ_1,

   $$\Sigma_t = [t\mathbf{t}_0^1 + (1-t)\mathscr{I}]\Sigma_0[t\mathbf{t}_0^1 + (1-t)\mathscr{I}] = t^2\Sigma_1 + (1-t)^2\Sigma_0 + t(1-t)[\mathbf{t}_0^1\Sigma_0 + \Sigma_0\mathbf{t}_0^1],$$

   where again $\mathbf{t}_0^1 = \Sigma_0^{-1/2}(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2})^{1/2}\Sigma_0^{-1/2}$.

2.2 Fréchet means of covariance operators

This section provides the first significant theoretical contribution of this thesis, going beyond Pigoli et al. [2014a]. It treats the empirical Fréchet mean of a collection of covariance operators $\Sigma_1, \ldots, \Sigma_n$ with respect to the Procrustes–Wasserstein metric. We recall that the empirical Fréchet mean is given by the minimiser of the Fréchet functional

$$F(\Sigma) = \frac{1}{2N}\sum_{i=1}^{N}\Pi^2(\Sigma,\Sigma_i) = \frac{1}{2N}\sum_{i=1}^{N}W^2(N(0,\Sigma),N(0,\Sigma_i)).$$

---

[2] In this way Tan_Σ contains all equivalent classes of bounded self-adjoint operators on $\mathscr{H}$, but also certain unbounded ones. For example, if $\Sigma^{1/3}$ is trace-class, then the tangent space inner product is well defined when taking $A = B = \Sigma^{-1/3}$, even though the latter is an unbounded operator.

Thanks to Theorem 7, the minimiser of the Fréchet functional F will also yield the barycenter of the corresponding Gaussian measures $\mu_1 \equiv \mathscr{N}(0,\Sigma_1),\dots,\mu_k \equiv \mathscr{N}(0,\Sigma_1)$.

For a collection of Gaussian measures, their Fréchet mean will still be Gaussian, so there is no need to minimize $F(\Sigma)$ over the full set of measures in $\mathscr{W}(\mathscr{H})$. See Section 1.5.1 for details.

The population Fréchet mean will minimize

$$F(\Sigma) = \frac{1}{2}\mathbb{E}\Pi^2(\Sigma,\mathscr{Y}),$$

for a random covariance operator $\mathscr{Y}$.
We recall (cf. Section 1.5.1) that existence and uniqueness of Fréchet means in general metric spaces are not granted, but we will give relatively weak conditions under which they can be established for Gaussian measures in Wasserstein spaces. In finite dimensions such conditions can be found in Agueh and Carlier [2011], but there is no straightforward counterpart for operators.

Results on existence of the Fréchet mean rely on the concept of *multicouplings*, which are an extension of the concept of coupling of measures (see Section 1.3) and are defined as follows.

Definition 13 (Multicoupling)**.** *Let* $\mu_1,\dots,\mu_N \in \mathscr{W}(\mathscr{H})$. *A multicoupling of* $(\mu_1,\dots,\mu_N)$ *is a Borel measure on* $\mathscr{H}^N$ *with marginals* $\mu_1,\dots,\mu_N$.

Multicouplings were studied in connection with the *multimarginal* Monge–Kantorovich problem, which is solved by minimising the functional

$$G(\pi) = \frac{1}{2N^2}\int_{\mathscr{H}^N}\sum_{i<j}\|x_i - x_j\|^2\,\mathrm{d}\pi(x_1,\dots,x_N) = \int_{\mathscr{H}^N}\frac{1}{2N}\sum_{i=1}^{N}\|x_i - \overline{x}\|^2\,\mathrm{d}\pi(x). \tag{2.6}$$

A multicoupling π is optimal if it is a minimiser of $G(\pi)$. See Gangbo and Święch [1998] and Zemel and Panaretos [2017] for a complete study of optimal multicouplings in $\mathbb{R}^d$.

When $N = 2$ we retrieve the Kantorovich problem of Section 1.3, and just as in $N = 2$ an optimal multicoupling π always exists. Moreover, if μ_1 is regular, an optimal multicoupling of $\mu^1,\dots,\mu^N \in \mathscr{W}$ is given by $(\mathfrak{I}, S_2,\dots,S_N)\#\mu_1$ for some functions $S_i : \mathbb{R}^d \to \mathbb{R}^d$, where

$$(\mathfrak{I}, S_2,\dots,S_N)\#\mu_1(B_1\times\ldots\times B_N) = \mu_1(\{x\in B_1 : S_2(x)\in B_2,\dots,S_N(x)\in B_N\}) = \mu_1\left(\bigcap_{i=1}^{N} S_i^{-1}(B_i)\right)$$

for any Borel-rectangle $B_1\times\ldots\times B_N$, and $S_1 = \mathfrak{I}$. Lemma 14 links the optimal multicoupling of the collection $\mu^1,\dots,\mu^N \in \mathscr{W}$ with the existence of their Fréchet mean. The result was originally proven in $\mathbb{R}^d$ by Agueh and Carlier [2011, Proposition 4.2], and here it is extended to infinite dimension.

Lemma 14 (Fréchet means and multicouplings)**.** *Let* $\mu^1,\dots,\mu^N \in \mathscr{W}$. *Then* μ *is a Fréchet mean of* $(\mu^1,\dots,\mu^N)$ *if and only if there exists an optimal multicoupling* $\pi \in \mathscr{W}(\mathscr{H}^N)$ *of* $(\mu^1,\dots,\mu^N)$

such that

$$\mu = M_N\#\pi, \qquad M_N : \mathscr{H}^N \to \mathscr{H}, \qquad M_N(x_1,\dots,x_N) = \overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

Proof. Let π be an arbitrary multicoupling of $(\mu^1,\dots,\mu^N)$ and set $\mu = M_N\#\pi$. The measures μ^i and μ are coupled by $(x \mapsto x_i, M_N)\#\pi$, so

$$\int_{\mathscr{H}^N} \|x_i - M_N(x)\|^2 \, \mathrm{d}\pi(x) \geq W^2(\mu,\mu_i).$$

Let $F(\mu)$ be the Fréchet functional and $G(\mu)$ be the multicoupling functional in (2.6). By summing over i in the above equation, we get $G(\pi) \geq F(\mu)$ and so $\inf G \geq \inf F$.

For the other direction, let $\mu \in \mathscr{W}$ be arbitrary. For each i, let π^i be an optimal coupling between μ and μ^i. Since all π^i share a common marginal μ, by the gluing lemma (Ambrosio and Gigli [2013, Lemma 2.1]) we can construct a measure η on $\mathscr{H}^{N+1}$ with marginals $\mu_1,\dots,\mu_N,\mu$ and such that projection π of η on $\mathscr{H}^N$ is a multicoupling of $\mu_1,\dots,\mu_N$.

Since $\mathscr{H}$ is a Hilbert space, the minimiser of $y \mapsto \sum \|x_i - y\|^2$ is $y = M_N(x)$. Therefore

$$F(\mu) = \frac{1}{2N}\int_{\mathscr{H}^{N+1}} \sum_{i=1}^{N} \|x_i - y\|^2 \, \mathrm{d}\eta(x,y) \geq \frac{1}{2N}\int_{\mathscr{H}^{N+1}} \sum_{i=1}^{N} \|x_i - M_N(x)\|^2 \, \mathrm{d}\eta(x,y) = G(\pi), \quad (2.7)$$

from which it follows that $\inf F \geq \inf G$. Combining the two inequalities, we get the correspondance between multicoupling and Fréchet mean, i.e., $\inf F = \inf G$.
To conclude the proof, we turn our attention to inequality (2.7) above. This inequality holds as equality if and only if $y = M_N(x)$ η-almost surely, which, when true, implies that $\mu = M_N\#\pi$. If this last condition fails and μ does not coincide with $M_N\#\pi$, then it cannot be optimal, as in this case, $F(\mu) > G(\pi) \geq F(M_N\#\pi)$. Finally, if π is optimal, then

$$F(M_N\#\pi) \leq G(\pi) = \inf G = \inf F$$

which completes the proof, yielding the required optimality of $\mu = M_N\#\pi$. □

Now that existence is established we can deal with uniqueness. A Gaussian measure μ with covariance Σ is regular if and only if Σ is injective. We will show that uniqueness is guaranteed if at least one of the measures is regular and we will use the fact that, if this regularity condition holds, the optimal map $\mathbf{t}_\mu^\nu$ exists for any $\nu \in \mathscr{W}(\mathscr{H})$ (including the cases with non-Gaussian ν).

Uniqueness of the Fréchet mean follows from the strict convexity of the Fréchet functional F. In this case, convexity is not intended with respect to the Wasserstein geometry, but rather with respect to set-wise addition (also known as mixing, in statistics): given probability measures μ and ν, we can consider the function defining by assigning to a Borel set A a measure $t\mu(A) + (1-t)\nu(A)$. Such functions define a probability measure $t\mu + (1-t)\nu$ for all

$t \in [0,1]$.

Proposition 15. *Let $\mu_1, \dots, \mu_N \in \mathscr{W}(\mathscr{H})$ and assume that μ_1 is regular. Then the Fréchet functional is strictly convex, and the Fréchet mean of $\mu_1, \dots, \mu_N$ is unique. In particular, the Fréchet mean of a collection of covariance operators is unique if at least one of the operators is injective.*

Proof. Let $\nu_1, \nu_2, \mu \in \mathscr{W}(\mathscr{H})$ and let π_i be an optimal coupling of ν_i and μ. For any $t \in (0,1)$ the linear interpolant $t\pi_1 + (1-t)\pi_2$ is a coupling of $t\nu_1 + (1-t)\nu_2$ and μ. We have

$$W^2(t\nu_1 + (1-t)\nu_2, \mu) \leq \int_{\mathscr{H}^2} \|x-y\|^2 \, \mathrm{d}[t\pi_1 + (1-t)\pi_2](x,y) = tW^2(\nu_1,\mu) + (1-t)W^2(\nu_2,\mu). \tag{2.8}$$

Equation (2.8) implies that the squared Wasserstein distance is weakly convex.
Notice that if the inequality was strict, it would lead to the strict convexity of the Fréchet functional, and therefore to the proof of the proposition, since in this case the Fréchet functional would be a sum of N squared Wasserstein distances that are all convex, and with one of them strictly convex.
To prove strict convexity, we procede by contradiction, invoking Theorem 2.9 in [Álvarez-Esteban et al., 2011], which holds true independently on dimension [3]. Observe first that for a regular μ, both π_1, π_2 are induced by well-defined maps $T_i = \mathbf{t}_\mu^{\nu_i}$. Now assume $\nu_1 \neq \nu_2$. Then by Álvarez-Esteban et al. [2011, Theorem 2.9] $t\pi_1 + (1-t)\pi_2$ cannot be the optimal coupling of $t\nu_1 + (1-t)\nu_2$ and μ, since there is no map T inducing $t\pi_1 + (1-t)\pi_2$ as some $(X, T(X))$ with marginals μ and $t\nu_1 + (1-t)\nu_2$. In particular, the inequality above is strict and so is the convexity of $W^2(\cdot, \mu)$. □

Uniqueness holds also at the population level for finite matrices, as long as the random covariance operator is injective with positive probability. Uniqueness for population Fréchet mean follows from an idea of Álvarez-Esteban et al. [2011], and it was proved in $\mathbb{R}^d$ by Bigot and Klein [2012] in a parametric setting, and Zemel and Panaretos [2017] in the non-parametric one.

We can now move to the Fréchet mean of Gaussian measures, specifically showing that a Gaussian–Fréchet mean is Gaussian as well.

2.2.1 Fréchet mean of Gaussian measures

Agueh and Carlier [2011, Section 6.3] showed that in $\mathbb{R}^d$ the Fréchet mean of Gaussian measures is Gaussian. Making use of the stability result of Section 2.2.2, we now extend their result to infinite dimensions.

---

[3] In our case $P_1 = P_2$ (in the notation of Álvarez-Esteban et al. [2011]), hence there is no need to work with densities.

Let $\Sigma_1,\dots,\Sigma_n$ be covariance operators with unique Fréchet mean $\bar{\Sigma}$. Let $\mathcal{P}_k$ be a sequence of finite-dimensional projections converging strongly to the identity. Let $\bar{\Sigma}^{(k)}$ denote the Fréchet mean of $(\mathcal{P}_k\Sigma_i\mathcal{P}_k)_{i=1}^n$. Following the reasoning in Agueh and Carlier [2011, Theorem 6.1], Brower's fixed-point theorem (e.g. [Karamardian, 2014]) implies that there exists a positive definite solution to the equation

$$F^{(k)}(\Sigma)=\Sigma, \tag{2.9}$$

where $F^{(k)}(\Sigma)$ is the functional

$$\sum_{i=1}^n \left(\Sigma^{1/2}(\mathcal{P}_k\Sigma_i\mathcal{P}_k)\Sigma^{1/2}\right)^{1/2}.$$

Now let $\bar{\Sigma}^{(k)}$ be a $k-$ranked positive definite solution of (2.9) and define $\bar{\mu}^{(k)}=\mathcal{N}(0,\bar{\Sigma}^{(k)})$. The optimal transport map between $\bar{\mu}^{(k)}$ and $\mu_i^{(k)}\equiv\mathcal{N}(0,(\mathcal{P}_k\Sigma_i\mathcal{P}_k))$ is the linear map

$$T_i^{(k)}=(\mathcal{P}_k\Sigma_i\mathcal{P}_k)^{1/2}\left((\mathcal{P}_k\Sigma_i\mathcal{P}_k)^{1/2}\bar{\Sigma}^{(k)}(\mathcal{P}_k\Sigma_i\mathcal{P}_k)^{1/2}\right)^{-1/2}(\mathcal{P}_k\Sigma_i\mathcal{P}_k)^{1/2}.$$

Always following Agueh and Carlier [2011], since $\sum_{i=1}^n T_i^{(k)}=\mathcal{I}$, $\bar{\Sigma}^{(k)}$ is a barycenter of $(\mathcal{P}_k\Sigma_i\mathcal{P}_k)_{i=1}^n$, making it the unique positive definite barycenter. Now notice that we have a sequence of Gaussian barycenters that, by the stability result in Section 2.2.2 and as k grows to infinity, must converge to the barycenter of the limit. This will give a sequence of mean-zero Gaussian measures with covariance operators $(\mathcal{P}_k\Sigma_i\mathcal{P}_k)$ which is convergent in Wasserstein norm, and by weak convergence of Gaussian measures the limit must be Gaussian. Agueh and Carlier [2011] make use of the invertibility of $\bar{\Sigma}_k^{1/2}$, which is not given for a finite-rank approximation to an infinite-dimensional operator. However, the range of $\bar{\Sigma}_k$ is completely contained in range$(\mathcal{P}_k)$, which is a finite-dimensional space, justifying the result.

In Chapter 1 it was mentioned that the Fréchet mean of Gaussian covariances can be found as a positive-definite solution Σ of the implicit equation [Knott and Smith, 1984, Rüschendorf and Uckelmann, 2002]

$$\Sigma=\frac{1}{n}\sum_{i=1}^n(\Sigma^{1/2}\Sigma_i\Sigma^{1/2})^{1/2}. \tag{2.10}$$

The result was extended by Agueh and Carlier [2011], who established existence and uniqueness, and indeed that the Fréchet mean of Gaussian measure is Gaussian. In Masarotto et al. [2018] we extended part of their results to infinite dimensions, as illustrated by the next Proposition.

Proposition 16. *Let* $\Sigma_1,\dots,\Sigma_n$ *be covariance operators. Then:*

1. *Any Fréchet mean* $\bar{\Sigma}$ *of* $\Sigma_1,\dots,\Sigma_n$ *satisfies* (2.10).
2. *If* (2.10) *holds, and* $\ker(\Sigma)\subseteq\bigcap_{i=1}^n\ker(\Sigma_i)$, *then* Σ *is a Fréchet mean.*

Proof. Let $\mathcal{P}_k$ be a sequence of finite-rank projections converging strongly to the identity $\mathcal{I}$. Then $\mathcal{P}_k\Sigma_i\mathcal{P}_k$ converges to Σ_i in Wasserstein distance (Section 2.1.2). The results in Section 2.1.1 thus imply that the convergence is also in trace norm. Now call $\overline{\Sigma}_k$ the Fréchet mean of the projected operators. If we replace Σ_i by $\mathcal{P}_k\Sigma_i\mathcal{P}_k$ in Agueh and Carlier [2011, Theorem 6.1], it follows that $\overline{\Sigma}_k$ satisfies (2.10). But $\overline{\Sigma}_k$ converge to $\overline{\Sigma}$ in trace norm due to the results in Section 2.1.1 and the stability Theorem 18 in the next Chapter. Therefore (2.10) holds true for $\overline{\Sigma}$ by continuity.

Let us now prove the other direction. From the conditions on the kernels it follows that $T_i = \mathbf{t}_{\Sigma}^{\Sigma_i} = \Sigma^{-1/2}(\Sigma^{1/2}\Sigma_i\Sigma^{1/2})^{1/2}\Sigma^{-1/2}$ exists and is defined on a dense subspace D_i of Σ-measure one (Proposition 4). Now $D = \bigcap_{i=1}^{n} D_i$ is a set of full measure and Equation (2.10) yields $\sum T_i = n\mathcal{I}$ on D_i. Theorem 21 implies that Σ is a Fréchet mean. $\square$

Lemma 14 shows that we can deduce the Fréchet mean of a collection of Gaussian measures from an optimal multicoupling. The following result from Zemel and Panaretos [2017] states that the proof of Lemma 14 can be used to prove the opposite direction, i.e., that one can deduce an optimal multicoupling from the Fréchet mean.

Lemma 17. *Let* $\Sigma_1,\dots,\Sigma_n$ *be covariances on* $\mathscr{H}$ *with injective Fréchet mean* $\overline{\Sigma}$. *Let* $Z \sim N(0,\overline{\Sigma})$ *and define a random Gaussian vector on* $\mathscr{H}^n$ *by*

$$(Y_1,\dots,Y_n), \qquad Y_i = \mathbf{t}_{\overline{\Sigma}}^{\Sigma_i}(Z) = \overline{\Sigma}^{-1/2}[\overline{\Sigma}^{1/2}\Sigma_i\overline{\Sigma}^{1/2}]^{1/2}\overline{\Sigma}^{-1/2}Z, \qquad i = 1,\dots,n.$$

Then the joint law of $(Y_1,\dots,Y_n)$ *is an optimal multicoupling of* $N(0,\Sigma_1),\dots,N(0,\Sigma_n)$.

2.2.2 Stability of the Fréchet mean

Much of the theory behind functional data analysis deals with continuously sampled data. In practice however, the observed data are most likely discretely sampled and only provide a finite-dimensional approximation to the infinite-dimensional objects. Moreover, the approximation is often hindered by the use of some kind of smoothing techniques for noise reduction (e.g. Yao et al. [2005a,b] or Descary [2017]). For this reason it is important to ascertain that the relevant inference remains stable across finer and finer approximations. In Section 2.1.2 we verified the stability of the Procrustes distance for progressively more refined finite rank approximations. The topological knowledge of the Wasserstein space makes it possible to deduce a similar conclusion for the stability of the Fréchet mean.

The next theorem is phrased in terms of covariance operators. It could be equivalently expressed in terms of Gaussian measures $\{N(0,\Sigma^i) : i \le N\}$ and finite rank sequences $\{N(0,\Sigma_k^i) : i \le N,\ k \ge 1\}$, and in terms of Wasserstein barycenters rather than Fréchet means.

Theorem 18 (Fréchet means and projections)**.** *Let* $\Sigma^1,\dots,\Sigma^N$ *be covariance operators with* Σ^1 *injective, and let* $\{\Sigma_k^i : i \le N,\ k \ge 1\}$ *be sequences such that* $\Sigma_k^i \xrightarrow{k\to\infty} \Sigma^i$ *in trace norm (equivalently,* *in Procrustes distance). Then the Fréchet mean of* $\Sigma_k^1,\dots,\Sigma_k^N$ *converges in trace norm to that of* $\Sigma^1,\dots,\Sigma^N$.

Proof. Call $\{\mu^i : i \le N\}$ and $\{\mu_k^i : i \le N,\ k \ge 1\}$ the sequences of centered Gaussian measures with corresponding covariance operators, i.e. $\{N(0,\Sigma^i) : i \le N\}$ and $\{N(0,\Sigma_k^i) : i \le N,\ k \ge 1\}$ and let F_k and F denote respectively the finite- and infinite-dimensional Fréchet functionals. The proof will develop by proving the following steps:

1. $(\overline{\mu}_k)$ is tight. From tightness paired with Gaussianity we will deduce that $(\overline{\mu}_k)$ is pre-compact in $\mathscr{W}(\mathscr{H})$.
2. Each of the limits of $(\overline{\mu}_k)$ is a minimiser of F.
3. There is only one minimiser of F. Therefore $\overline{\mu}_k$ must converge to the minimiser of F.

Step 1: tightness of $(\overline{\mu}_k)$. The equivalent results in Section 2.1.1 provide tightness of the entire collection $\mathscr{K} = \{\mu_k^i\}$, since all the sequences converge in distribution. For any $\epsilon > 0$, there exists a compact $K_\epsilon \subset \mathscr{H}$ such that $\mu(K_\epsilon) \ge 1 - \epsilon/N$ for all $\mu \in \mathscr{K}$. We can then assume K_ϵ to be convex by replacing it with its closed convex hull [Masarotto et al., 2018, Lemma 21].

Let π_k be any multicoupling of $(\mu_k^1,\dots,\mu_k^N)$. Then the marginal constraints of π_k imply that $\pi_k(K_\epsilon^N) \ge 1-\epsilon$. By Lemma 14, $\overline{\mu}_k$ must take the form $M_N\#\pi_k$ for some multicoupling π_k. Convexity of K_ϵ implies that $M_N^{-1}(K_\epsilon) \supseteq K_\epsilon^N$, and so

$$\overline{\mu}_k(K_\epsilon) = \pi_k(M_N^{-1}(K_\epsilon)) \ge \pi_k(K_\epsilon^N) \ge 1-\epsilon.$$

In particular, the sequence $(\overline{\mu}_k)$ is tight and up to subsequences. We may assume that it converges in distribution to a limit $\overline{\mu}$. By the results in Section 2.1.1 paired with Gaussianity, they converge as well in Wasserstein distance.

Step 2: a moment bound for $\overline{\mu}_k$. Let $R^i = \int_{\mathscr{H}} \|x\|^2 \, \mathrm{d}\mu^i(x)$ denote the second moment of μ^i. For a general measure ν,

$$W_2^2(\nu,\delta_0) = \inf_{\{(X,Y):\mathrm{law}(X)=\mu,\mathrm{law}(Y)=\delta_0\}} \mathbb{E}\Pi(X,Y) = \inf_{X\sim\mu} \Pi(X,0),$$

so the squared Wasserstein distance to the Dirac mass at 0 can be interpreted as a second moment (see also Theorem 7.12 in Villani [2003]). In particular, the second moment of μ_k^i converges to R^i, and for sufficiently large k is bounded above by $R^i + 1$. By Masarotto et al. [2018, Corollary 9], for k large

$$\int_{\mathscr{H}} \|x\|^2 \, \mathrm{d}\overline{\mu}_k(x) \le \frac{1}{N}\sum_{i=1}^N R^i + 1 \le \max(R^1,\dots,R^N) + 1 := R+1. \tag{2.11}$$

Step 3: the limit $\overline{\mu}$ is a Fréchet mean of (μ^i). We start by proving that the for large k the Fréchet functionals F_k are uniformly Lipschitz on the Wasserstein ball

$$B = \{\mu \in \mathscr{W} : W^2(\mu, \delta_0) \leq R + 1\},$$

with δ_0 a Dirac measure at the origin.
By the moment bound in (2.11), for any sufficiently large k, the Fréchet means $\overline{\mu}_k$ fall into B. If $\mu, \nu \in B$ then, since $\mu_k^i \in B$ for k large,

$$|F_k(\mu) - F_k(\nu)| \leq \frac{1}{2N} \sum_{i=1}^{N} [W(\mu, \mu_k^i) + W(\nu, \mu_k^i)] W(\mu, \nu) \leq 2\sqrt{R+1}\, W(\mu, \nu),$$

which yields the Lipschitz property. Assume that $\overline{\mu}_k \to \overline{\mu}$ in $\mathscr{W}$ and let $\mu \in B$, $\epsilon > 0$ and k_0 such that $W(\overline{\mu}_k, \overline{\mu}) < \epsilon/(2\sqrt{R+1})$ for all $k \geq k_0$. Since $F_k \to F$ point-wise we may assume that $|F(\mu) - F_k(\mu)| < \epsilon$ when $k \geq k_0$. The same holds for $\mu = \overline{\mu}$. Then for all $k \geq k_0$ and arbitrarily small ϵ,

$$\epsilon + F(\mu) \geq F_k(\mu) \geq F_k(\overline{\mu}_k) \geq F_k(\overline{\mu}) - \epsilon \geq F(\overline{\mu}) - 2\epsilon$$

implying that $\overline{\mu}$ minimises F over B and hence over the entire Wasserstein space $\mathscr{W}(\mathscr{H})$ and concluding the proof. □

Theorem 18 makes it possible to profit from finite-dimensional results, as we can exploit the stability of the distance and of the Fréchet mean to lift them to infinite dimensions. For example, we can deduce that the operator mean does not “swell”. For covariance matrices, swelling means that the determinant of the arithmetic mean can be larger than the determinant of its parts. One of the advantages in applications of Wasserstein means is that they do not swell, as Euclidean estimators occasionally do (see e.g. Masak [2017], Kalunga et al. [2015]). The next theorem formalizes this result for general covariance operators.

Theorem 19. *Let $\overline{\Sigma}$ be a unique Fréchet mean of $\Sigma_1, \dots, \Sigma_n$. Then $(\Sigma_1 + \cdots + \Sigma_n)/n - \overline{\Sigma}$ is non-negative.*

Proof. For positive Σ_i, the result can be found in Bhatia et al. [2018, Theorem 9]. If Σ_i are non-negative operators in finite dimensions, we can employ a regularisation to exploit the result for positive Σ_i: let $\overline{\Sigma}_\epsilon$ be the Fréchet mean of the positive operators $\Sigma_i + \epsilon\mathscr{I}$. Then $(\Sigma_1 + \cdots + \Sigma_n)/n + \epsilon\mathscr{I} - \overline{\Sigma}_\epsilon$ is non-negative for all $\epsilon > 0$. The result follows by taking $\epsilon \to 0$. This proves the result for finite dimensions.

The infinite-dimensional case shares the same ideas as the proof in Section 2.2.1 and can be deduced from the stability result.
Let $\mathscr{P}_k$ be a sequence of finite-dimensional projections that converges strongly to the identity and let $\overline{\Sigma}_k$ be any Fréchet mean of $(\mathscr{P}_k \Sigma_i \mathscr{P}_k)_{i=1}^n$. Any multicoupling assigns probability one to $[\text{range}(\mathscr{P}_k)]^n$, and by Lemma 14 the range of $\overline{\Sigma}_k$ is included in the finite-dimensional space

range($\mathcal{P}_k$). The sequence

$$(\mathcal{P}_k\Sigma_1\mathcal{P}_k + \mathcal{P}_k\Sigma_n\mathcal{P}_k)/n - \overline{\Sigma}_k$$

is non-negative for all k. Letting $k \to \infty$ gives the result. □

2.3 Computing the Fréchet mean through gradient descent

Once existence and uniqueness of the Fréchet mean are established, the next relevant question is how it can be computed. The main issue in this context is that Fréchet means rarely admit closed-form expressions. Even in the finite-dimensional Gaussian case, in general a closed-form expression is obtainable only for a sample of size two, in which case the Wasserstein mean between Σ_1 and Σ_2 is exactly the mid-point of the Wasserstein geodesics connecting the them (see Bhatia et al. [2018] for matrices and Zemel [2017] for operators).
The lack of a closed-form expression entails a need for some sort of numerical computational scheme, to be applied on some (possibly smoothed) finite-dimensional version of the operators.

Let $\Sigma_1, \ldots, \Sigma_N$ be covariance operators on a finite dimensional subspace $\mathcal{H}' \subset \mathcal{H}$. Assume we wish to compute a Fréchet mean $\overline{\Sigma}$ for $\Sigma_1, \ldots, \Sigma_N$.
Numerical methods to compute the Fréchet mean have filled the line of work of several authors. Álvarez-Esteban et al. [2016] and Bhatia et al. [2018] exploited equation (2.10) to propose a fixed-point iteration algorithm for Gaussian measures. Pigoli et al. [2014a] suggested an iterative procedure to find $L = \overline{\Sigma}^{1/2}$ based on generalised Procrustes analysis (Gower [1975] and Dryden and Mardia [1998]), and which is summarised in Section 1.2.5. We adopt a novel perspective, made possible by the knowledge of the Wasserstein geometry. A similar perspective will be embraced in Chapter 3 and will be of key importance in applications. Rather than limiting the algorithm to the manifold of covariance operators, we lift the problem of computing the mean to its tangent space. More specifically, the algorithm lifts all observations to the tangent space centered at an initial guess of the Fréchet mean via the log map, and then averages all the lifted-operators linearly on the tangent space. The retraction through the exponential map of this linear average onto the manifold provides the next iterate.

The algorithm effectively implements a series of Procrustes-type iterations in the sense of Gower [1975] and Dryden and Mardia [1998] and was proposed in this tangent space form by Zemel and Panaretos [2017]. In finite dimensions, these iterations are proved to converge to the unique Fréchet mean $\overline{\Sigma}$ of $\Sigma_1, \ldots, \Sigma_N$ as soon as one of the Σ_i is injective, and independently of the initial point, provided that the initial point is chosen to be injective. Convergence has been independently proved by Zemel and Panaretos [2017] and Álvarez-Esteban et al. [2016].

We now give a high-level description of the algorithm. We present it in terms of covariance operators (or equivalently, centered Gaussian measures), however the same procedure applies to any finite collection of measures in Wasserstein space.

- Initialise the procedure at some injective covariance $\Sigma^0 : \mathscr{H}' \to \mathscr{H}'$.
- Denote as Σ^k the current iterate at step k.
- For each i compute the optimal maps from Σ^k to each of the operators $\Sigma_i : \mathscr{H}' \to \mathscr{H}'$ via

  $$\mathbf{t}_{\Sigma^k}^{\Sigma_i} = (\Sigma^k)^{-1/2}[(\Sigma^k)^{1/2}\Sigma_i(\Sigma^k)^{1/2}]^{1/2}(\Sigma^k)^{-1/2}.$$

- Compute the average $T_k = N^{-1}\sum_{i=1}^N \mathbf{t}_{\Sigma^k}^{\Sigma_i}$, which is itself a non-negative matrix on $\mathscr{H}'$.
- Set the next iterate to $\Sigma^{k+1} = T_k\Sigma^k T_k$ (guaranteed to be injective on $\mathscr{H}'$ if at least one Σ_i is so).

Álvarez-Esteban et al. [2016] show $\mathrm{tr}\Sigma^{\mathrm{k}}$ to be increasing in k, and Zemel and Panaretos [2017] show that the optimal maps $\mathbf{t}_{\Sigma^k}^{\Sigma_i}$ converge uniformly over compacta to $\mathbf{t}_{\overline{\Sigma}}^{\Sigma_i}$ as $k \to \infty$.

The Wasserstein–Procrustes algorithm shares a connection with gradient descent. In order to see it, write the finite-dimensional Fréchet functional at iteration k as

$$F(\Sigma^k) = \frac{1}{2N}\sum_{i=1}^N \Pi(\Sigma^k, \Sigma_i).$$

By Zemel and Panaretos [2017, Theorem 1], its gradient is

$$F'(\Sigma^k) = -\frac{1}{N}\sum_{i=1}^n \log_{\Sigma^k}(\Sigma_i) = -\frac{1}{N}\sum_{i=1}^N \left(\mathbf{t}_{\Sigma^k}^{\Sigma_i} - \mathcal{I}\right),$$

i.e., exactly the (negative) average that enters into the third step of the algorithm. This makes the Wasserstein–Procrustes algorithm effectively a steepest descent in the space of covariances endowed with the Procrustes metric, a result that was proved by Zemel and Panaretos [2017]. Finally, it is worth mentioning that when the covariance operators commute, either algorithm converges to the Fréchet mean after a single iteration. In the Wasserstein-inspired algorithm this requires to initialise the algorithm at a positive linear combination of the operators in the sample, or a positive power thereof.

2.4 Phase variation and generative models for deformations

Section 1.2.3 gave the geometrical intuition that led to the choice of the Procrustes distance as a metric for the space of covariances, namely, the link between Gram matrices and square roots of covariances, and the consequent connection to shape theory. Throughout this thesis, we explained how covariance operators can be raised to be the main object of interest in the statistical analysis. We stressed the importance of picking a suitable non-linear metric, as to respect the curved geometry on the manifold of covariances. What it might be less clear to the reader, is how this translates into a motivation to chose to endow the space of covariance operators specifically with the Procrustes-Wasserstein distance. This section aims to make clear the motives behind this choice. More precisely, we show how the connection between the Procrustes distance and the registration of shapes is echoed by the connection between the Wassestein distance and the registration of curves, thus making the Wasserstein metric “natural” when we assume a generative mechanism for the data based on random deformations.

While dealing with several populations exhibiting second-order variation, it becomes relevant to be able to quantify and understand this variation. We mentioned in the Introduction and in Chapter 1 how it is very common to perform statistical analysis on covariances by embedding the operators into a larger, linear, Hilbert–Schmidt space. Inference is then performed according to the corresponding metric (there is a multitude of references dealing with Hilbert–Schmidt embedding. For example Benko et al. [2009], Panaretos et al. [2010a]),Horváth et al. [2013], Paparoditis and Sapatinas [2014], Gabrys et al. [2010], Fremdt et al. [2013], Horváth and Kokoszka [2012], Jarušková [2013], Coffey et al. [2011], Kraus [2014]). We now try to give an idea of more subtleties involved in endowing a space with a particular metric. In statistics, a specific choice of a distance is intimately connected with some sort of model generator for the data at hand. By imbedding the operators into a Hilbert–Schmidt space, and using the Hilbert–Schmidt distance, the model that is implicitly assumed considers that the sample of (different) covariance operators arises as a linear perturbation of an underlying operator, i.e.,

$$\Sigma_k = \Sigma + E_k. \tag{2.12}$$

Here E_k is a random zero-mean, self-adjoint, trace-class operator, and the equation is constrained so that the left hand side is non-negative definite.
As a random trace-class self-adjoint operator, E admits its own Karhunen–Loève expansion, which is what usually is employed in the statistical analysis.
However we know that covariance operators lie in a fundamentally non-linear space, and therefore are not closed under linear perturbations such as (2.12).

One of the main advantages of the Wasserstein distance is having a canonical connection with the problem of warping (or phase variation). Due to this, the natural “Wassestein” generative model is the one of random deformations. Let X be a Gaussian process. We have a phase-variation if instead of X we observe $X^*(\cdot) = X(T^{-1}(\cdot))$ for a random invertible function T called a *deformation* or *warping* function (cfr. Section 1.1.4). For data curves, phase-variation is normally regarded as a “variation in the x-axes”, as opposite to amplitude variation, which can be viewed as a variation on the y-axes. See for example the plot of the age curves in the Berkeley growth dataset in Section 1.1.4, where it is possible to identify the two types of variations quite clearly.

If now X has covariance Σ, and given a random[4] non-negative bounded operator T on $\mathscr{H}$, we know that TX is a Gaussian process with covariance $T\Sigma T^*$, conditional upon T. We would

---

[4] In the sense that T is Bochner measurable from a probability space Ω to $B(\mathscr{H})$. In particular $T(\Omega)$ will be a separable subset of (the non-separable) $B(\mathscr{H})$ [Panaretos and Zemel, To appear].

like Σ to be the "true average" covariance of TX. A natural (and indeed necessary) assumption in that direction is that the expected value of the perturbation T is the identity, i.e. $\mathbb{E}T = \mathfrak{I}$. If this holds, given Σ together with a collection $T_1, \dots, T_N$, the conjugation perturbations

$$\Sigma_k = T_k \Sigma T_k^*, \qquad k = 1, \dots, N$$

return a generative model for $\Sigma_1, \dots, \Sigma_N$ which is non-linear on the operator space, but linear on the tangent space (through the T_k), and somehow canonical for Procrustes metric. The next theorem formalises the result.

Theorem 20 (Generative Model)**.** *Let Σ be a covariance operator and let $T : \mathscr{H} \to \mathscr{H}$ be a random non-negative linear map with $\mathbb{E}\|T\|_\infty^2 < \infty$ and $\mathbb{E}T = \mathfrak{I}$. Then the random operator $T\Sigma T^*$ has Σ as a Fréchet mean in the Procrustes metric,*

$$\mathbb{E}[\Pi^2(\Sigma, T\Sigma T^*)] \le \mathbb{E}[\Pi^2(\Sigma', T\Sigma T^*)],$$

for all non-negative nuclear operators Σ'.

Proof. We follow Zemel and Panaretos [2017, Theorem 5] and replicate their argument, which uses the Kantorovich duality (Villani [2003, Theorem 5.10]). Define the function $\varphi(x) = \langle Tx, x\rangle/2$ and its Legendre transform $\varphi^*(y) = \sup_{x\in\mathscr{H}} \langle x, y\rangle - \varphi(x)$. With an abuse of notation we write $d\Sigma(x)$ for integration with the corresponding measure with that covariance. The strong and weak Kantorovich dualities yield

$$\begin{aligned}
\frac{1}{2}W^2(N(0,\Sigma), N(0, T\Sigma T)) &= \int_{\mathscr{H}} \left(\frac{1}{2}\|x\|^2 - \varphi(x)\right) \mathrm{d}\Sigma(x) + \int_{\mathscr{H}} \left(\frac{1}{2}\|y\|^2 - \varphi^*(y)\right) \mathrm{d}T\Sigma T(x); \\
\frac{1}{2}W^2(N(0,\Sigma'), N(0, T\Sigma T)) &\ge \int_{\mathscr{H}} \left(\frac{1}{2}\|x\|^2 - \varphi(x)\right) \mathrm{d}\Sigma'(x) + \int_{\mathscr{H}} \left(\frac{1}{2}\|y\|^2 - \varphi^*(y)\right) \mathrm{d}T\Sigma T(x).
\end{aligned}$$

Now $\mathbb{E}\varphi(x) = \|x\|^2/2$ because $\mathbb{E}T = \mathfrak{I}$. Taking expectations in the equalities above and using Fubini's theorem proves the result. Note that when T takes finitely many values, this provides a proof for empirical Fréchet means.

We modify the construction in Zemel and Panaretos [2017], adapting it to the unboundedness of the spaces. Let Ω be the underlying probability space and $B(\mathscr{H})$ the set of bounded operators on $\mathscr{H}$ with the operator norm topology. Assume that $T : \Omega \to B(\mathscr{H})$ is Bochner measurable with (Bochner) mean $\mathfrak{I}$. Notice that

$$W^2(N(0, S\Sigma S^*), N(0, T\Sigma T^*)) \le \int_{\mathscr{H}} \|S(x) - T(x)\|^2 \,\mathrm{d}\Sigma(x) = \mathrm{tr(S-T)\Sigma(S^*-T^*)} \le \|\mathrm{S-T}\|_\infty^2 \mathrm{tr}\Sigma.$$

Similarly,

$$\left|\int_{\mathscr{H}} \langle (T-S)x, x\rangle \,\mathrm{d}\Sigma'(x)\right| = \left|\mathrm{tr(T-S)}\Sigma'\right| \le \|T - S\|_\infty \mathrm{tr}\Sigma'$$

so the integrals with respect to φ are measurable (from Ω to $\mathbb{R}$) for all Σ', and integrable because

$\mathbb{E}\|T\|_\infty < \infty$. In particular (the measure corresponding to) $T\Sigma T : \Omega \to \mathscr{W}(\mathscr{H})$ is measurable, because it is a continuous Lipschitz function of T.
Furthermore, the integrals with respect to φ^* are measurable and integrable, because so is $W^2(\Sigma', T\Sigma T)$, and because they are given by the difference between integrable functionals. To conclude the proof it remains to show that for all Σ'

$$\mathbb{E}\int_{\mathscr{H}} \langle Tx, x\rangle \, \mathrm{d}\Sigma'(x) = \int_{\mathscr{H}} \langle (\mathbb{E}T)x, x\rangle \, \mathrm{d}\Sigma'(x).$$

This is clearly true if T is simple (takes finitely many values). Otherwise, we can find a sequence of simple $T_n : \Omega \to B(\mathscr{H})$ such that $\|T_n - T\|_\infty \to 0$ almost surely and in expectation. This Fubini equality holds for T_n and

$$\begin{aligned}\left|\mathbb{E}\int_{\mathscr{H}} \langle Tx, x\rangle \, \mathrm{d}\Sigma'(x) - \mathbb{E}\int_{\mathscr{H}} \langle T_n x, x\rangle \, \mathrm{d}\Sigma'(x)\right| &= |\mathbb{E}\mathrm{tr}(\mathrm{T}-\mathrm{T_n})\Sigma'| \le \mathrm{tr}\Sigma'\mathbb{E}\|\mathrm{T}-\mathrm{T_n}\|_\infty \\ \left|\int_{\mathscr{H}} \langle (\mathbb{E}T)x, x\rangle \, \mathrm{d}\Sigma'(x) - \int_{\mathscr{H}} \langle (\mathbb{E}T_n)x, x\rangle \, \mathrm{d}\Sigma'(x)\right| &= |\mathrm{tr}(\mathbb{E}\mathrm{T}-\mathbb{E}\mathrm{T_n})\Sigma'| \le \mathrm{tr}\Sigma'\|\mathbb{E}\mathrm{T}-\mathbb{E}\mathrm{T_n}\|_\infty.\end{aligned}$$

By approximation the Fubini equality holds for T, completing the proof. □

We would like to stress the fact that the assumption $\mathbb{E}\|T\|_\infty^2 < \infty$ guarantees that the Fréchet functional $\mathbb{E}\Pi^2(A, T\Sigma T)$ is finite for any covariance[5] operator Σ. When the measures are defined on $\mathbb{R}^d$ and have compact support, the result in Theorem 20 holds in a more general Wasserstein setup, where μ is a fixed measure and T is a random optimal map with mean identity (Bigot and Klein [2012]; Zemel and Panaretos [2017]).

The proof of Theorem 20 does not make specific use of linearity of T or Gaussianity. When Gaussianity is assumed however, the Fréchet functional can be evaluated explicitly due to the Wasserstein distance formula. In finite dimensions, this allows a more constructive proof of Theorem 20. We report it in the next paragraph.

Alternative proof of Theorem 20 in finite dimensions. We first evaluate the term $\mathbb{E}\mathrm{tr}(\mathrm{T}\Sigma\mathrm{T}^*) = \mathbb{E}\mathrm{tr}(\mathrm{T}\Sigma\mathrm{T})$ in the Wasserstein distance, using $\mathbb{E}T = \mathfrak{I}$, as

$$\mathbb{E}\mathrm{tr}(\mathrm{T}-\mathfrak{I})\Sigma(\mathrm{T}-\mathfrak{I}) + \mathbb{E}(\mathrm{tr}\Sigma\mathrm{T}) + \mathbb{E}(\mathrm{tr}\mathrm{T}\Sigma) - \mathbb{E}\mathrm{tr}\Sigma = \mathrm{tr}[\mathrm{Cov}(\mathrm{T})\Sigma] + \mathrm{tr}(\Sigma),$$

where $\mathrm{Cov}T = \mathbb{E}[(T-\mathfrak{I})(T-\mathfrak{I})]$. Consequently, the Fréchet functional at Σ equals

$$\begin{aligned}\mathbb{E}W^2(T\Sigma T^*, \Sigma) &= \mathrm{tr}(\Sigma) + \mathbb{E}\mathrm{tr}(\mathrm{T}\Sigma\mathrm{T}^*) - 2\mathbb{E}\mathrm{tr}(\Sigma^{1/2}\mathrm{T}\Sigma\mathrm{T}^*\Sigma^{1/2})^{1/2} \\ &= 2\mathrm{tr}(\Sigma) + \mathrm{tr}[\mathrm{Cov}(\mathrm{T})\Sigma] - 2\mathbb{E}\mathrm{tr}(\Sigma^{1/2}\mathrm{T}\Sigma^{1/2}) \\ &= 2\mathrm{tr}(\Sigma) + \mathrm{tr}[\mathrm{Cov}(\mathrm{T})\Sigma] - 2\mathrm{tr}(\Sigma) \\ &= \mathrm{tr}[\mathrm{Cov}(\mathrm{T})\Sigma].\end{aligned}$$

---

[5] actually, non-negative and nuclear

We now compute the functional at an arbitrary Σ':

$$\begin{aligned}\mathbb{E}W^2(T\Sigma T^*,\Sigma') &= \mathrm{tr}(\Sigma') + \mathbb{E}\mathrm{tr}(\mathrm{T}\Sigma\mathrm{T}^*) - 2\mathbb{E}\mathrm{tr}(\Sigma'^{1/2}\mathrm{T}\Sigma\mathrm{T}^*\Sigma'^{1/2})^{1/2}\\ &= \mathrm{tr}(\mathrm{Cov}(\mathrm{T})\Sigma) + \mathbb{E}\Big\{\mathrm{tr}(\Sigma'^{1/2}\mathrm{T}\Sigma'^{1/2}) + \mathrm{tr}(\Sigma^{1/2}\mathrm{T}\Sigma^{1/2}) - 2\mathrm{tr}(\Sigma'^{1/2}\mathrm{T}\Sigma\mathrm{T}\Sigma'^{1/2})^{1/2}\Big\}.\end{aligned}$$

To prove that $F(\Sigma') \geq F(\Sigma)$ we just need to show that the term inside the expectation is non-negative. We will do this by interpreting it as the Wasserstein distance between $B = T^{1/2}\Sigma' T^{1/2}$ and $A = T^{1/2}\Sigma T^{1/2}$. Write $B_1 = \Sigma'^{1/2} T \Sigma'^{1/2}$, $A_1 = \Sigma^{1/2} T \Sigma^{1/2}$. The Wasserstein distance is non-negative, so

$$2\mathrm{tr}(\mathrm{A}^{1/2}\mathrm{B}\mathrm{A}^{1/2})^{1/2} \leq \mathrm{trA} + \mathrm{trB} = \mathrm{trA}_1 + \mathrm{trB}_1.$$

Therefore we must show that $\mathrm{tr}(\mathrm{A}^{1/2}\mathrm{B}\mathrm{A}^{1/2})^{1/2} = \mathrm{tr}(\Sigma'^{1/2}\mathrm{T}\Sigma\mathrm{T}\Sigma'^{1/2})^{1/2}$.
We point out that the next step is what makes the proof intrinsically finite-dimensional, as every argument given so far holds in infinite dimensions as well. We will prove that $\mathrm{tr}(\mathrm{A}^{1/2}\mathrm{B}\mathrm{A}^{1/2})^{1/2} = \mathrm{tr}(\Sigma'^{1/2}\mathrm{T}\Sigma\mathrm{T}\Sigma'^{1/2})^{1/2}$ by showing that these *matrices* are conjugate. Assume first that Σ, Σ' and T are invertible and write

$$D = \Sigma'^{1/2} T \Sigma T \Sigma'^{1/2} = \Sigma'^{-1/2} T^{-1/2} [BA] T^{1/2} \Sigma'^{1/2} = \Sigma'^{-1/2} T^{-1/2} A^{-1/2} [A^{1/2} B A^{1/2}] A^{1/2} T^{1/2} \Sigma^{1/2}.$$

Thus the non-negative matrices D and $A^{1/2}BA^{1/2}$ have the same eigenvalues. This stays true for their square roots, that consequently have the same trace. If now the matrices are singular, we just need to notice that singular matrices can be approximated by non-singular ones, making the proof valid without restriction in finite dimensions. □

If the law of the random deformation T is finitely supported, Theorem 20 can be strengthened and we can remove the boundness assumption on T. This is a significant improvement, as optimal maps are not in general bounded (see Section 3.1).

Theorem 21. *Let Σ be a covariance operator corresponding to a centred Gaussian measure $\mu \equiv N(0,\Sigma)$ and let $D \subseteq \mathscr{H}$ be a dense linear subspace of μ-measure one. If $T_1,\dots,T_n : D \to \mathscr{H}$ are (possibly unbounded) non-negative operators such that $T_i \in \mathscr{L}_2(\mu)$ for all $i = 1,\dots,n$ and $\sum T_i(x) = nx$ for all $x \in D$, then Σ is a Fréchet mean of the finite collection $\{T_i\Sigma T_i : i = 1,\dots,n\}$.*

Proof. Direct calculations show that the functions $\varphi_i(x) = \langle T_i x, x\rangle/2$ are convex on D and $T_i x$ is a sub-gradient of φ_i for any i and any $x \in D$. We can therefore use the duality of the proof of Theorem 20 with the integrals involving φ_i taken on D rather than on the whole of $\mathscr{H}$. Now φ and φ^* are non-negative functions and the Wasserstein distance is finite, which make these integrals finite too (the integral of $\|x\|^2/2 - \varphi(x)$ with respect to Σ' may be negative infinite, but this does not hinder the validity of the arguments). Since there are finitely many integrals, there are no measurability issues and we have $F(\mu) \leq F(\nu)$ whenever $\nu(D) = 1$. Now D is dense in $\mathscr{H}$ and continuity arguments yield that $F(\mu) \leq F(\nu)$ for all $\nu \in \mathscr{W}(\mathscr{H})$, implying that μ (hence Σ) is a Fréchet mean. □

3 Applications

Once data come subdivided into several populations, each having its own sample mean and covariance operators, we can inquire about the possible variations across these populations. This chapter contains proposals and results regarding the second-order variational analysis on a dataset of populations of covariance operators. We consider testing the hypothesis of equality of covariances across different populations in Section 3.1. If the null hypothesis is rejected, Principal Component Analysis (PCA) offers a valid tool to understand the differences within the data and describe the main mode(s) of variation. The understanding of the Wasserstein geometry allows to perform PCA on the tangent space, and the details are discussed in Section 3.2. Clustering of operators is treated in Section 3.3. Two different clustering methods for covariances are presented, one based on lifting the covariances on the tangent space and using the classical K-means algorithm, the second, coined *soft* clustering, using directly the Wasserstein distance.

Each section in this chapter contains a description of the methodology employed and a part of numerical simulations and data analysis. For convenience, we collected the descriptions of the simulation setups and the dataset employed in the analysis in Section 3.1.1.

3.1 Transportation-based functional ANOVA of Covariances

Assume we have N populations of curves. Namely, let $\{X_{i,1}\}_{i=1}^{n_1}, \dots, \{X_{i,N}\}_{i=1}^{n_N}$ be N independent samples of i.i.d. random elements in a separable Hilbert space $\mathscr{H}$. We aim to investigate the fluctuations not in the mean structure of the $\{X_{i,j}\}$ but rather *around* their mean, as j varies. Similar problems have been studied before. For example, Panaretos et al. [2010b], Kraus and Panaretos [2012], and Tavakoli and Panaretos [2016] considered several groups of DNA mini-circles vibrating in solution. Here, each sequence corresponds to a different group and different vibration properties would highlight a dependence of the mechanical properties on the base pair sequence. Another example lies in the analysis of handwriting [Ramsay, 2000]. Here different authors of different handwritings give rise to populations of written words (seen as curves), and one may wish to see whether (specific repetitions of) written words come from

the same author or not.
In analogy to the functional analysis of variance (Benko et al. [2009], Zhang [2013]), which treats the topic of fluctuations in the mean structure, we named our problem a functional covariance ANOVA (or transportation ANOVA, for reasons which will soon become clear).

Mathematically, given $\{X_{i,1}\}_{i=1}^{n_1}, \dots, \{X_{i,N}\}_{i=1}^{n_N}$ centered processes with respective (well-defined) covariances $\{\Sigma_j\}_{j=1}^{N}$, we are interested in testing the hypothesis

$$H_0 : \Sigma_1 = \Sigma_2 = \ldots = \Sigma_N \tag{3.1}$$

on the basis of the observations $\{X_{i,j}\}$ against

$$H_1 : \big\{\text{At least one operator is different}\big\}. \tag{3.2}$$

Most work in this topic so far employed the Hilbert–Schmidt metric, as seen for example in Boente et al. [2018], Panaretos et al. [2010b], Fremdt et al. [2013]. The issue with employing the Hilbert–Schmidt metric is that it implicitly imbeds covariance operators in a larger linear space, the Hilbert–Schmidt space, while covariances are not closed under linear operations, as we remarked in Chapter 2.
Driven by the need to analyse cross-linguistic variation of phonetics in Romance languages, Pigoli et al. [2014a] were the first to consider nonlinear metrics in the two-sample testing. In finite dimensions, some of these ideas can be found in Dryden et al. [2009].
The testing procedure of Pigoli et al. [2014a] was insightful in the taking into account the structural geometry of the space. However, it did not bring significant improvements compared with the results of Panaretos et al. [2010b].
The 2-sample testing procedure of Pigoli et al. [2014a] has been generalised to k-sample testing by Cabassi et al. [2017]. Their idea is to perform a k-sample permutation test as a series of partial 2-sample tests, where the partial test statistics are combined through the non-parametric combination algorithm of Pesarin and Salmaso [2010]. As test statistics, they consider $T_{ij} = d(\hat{\Sigma}_i, \hat{\Sigma}_j)$, where $\hat{\Sigma}_i$ and $\hat{\Sigma}_j$ are the sample covariance operators of the corresponding groups, and d can be any metric on the operator space. After performing the global test, if the null hypothesis H_0 is rejected, Cabassi et al. [2017] propose a subsequent analysis to investigate which covariance operators are indeed different through the techniques explained in Pesarin and Salmaso [2010]. Their methodology gave rise to the R-package `fdcov`.
Cabassi et al. [2017] compare their method favourably against other methods, such as the $k-$sample test via concentration inequalities of Kashlak et al. [2016], demonstrating an increase in power and illustrating how their performance can be considered state-of-the-art.

We take a completely different approach. Taking advantage of our understanding of the Wasserstein geometry, we can construct a 2-sample test that seems more powerful than other approaches, while still respecting the geometry of the data. The test focusses on the transport maps giving rise to the Procrustes distance, rather than the distance itself. Furthermore, it has a natural k-sample analogue that comes from the optimal multicoupling problem, and allows

for a testing procedure that is more powerful than the pairwise method of Cabassi et al. [2017], at least in the functional case.
The key observation is that under suitable conditions, testing the equality of covariance operators $\{\Sigma_1,\dots,\Sigma_N\}$ translates into a testing problem concerning the existence of a deterministic multicoupling of the collection of centred Gaussian measures $\{\mathcal{N}(0,\Sigma_1),\dots,\mathcal{N}(0,\Sigma_N)\}$. In the following, we show that such a multicoupling can be realised through deterministic optimal transport maps, and that the problem of testing the equality between covariance operators can be solved by comparing these transport maps to the identity.

Recall from Section 2.2 that a Gaussian optimal multicoupling is manifested as a joint distribution of a collection of Gaussian processes, such that the marginals are pairwise coupled as tightly as possible. An optimal multicoupling of a collection of N Gaussian processes $(X_1,\dots,X_N)$ with $X_j \sim \mathcal{N}(0,\Sigma_j)$ is *deterministic* if $(X_1,\dots,X_N)$ arise as the image of a single process X through a collection of deterministic maps $\mathbf{t}_j : \mathcal{H} \to \mathcal{H}$,

$$(X_1,\dots,X_N) \overset{d}{=} (\mathbf{t}_1(X),\dots,\mathbf{t}_N(X)).$$

Masarotto et al. [2018] show that an optimal multicoupling of Gaussian measures always exists, yet the same cannot be said for a deterministic multicoupling. In other words, the multicoupling may not have only "one degree of freedom". Nevertheless, Masarotto et al. [2019] show that the optimal Gaussian multicoupling can always be manifested by deterministic transport maps, and indeed by bounded linear maps. Furthermore, under the null H_0 in 3.1, this multicoupling becomes the "trivial" one where all maps coincide. We formally state both results.

Lemma 22. *Hypothesis* (3.1) *holds true if and only if the (unique) optimal multicoupling of* $(\gamma_1,\dots,\gamma_N)$ *can be achieved by transport maps satisfying* $\mathbf{t}_1 = \cdots = \mathbf{t}_N$.

Theorem 23. *Let* $\{\gamma_1,\dots,\gamma_N\}$ *be an arbitrary finite collection of Gaussian measures on* $\mathcal{H}$ *with mean zero. Then there exists an optimal multicoupling of* $\{\gamma_j\}_{j=1}^N$ *manifested by deterministic transport maps* $\mathbf{t}_j : \mathcal{H} \to \mathcal{H}$ *that are bounded non-negative linear operators satisfying* $\|\mathbf{t}_j\|_\infty \leq N$*, for all* $j \leq N$.

Theorem 23 allows us to define our testing method. Before doing that, notice that once a sample $\{\Sigma_j\}_{j=1}^N$ of covariances is given together with their Fréchet mean $\bar{\Sigma}$, then each element Σ_j can be identified with the transport map $\mathbf{t}_{\bar{\Sigma}}^{\Sigma_j}$. If moreover at least one Σ_j is injective, then their Fréchet mean is unique, and the maps $\mathbf{t}_{\bar{\Sigma}}^{\Sigma_j}$ are computable in closed form (see Section 2.2). The proof of Theorem 23 establishes that the optimal maps $\mathbf{t}_{\bar{\Sigma}}^{\Sigma_j}$ exist and are bounded [Masarotto et al., 2019], and give rise to the optimal deterministic multicoupling of $X_1 \sim \mathcal{N}(0,\Sigma_1),\dots,X_N \sim \mathcal{N}(0,\Sigma_N)$. These maps will be also centered around the identity, i.e.

$$\frac{1}{N}\sum_{j=1}^{N} \mathbf{t}_{\bar{\Sigma}}^{\Sigma_j} = \mathfrak{I}. \tag{3.3}$$

To lighten the notation, we denote the optimal maps by $\mathbf{t}_j$.

Theorem 23 suggests that rather than considering the null hypothesis (3.1), we can test the equivalent hypothesis

$$H_0' : \mathbf{t}_1 - \mathfrak{I} = \ldots = \mathbf{t}_N - \mathfrak{I} = 0 \tag{3.4}$$

where $\mathfrak{I}$ is the identity on $\mathscr{H}$. Under H_0, all differences $\Delta_j := \mathbf{t}_j - \mathfrak{I}$ are equal to 0, while at least one of them will be different from 0 under the alternative.

Theorem 23 implies that the Δ_j are bounded, thus their difference from 0 can be quantified in terms of their operator norm. Notice how this establishes a parallel with ANOVA. We can also imagine to contrast the Δ_j to 0 by means of a stronger norm, if such norm is finite. By using the Hilbert–Schmidt norm for example, or again the trace-norm, one can imagine that one might be able to detect finer differences, and identify weaker departures from H_0.

We move now to a description of the implementation procedure. Consider N independent groups of functional data $\{X_{ij},\ j = 1,\ldots,N,\ i = 1,\ldots,n_j\}$ and let Σ_j, $j = 1,\ldots,N$ be the covariance operator for each group.
In practice, we only have access to estimated covariances (e.g., the empirical covariance of smoothed versions of the $\{X_{ij}\}$ (Ramsay and Silverman [2005b]), or PACE-type estimators (Yao et al. [2005a])). Denote as $\{\hat{\Sigma}_j\}$ these empirical estimates of the full covariances Σ_j, computed from the n_j-sized sample in each group and by $\hat{\Sigma}$ their unique empirical (weighted) Fréchet mean (computed as in Section 2.3):

$$\hat{\Sigma} = \underset{\mathbb{R}^{q\times q}\ni\Gamma\geq 0}{\arg\min} \sum_{j=1}^{N} n_j \Pi^2(\hat{\Sigma}_j, \Gamma).$$

The corresponding optimal maps are denoted by

$$\hat{\mathbf{t}}_j = \hat{\Sigma}^{-1/2}(\hat{\Sigma}^{1/2}\hat{\Sigma}_j\hat{\Sigma}^{1/2})^{1/2}\hat{\Sigma}^{-1/2}, \quad j = 1,\ldots,N,$$

while we write the empirical deviations of these optimal maps from the identity as

$$\hat{\Delta}_j = \hat{\mathbf{t}}_j - \mathfrak{I}_{q\times q}, \qquad j = 1,\ldots,N.$$

Our test statistics is then

$$T_r = \sum_{j=1}^{N} n_j \|\Delta_j\|_r^2. \tag{3.5}$$

where $r \in \{1,2,\infty\}$ to consider the operator norm as well as the Hilbert–Schmidt and the trace-class norms.
Since we want to perform the test without any parametric assumption on the sample, we compute the P-value via permutation. The procedures is:

- reassign the $\left(\sum_{j=1}^k n_j\right)$ curves $\{X_{i,j},\ i = 1,\dots,n_j,\ j = 1,\dots,N\}$ into N groups, respecting the sizes of the initial groups. Call these new "data" $X^*_{i,j}$.
- Construct the empirical covariance $\hat\Sigma^*_j$ for the jth group $\{X^*_{i,j}\}_{i=1}^n$, $j = 1,\dots,N$.
- Compute the empirical (weighted) Fréchet mean $\hat\Sigma^*$ of $\{\hat\Sigma^*_1,\dots,\hat\Sigma^*_N\}$.
- Construct

  $$\hat{\mathbf{t}}^*_j = (\hat\Sigma^*)^{-1/2}\big((\hat\Sigma^*)^{1/2}\hat\Sigma^*_j(\hat\Sigma^*)^{1/2}\big)^{1/2}(\hat\Sigma^*)^{-1/2}$$

  and compute

  $$T^*_r = \sum_{j=1}^N n_j\|\hat{\mathbf{t}}^*_j - \mathfrak{I}_{q\times q}\|_r^2 = \sum_{j=1}^N n_j\|\hat\Delta^*_j\|_r^2.$$

Iterating this procedure for all possible re-assignments of the indexes gives the distribution of the permuted statistics T^*_r, which in turn can be used to generate a p-value for T_r under the null hypothesis. Under H_0, all possible permutations of the operators labels have equal probability $p = 1/K!$. Obtaining an exact test would thus require $K!$ permutations of the labels, making it computationally prohibitive for large K. Therefore, rather than computing an exact p-value, we resort to a Monte Carlo sample of permutations.

Similar steps would allow for the implementation of a bootstrap-type procedure, simply by randomly permuting indices with replacement. However, the exchangeability of the permutation labels under H_0 guarantees the exactness of the level of the permutation test for finite samples (Pesarin and Salmaso [2010]), so we focus on the permutation test only.

3.1.1 Simulation scenarios

Before reporting the simulation results, we describe the synthetic and real datasets employed in the simulations.

- **(Perturbations of) Berkeley growth study data, Jones and Bayley [1941].** This simulation scenario is taken directly from Cabassi et al. [2017]. It is only considered in the k-sample testing application, in order to compare our method with Cabassi et al. [2017] on the same benchmark scenario.

  The data set contains the heights of 39 boys and 54 girls from age 1 to 18 and the ages at which they were collected. It is contained in the R-package `fda`. We generate N populations whose covariance is a perturbation on known operators Σ_f and Σ_m, the sample covariances operators of the male and female subjects in the Berkeley growth data set (Jones and Bayley [1941]). The perturbations take two different forms:

1. *geodesic perturbations:* $k_1 < N$ groups have covariance operator Σ_m. The other $k_2 = N - k_1$ groups have covariance operator

   $$\Sigma(\gamma) = [\Sigma_m^{1/2} + \gamma(\Sigma_f^{1/2} R - \Sigma_m^{1/2})][\Sigma_m^{1/2} + \gamma(\Sigma_f^{1/2} R - \Sigma_m^{1/2})]^*, \tag{3.6}$$

   where R the operator minimising the Procrustes–Wasserstein distance between Σ_f and Σ_m, and $\gamma \in [0,5]$ is a parameter which controls how far this covariance operator is from Σ_m.

2. *additive perturbations:* $k_1 < N$ groups have covariance operator Σ_m. The other $k_2 = N - k_1$ groups have covariance operator

   $$\Sigma(\gamma) = (1+\gamma)\Sigma_m, \tag{3.7}$$

   with $\gamma \in [0,5]$.

In practice, Σ_f and Σ_m are estimated from spline-smoothed growth curves via the R-package `fda`. We chose this smoothing method because we wish to exactly replicate the simulation scenario in Cabassi et al. [2017], which employs smoothed growth curves to compute Σ_f and Σ_m. Specifically, the original observations are evaluated as a linear combination of 12 B-spline basis functions via the function `create.bspline.basis`. They are smoothed on 31 knots, equally distributed on the range of the curves. Finally, the use of the command `var.fd` yields the empirical covariances $\hat{\Sigma}_m$ and $\hat{\Sigma}_f$, estimated from the smoothed curves.

- **Generative model.** The choice of any metric, and in particular of the Wasserstein distance, is intrinsically related to the model assumed for the generation of the sample at hand. The Wasserstein distance links naturally with a generation procedure based on random deformations, as described in Section 2.4. We report it here for the reader's convenience: if $T_1, \dots, T_N$ are non-negative operators with mean identity, then any covariance operator Σ is the Fréchet mean of $\{\Sigma_j = T_j \Sigma T_j\}_{j=1}^N$, and the maps $\mathbf{t}_j$ in (3.10) are exactly T_j (on the closed range of Σ).

  For testing purposes, it is convenient to produce a simulation setup where the dataset occurs as a series of known perturbations of a (known) Fréchet mean. We also wish for such perturbations to move away from the commutative case, since assuming that the generative maps commute would make the computation trivial (cf. Section 2.1).
  We build a sample of covariance operators as perturbations of an underlying "true" Fréchet mean $\overline{\Sigma}$. These perturbations are given by the optimal maps T_i, with $\mathbb{E}T = \mathscr{I}$. We decided to produce a generative setting for such optimal maps $\{T_1, \dots, T_N\}$ as follows:

  $$T_j = k^{-1} \sum_{n=1}^{\infty} \delta_n^{(j)} \sin(2n\pi t - \theta^{(j)}) \otimes \sin(2n\pi t - \theta^{(j)}), \quad j \in \{1, \dots, N\}, \quad \delta_n^{(j)} \overset{iid}{\sim} \chi_k^2, \tag{3.8}$$

  where the $\delta_n^{(j)}$ are independent of the $\theta^{(j)}$, and $k > 0$.
  This construction guarantees that $\mathbb{E}[T_j] = \mathbb{E}[\mathbb{E}[T_j | \theta^{(j)}] = \mathscr{I}$ regardless of the distribution

of $\theta^{(j)}$. The parameter k controls the concentration of $T^{(j)}$ around the identity.

In practice, we exploit the fact that $\{1, \sqrt{2}\sin(2\pi n x), \sqrt{2}\cos(2\pi n x) : n \in \mathbb{N}\}$ is an orthonormal basis for $L^2([0,1], \mathbb{R})$, and in particular

$$\sum_{n,m=1}^{N} \sqrt{2}\sin(2\pi n x) \cdot \sqrt{2}\sin(2\pi n x) = N.$$

The numerical model for the generation corresponds to

$$T^{(j)} = \sum_{n=1}^{N} \delta_n^{(j)} \sin(2n\pi t - \theta^{(j)}) \sin(2n\pi t - \theta^{(j)})$$

whose $[k, l]$ entry is

$$T^{(j)}[k, l] = \sum_{n=1}^{N} \delta_n^{(j)} \sin(2n\pi t[k] - \theta^{(j)}) \sin(2n\pi t[l] - \theta^{(j)}),$$

successively scaled by $2/N$ so that the mean of the $T^{(i)}$ converges to the identity. Notice that the mean of any given collection $T_1, \ldots, T_N$ will not be exactly the identity, but will be an approximation of it, if k and N are not too small. Moreover, since $\theta_i = 0$ would return us the commutative case, one can imagine that the θ_i serve as indicators on how far we are from this. On this note, a parametric model can be assumed for the θ_i. We chose the θ_i to be sampled from a von Mises distribution with mean 0 and measure of concentration $1/\sigma$, with the degenerate case of $\sigma \to \infty$ yielding commutativity.

After we generate the collection $T_1, \ldots, T_N$, we obtain the subsequent empirical covariances as $\hat{\Sigma}_j = T_j \bar{\Sigma} T_j^{\star}$. The "population" Fréchet mean $\bar{\Sigma}$ is inspired by the simulation scenario in Kashlak et al. [2016] and chosen to be a matrix with eigenvalue decay rate of $O(n^{-4})$:

$$\bar{\Sigma} = U \left[\sum_{n=1}^{\infty} n^{-4} \sin(2n\pi t) \otimes \sin(2n\pi t) \right] U^{*} \tag{3.9}$$

where U is a randomly-generated unitary operator.

- **Phoneme dataset and expanded Phoneme dataset.** This dataset includes the collection of over 4509 phonemes as in Ferraty and Vieu [2004], Hastie et al. [1995]. The file is available at

  https://web.stanford.edu/ hastie/ElemStatLearn/.

  The dataset consists of 4509 log-periodograms of length 256, each computed from continuous speech frames of 50 male speakers with known class (phoneme) memberships. Each speech frame is 32msec long, sampled at a rate of 16kHz and represents one of the following five phonemes: "aa" (as in "dark", nasal a), "ao" (as in "water"), "iy" (as in "she"), "sh" (as in "she"), "dcl" (as in "dark", "british" d). The dataset contains 256

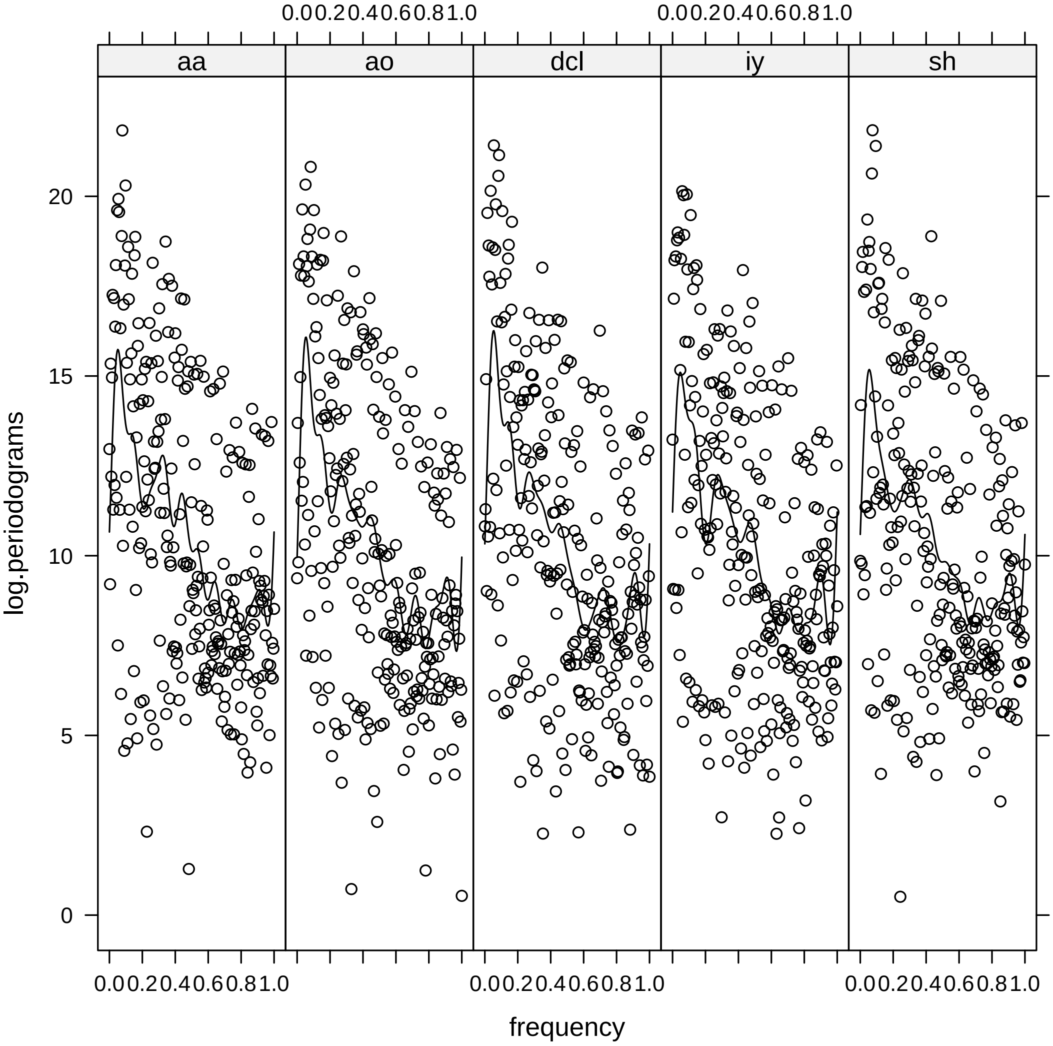


Figure 3.1: Original and smoothed log-periodograms for each phoneme

columns labelled $x.1$ to $x.256$, corresponding to the frequencies, a response column labelled "g", and a column labelled "speaker" identifying the different speakers.

The log-periodograms, which are quite noisy, are smoothed using a Fourier basis (21 basis functions) and digitalised to 256 equispaced frequencies, giving rise to a 4509×256 matrix. We perform the analysis of the collection of curves identified by the rows of this matrix.

Figure 3.1 shows the original and smoothed log-periodograms in the original dataset, while Figure 3.2 shows the mean log-periodograms. Estimated covariance operators for the 5 phonemes are shown in Figure 3.3. In order to produce a more realistic setup for classification and PCA, we consider a further scenario in which we artificially enlarge the `phonemes` dataset, and produce replicas of the 5 covariance operators. These replicas

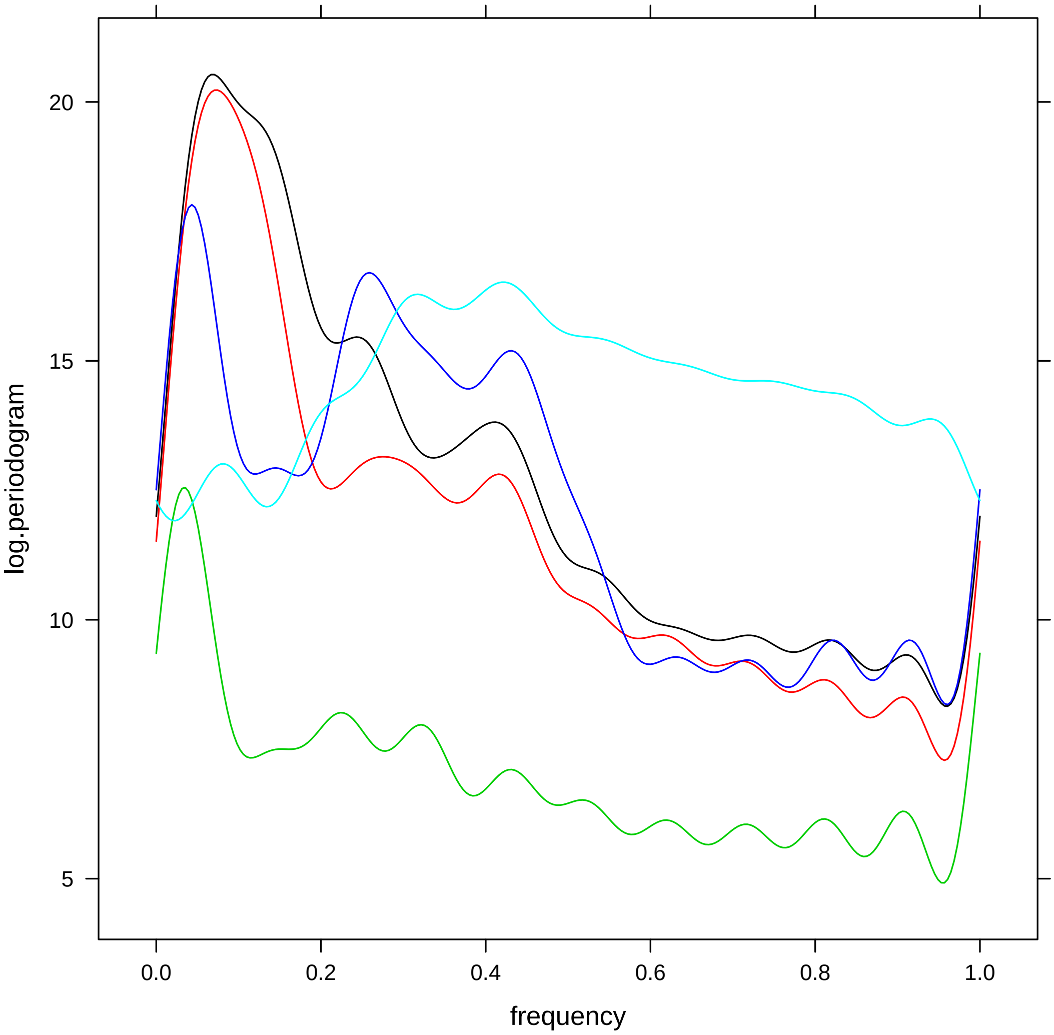


Figure 3.2: Mean Log-periodograms of the phonemes. Colours are as follows: black for “sh”, red for “iy”, green for “dcl”, blue for “aa”, cyan for “ao”.

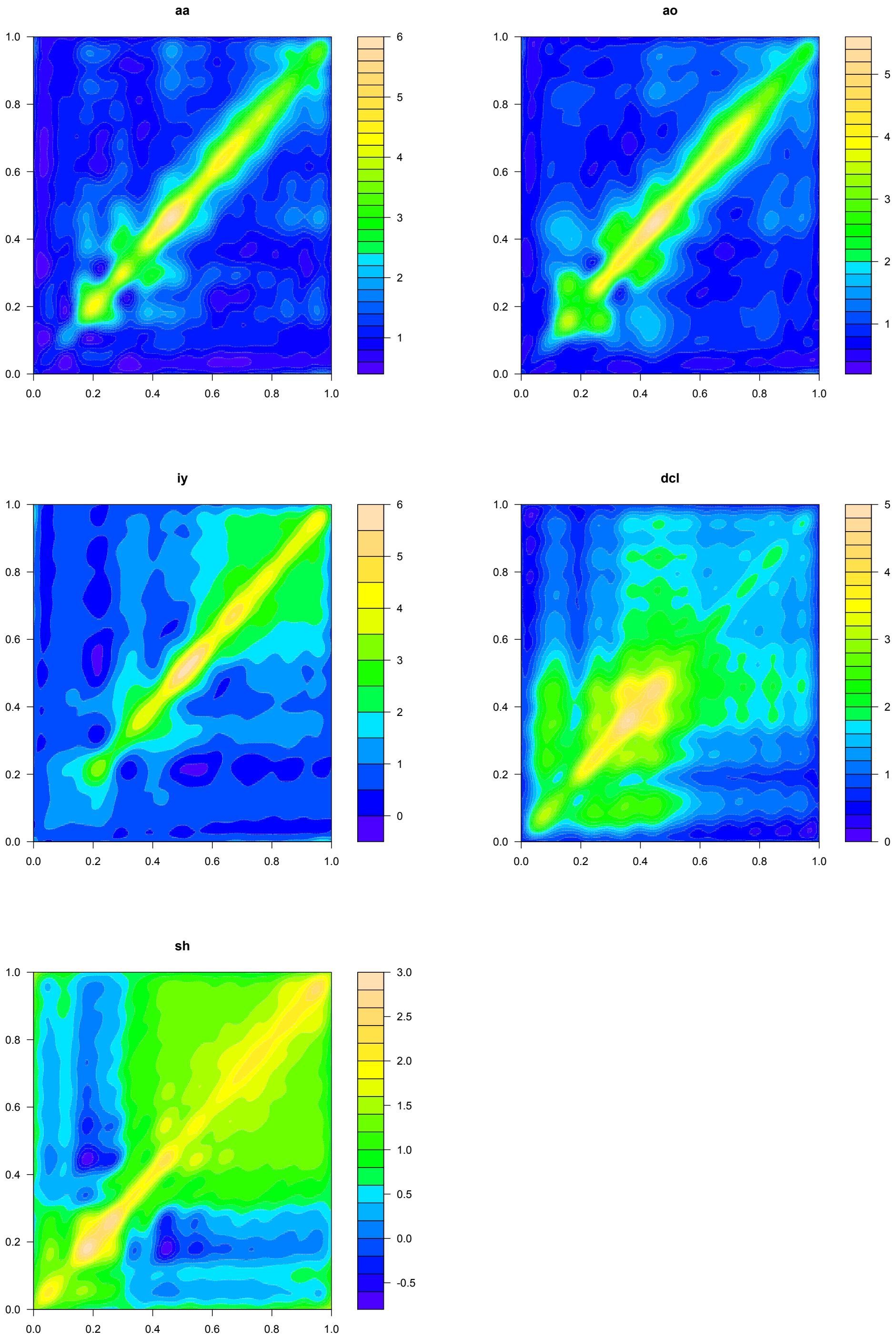


Figure 3.3: Estimated covariances of the log-periodograms in the five phoneme groups.

are estimated using only a subset of the whole collection of curves. Specifically, for each of the 5 phonemes, we estimate the covariance operator using 12 distinct subsamples of 50 log-periodograms, for a total count of 60 collections of 50 curves each. The estimation of the covariance operators for each of these subgroups yields 60 covariance operators, classified in 5 groups. We then carry out the analysis on these 60 covariances.

3.1.2 Numerical simulations

In this section, we compare our test with the pairwise k-sample permutation test of Cabassi et al. [2017] and with the concentration-based test in Kashlak et al. [2016]. To perform the two tests, we used the R-package `fdcov`. In the comparison, we have considered the version based on the Procrustes distance of the pairwise test.

Figures 3.4-3.7 show the empirical powers of the different procedures when applied to the Berkeley growth data, i.e., in the same simulation scenario considered by Cabassi et al. [2017] and described in the last Section.

For these particular simulations, we chose the dimensions of the covariance matrices to be 31×31. For each covariance, 20 curves are generated at each replication to estimate the empirical covariances on which the test is performed. These curves are sampled from mean-zero processes with the corresponding covariance. We carry out the test both on a Gaussian and on a t-Student distribution with 5 degrees of freedom. The number of permutation is 100. The power is estimated from a total of 1000 replications. The test statistics employ the Hilbert–Schmidt norm ($r = 2$ in Equation (3.5)). The probabilities of false positive (I type error) are estimated using all the available replications when $\gamma = 0$.

The x-axis in Figures 3.4-3.7 represents the value of the γ parameter (see (3.6) and (3.7)), while on the y-axes is displayed the empirical power. It is evident that our procedure is more powerful than the competitors. Moreover the plot shows that we achieve near perfect power, as opposed to the other tests that have nearly no power, for small values of γ, i.e. against local alternatives.

Furthermore, notice that without knowing the null distribution is not possible to use the calibration procedure of Kashlak et al. [2016], and that, under these circumstances, the concentration test is too conservative and does not respect the nominal level of 0.05 under H_0. However, Cabassi et al. [2017] show that their test outperforms that of Kashlak et al. [2016]. For this reason, in all the rest of this section we drop the comparison with Kashlak et al. [2016].

One might also wonder what happens when the intensity of the differences Δ_i's is measured with respect to the norm induced by the tangent space inner product $\langle\cdot,\cdot\rangle_\Sigma$ at $\bar{\Sigma}$ (cf. Section 1.4.3):

$$\langle A, B\rangle_\Sigma = \mathrm{tr}(A^*\Sigma B).$$

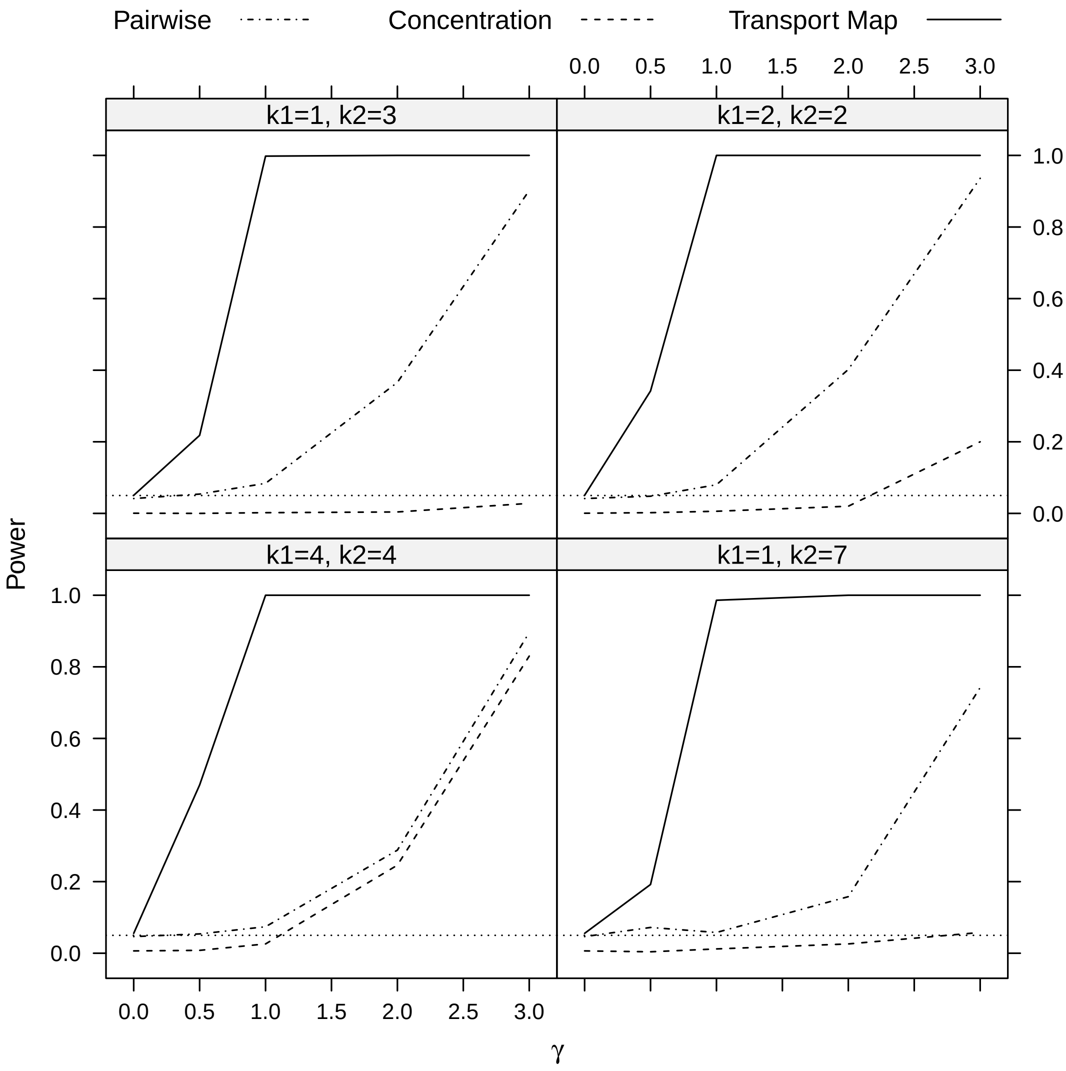


Figure 3.4: Comparison of our method against the Pairwise test of Cabassi et al. [2017] and the Concentration-based test of Kashlak et al. [2016] with Gaussian data and geodesic perturbations. Dotted horizontal line gives the nominal level under H_0.

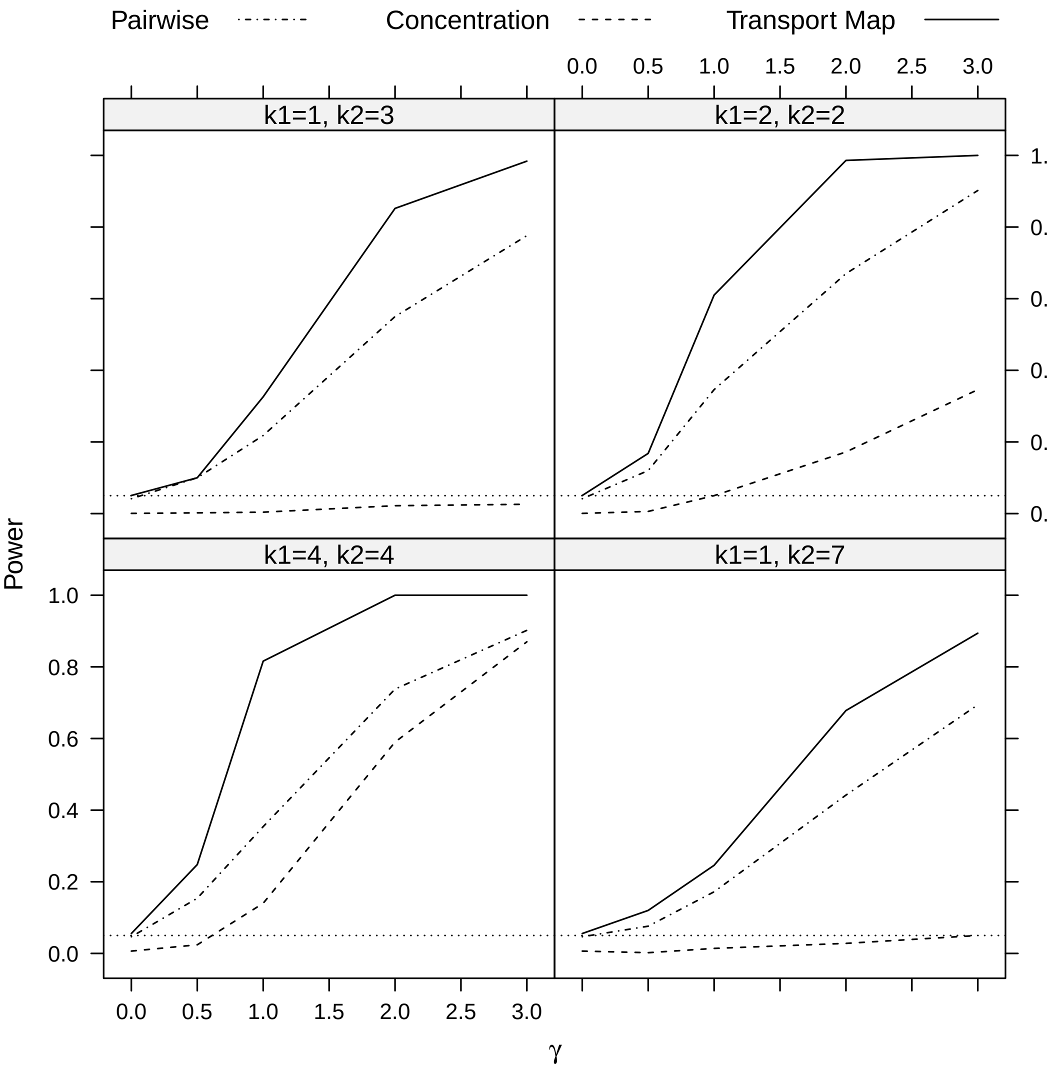


Figure 3.5: Comparison of our method against the Pairwise test of Cabassi et al. [2017] and the Concentration-based test of Kashlak et al. [2016] with Gaussian data and additive perturbations. Dotted horizontal line gives the nominal level under H_0.

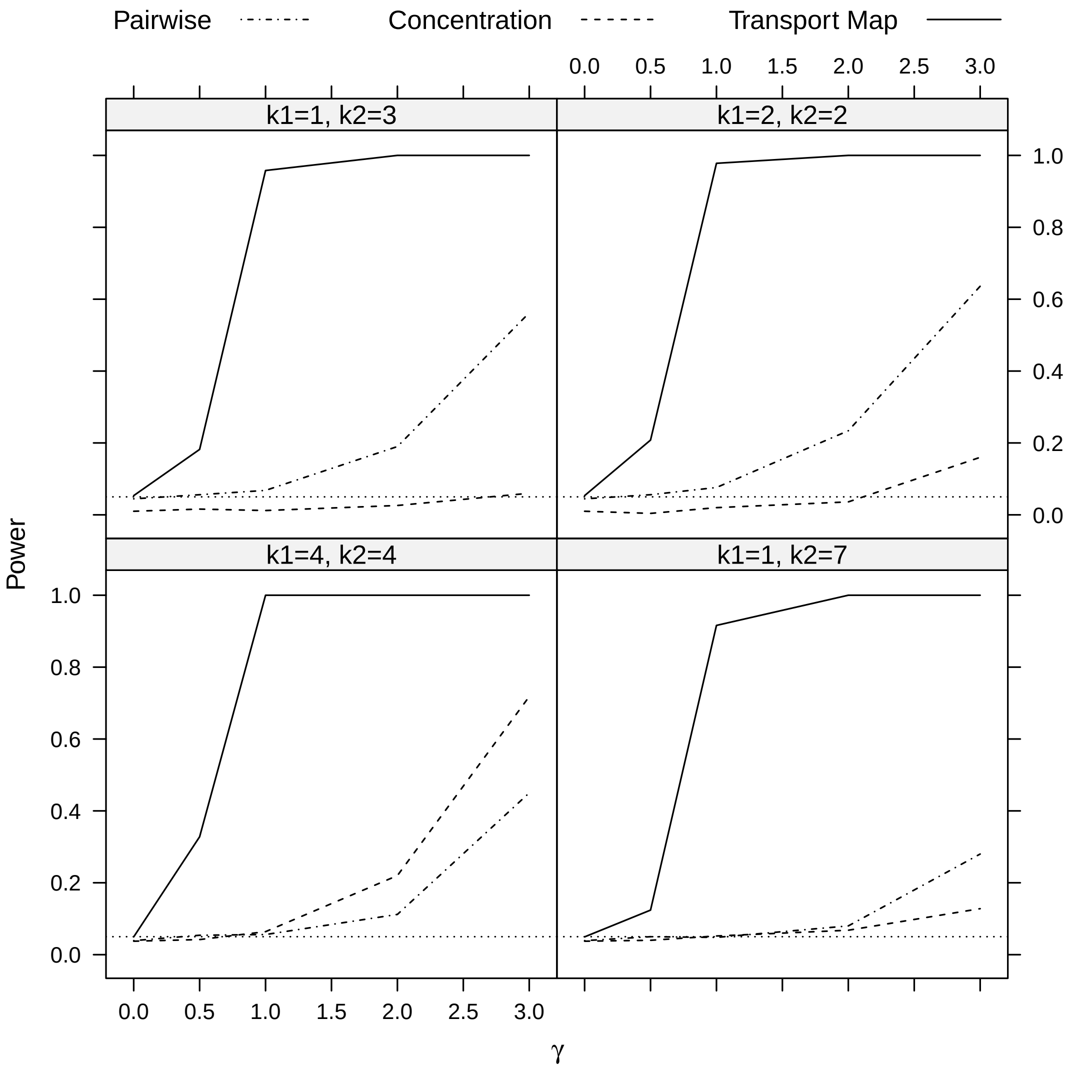


Figure 3.6: Comparison of our method against the Pairwise test of Cabassi et al. [2017] and the Concentration-based test of Kashlak et al. [2016] with t-Student data and geodesic perturbations. Dotted horizontal line gives the nominal level under H_0.

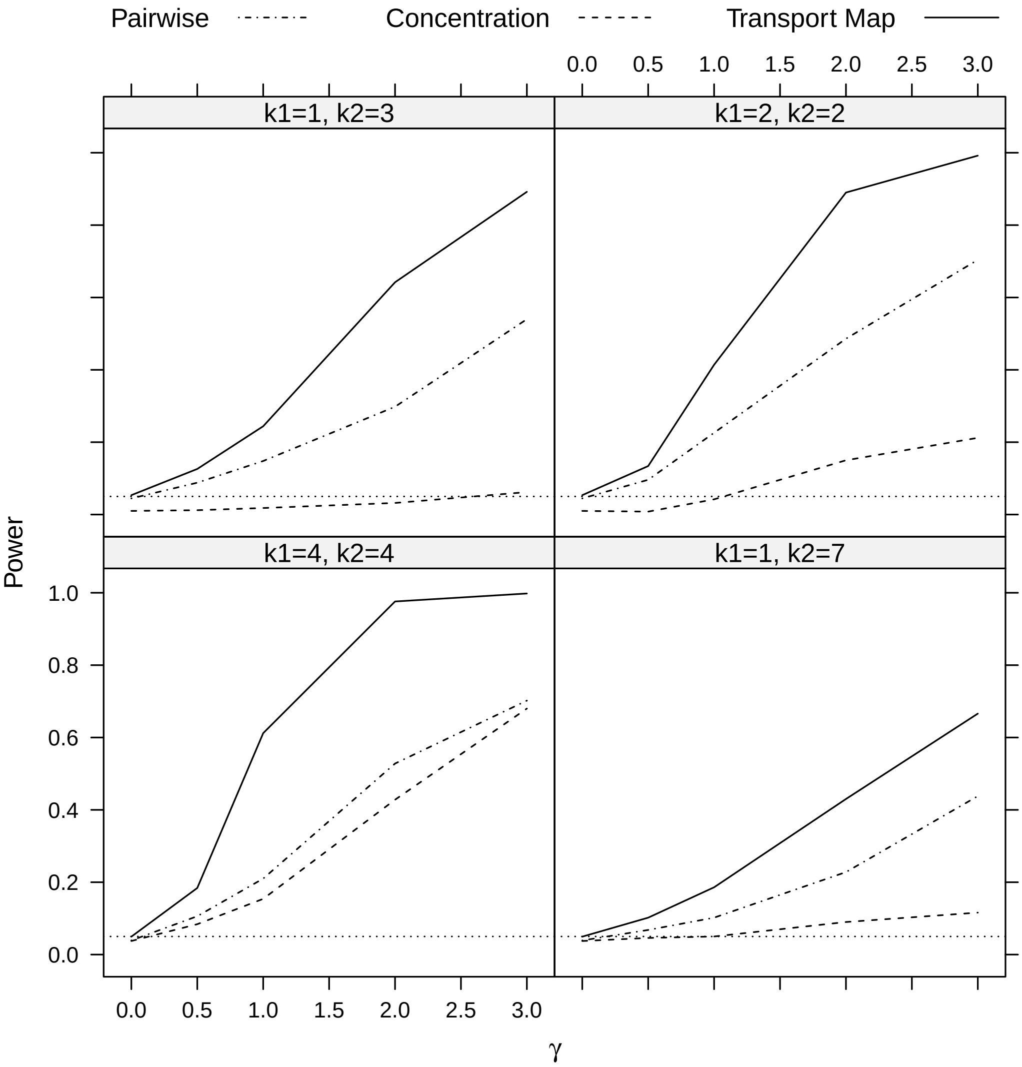
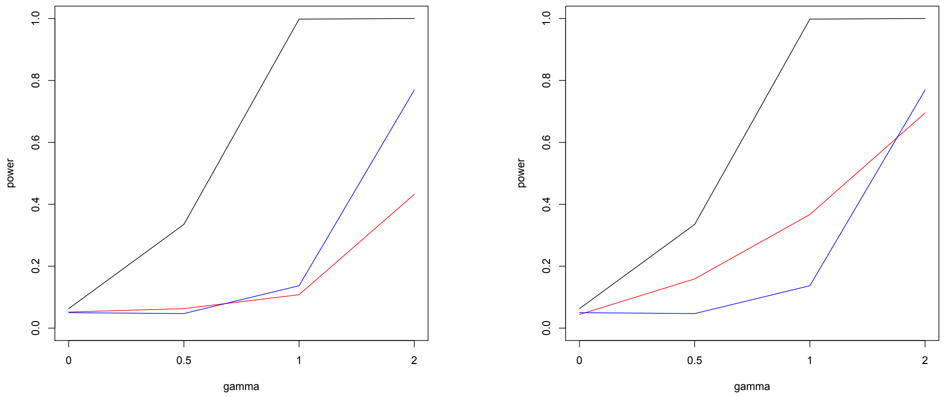


Figure 3.7: Comparison of our method against the Pairwise test of Cabassi et al. [2017] and the Concentration-based test of Kashlak et al. [2016] with t-Student data and additive perturbations. Dotted horizontal line gives the nominal level under H_0.



(a) Geodesic perturbations

(b) Additive perturbations

Figure 3.8: Empirical power comparison, in Berkeley growth data scenario, for the transport functional ANOVA test, in the case where the maps are contrasted using the tangent space inner product rather than the Hilbert–Schmidt one. The black line corresponds to the transport-based ANOVA, the red line is the pairwise test from Cabassi et al. [2017] and blue line is the transport based ANOVA with respect to the tangent space metric. Power is estimated via permutations.

This would imply that instead of computing

$$\mathrm{tr}(T-\mathfrak{I})^*(T-\mathfrak{I})$$

we would compute

$$\mathrm{tr}(T_i-\mathfrak{I})^*\bar{\Sigma}(T_i-\mathfrak{I}),$$

the latter equation yielding precisely the Procrustes distance between T and $\bar{\Sigma}$ i.e., it corresponds to the testing method of Pigoli et al. [2014a] (at least in the 2-sample case).

Figure 3.8 shows that the functional ANOVA picks up power when the comparison is made with respect to the Hilbert–Schmidt norm rather that the tangent space one. This is to be expected, as our test is intrinsically functional (see Section 3.1.3 later) and is particularly responsive to high-frequency departures from the null. Considering the tangent space inner product corresponds to pre- and post-multiplying the Δ_j by $\bar{\Sigma}^{1/2}$, and this might cancel the difference in the tails of the operators.

Ultimately our goal in testing is to find a way to compare operators, and our testing procedure does not need to directly make use of the tangent space geometry. Rather, it uses the notion of multi-transport and that of transport maps.

In Section 3.2 and Section 3.3, we will see how the use of the tangent space inner product will be of crucial importance for a different purpose, since both PCA and clustering are intrinsically geometrical methodology.

Finally, equation (3.5) in Section 3.1 suggests that we can use yet another different norm to contrast the Δ_j to zero. One might expect that the stronger the norm is, the more it will detect finer differences between the tail of the spectra of the operators.
We ran simulations to this purpose on the Berkeley growth data. Using a stronger norm did not bring a significant improvement to the power of the test. This is consistent with the intrinsic finite dimensionality of the dataset and the relatively small dimension of the matrices, as in finite dimensions all norms are equivalent. A more significant difference is manifested when we apply the method on the generative model (Section 3.1.3), as we could better emulate a purely functional setting.

Figure 3.9 shows the results for the geodesic perturbations, and in Figure 3.10 for the additive ones. As the differences are not so visible from the pictures, the power values are also collected into Tables 3.2 and 3.1.
Since using different norms does not show a significant difference in practice, the next comparisons under this scenario will be limited to the Hilbert–Schmidt norm, as it lies "in between" the trace-class and the operator norm, as well as allowing an easier parallel with the Euclidean literature.

	γ=0	γ=0.5	γ=1	γ=2
k1=1, k2=2				
Trace norm	0.04	0.24	0.99	1.00
HS norm	0.04	0.23	0.99	1.00
Operator norm	0.04	0.22	0.99	1.00
k1=4, k2=4				
Trace norm	0.04	0.24	0.99	1.00
HS norm	0.04	0.23	0.99	1.00
Operator norm	0.04	0.22	0.99	1.00
k1=1, k2=3				
Trace norm	0.04	0.24	0.99	1.00
HS norm	0.04	0.23	0.99	1.00
Operator norm	0.04	0.22	0.99	1.00
k1=1, k2=7				
Trace norm	0.04	0.24	0.99	1.00
HS norm	0.04	0.23	0.99	1.00
Operator norm	0.04	0.22	0.99	1.00

Table 3.1: Comparison of the three different norms, geodesic perturbations, Berkeley growth data.

3.1.3 Comparison on generative model

We now apply our testing procedure on the generative model as described in Section 3.1.1 and 2.4. Recall that the generative model states that if a collection of nonnegative maps $T_1, \dots, T_N$ has mean identity, then any covariance operator Σ is the Fréchet mean of $\{\Sigma_j = T_j \Sigma T_j\}_{j=1}^N$,

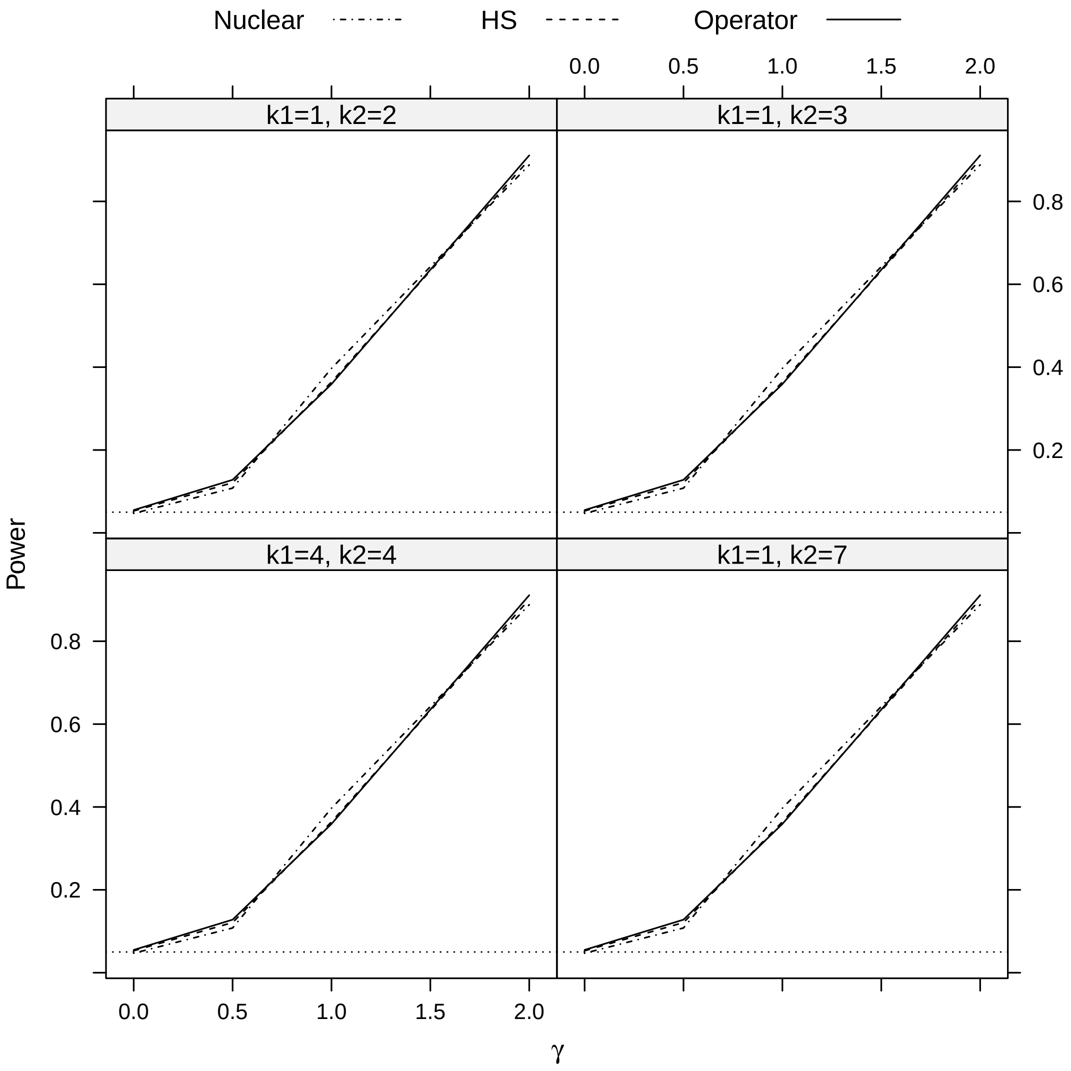


Figure 3.9: Comparison of empirical powers computed with respect to the the three different norms (Nuclear, Hilbert–Schmidt and operator) with geodesic perturbations using the Berkeley growth data. Dotted horizontal line gives the nominal level under H_0.

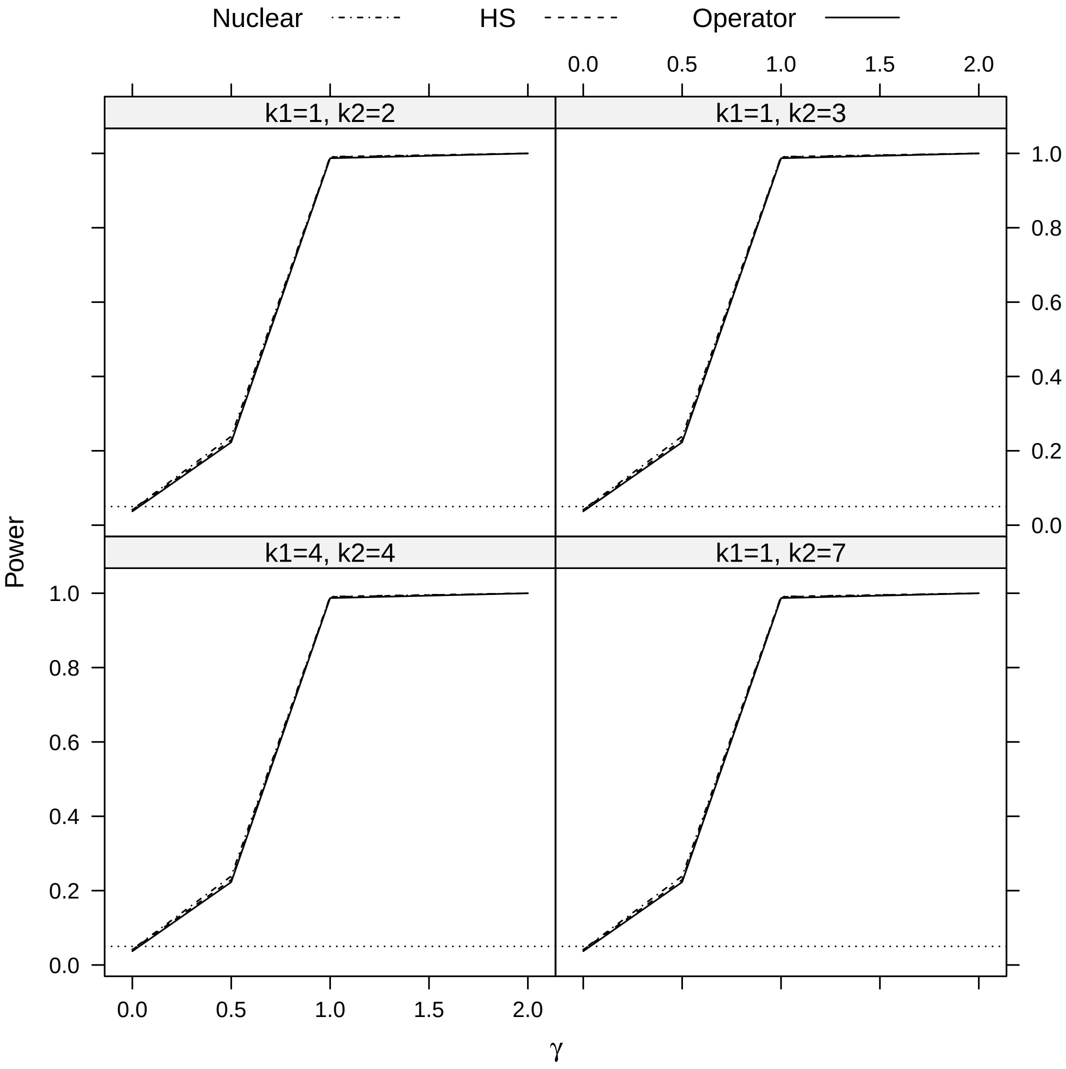


Figure 3.10: Comparison of empirical powers computed with respect to the the three different norms (Nuclear, Hilbert–Schmidt and operator) with additive perturbations using the Berkeley growth data. Dotted horizontal line gives the nominal level under H_0.

	γ=0	γ=0.5	γ=1	γ=2
k1=1, k2=2				
Trace norm	0.05	0.11	0.40	0.89
HS norm	0.05	0.12	0.36	0.90
Operator norm	0.06	0.13	0.36	0.91
k1=4, k2=4				
Trace norm	0.05	0.11	0.40	0.89
HS norm	0.05	0.12	0.36	0.90
Operator norm	0.06	0.13	0.36	0.91
k1=1, k2=3				
Trace norm	0.05	0.11	0.40	0.89
HS norm	0.05	0.12	0.36	0.90
Operator norm	0.06	0.13	0.36	0.91
k1=1, k2=7				
Trace norm	0.05	0.11	0.40	0.89
HS norm	0.05	0.12	0.36	0.90
Operator norm	0.06	0.13	0.36	0.91

Table 3.2: Comparison of the three different norms, additive perturbations, Berkeley growth data.

and the maps $\mathbf{t}_j$ in (3.10) must equal T_j (on the closed range of Σ).

Figure 3.11 shows the result of the test. The x-axis represents the value of the von Mises parameter σ. To simulate a functional case, we increased the size of the matrices with respect to the scenario on the Berkeley growth data (where the dimension was 31×31), and we have chosen for the matrices to have size 70×70. The power is estimated from 1000 replications. The number of permutations is 100. At each replication, we generate two optimal maps T_1 and T_2 via the generative model, and two corresponding covariances $\Sigma_1 = T_1\Sigma T_1$ and $\Sigma_2 = T_2\Sigma T_2$, with Σ given by equation (3.9). For each Σ_i, $i = 1,2$, we sample 50 observations of a Gaussian process with mean-zero and covariance Σ_i. The empirical covariance computed from these observations will yield a replica of Σ_i. We repeat this as to obtain k_1 replicas of Σ_1, and k_2 replicas of Σ_2, for a total of $k_1 + k_2$ covariances divided into two groups of size k_1 and k_2 respectively. The values of the pair (k_1, k_2) are (1,2), (1,3), (1,7) and (4,4). This procedure is repeated for several values of the Von Mises parameter σ, namely $\sigma = (0.1, 1, 15)$. Recall that σ can be seen as an inverse variance, therefore a smaller value of σ will imply a larger dispersion of the generated optimal maps T_i's.

From Figure 3.11 we can see the our test still outperforms the test of Cabassi et al. [2017], albeit by less. This is consistent with all tests seeming to be more powerful, and in fact achieving near perfect power, than in the Berkeley growth data scenario. Indeed, the generative model produces matrices which are better separated among each others (in contrast, for example, with data generated for a small value of γ in the Berkeley case), therefore allowing for a better discrimination against the null. Finally, notice how the dimension of the matrices allows us to visualize a difference between the norms employed in the test. The operator norm appears to

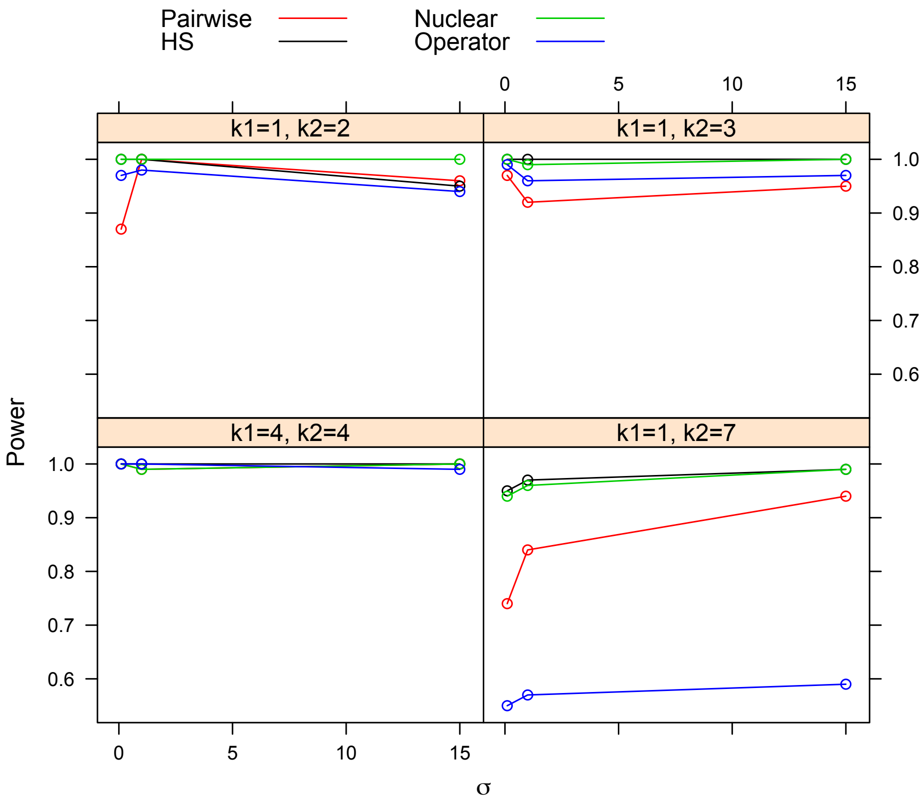


Figure 3.11: Empirical power of our and Cabassi et al. [2017]'s tests as a function of the dispersion parameter σ of the von Mises distribution in the generative model of Equation (3.8). Our test is performed using three different norms: trace-class, Hilbert–Schmidt and operator norm.

be the weakest, as expected, and this is particularly evident in the "difficult" case where most of the matrices are equal.

A genuinely functional test

There is an important observation to be made about the transport-based ANOVA. Specifically, our test owes its power to the intrinsic functional nature of the data. By considering the maximum distance, the procedure of Cabassi et al. [2017] is less sensitive to differences in the tails of the spectra of the covariances, whereas our procedure is able to detect departures from the null even when they only occur at very high frequencies. In turn, if the data are truncated, or corrupted by noise, this reflects on the test performance.

The smoothness of the functions is intrinsically connected with the spectral decomposition of the covariance operators. In applications, we smooth functional data by expressing them as a linear combinations of basis functions. For clarification, we can keep the notation of Ramsay and Silverman [2005b] and write the expansion into basis functions in the form

$X(t) = \sum_{k=1}^{K} c_k \phi_k$, where the c_k are a vector of weights and the ϕ_k are basis elements (cf. Ramsay and Silverman [2005b, Section 3.3]. See also Section 1.1.3). The degree to which the data are smoothed (as opposed to interpolated) is regulated by the number K of basis functions. If the data functions are sufficiently smooth, we can assume that they can be expressed as a linear combination of a limited number of basis elements.

From an implementation point of view, the transport-based ANOVA compute an estimate of the transport maps $\mathbf{t}_j$ going from a Fréchet mean to the data curves at every replication. The level of smoothness of the curves affects as well the number of observations needed to estimate the t_j, and in particular, needed to achieve a high power of the test. When we contrast the $\Delta_i = \mathbf{t}_j - \mathfrak{I}$ to 0, we are essentially contrasting the eigenvalues of $\mathbf{t}_j$ to 1. Therefore, we need for the number of n observations from which the sample of $\hat{\Sigma}_j$'s is estimated, to be greater than their rank[1], or better, that our curves are in fact linear combinations of a smaller number of curves. Simulations showed that we need to have at least about 20 observations more of the number of significant eigenvalues of the $\mathbf{t_j}$.

We tested our transportation ANOVA in a scenario which resembles a multivariate case, rather than a functional one. Differently than in the previous paragraph, the "true barycenter" of the model $\bar{\Sigma}^*$ is taken to be a randomly generated Wishart distribution with $d = 70$ degrees of freedom. As such, $\bar{\Sigma}^*$ (and consequently corresponding sample $\Sigma_1, \ldots, \Sigma_N$) will not display fast eigenvalue decay as that in Equation (3.9), and which is characterising of trace-class operators. Figure 3.12 compares the transport ANOVA with the k-sample test of Cabassi et al. [2017] under the generative model with true barycenter $\bar{\Sigma}^*$. The power was estimated via 1000 replications. The number of permutations is 100. The procedure follows step-by-step the one described above for the generative model: the dimension of the matrices is 70×70, and once we obtained Σ_1 and Σ_2 via the generative model, we estimate, for each $j = 1, 2$, k_j replicas from 50 observations from a mean-zero Gaussian process with covariance Σ_j. Notice how the number of eigenvalues of the Σ_j which are significantly different from 0 is exactly equal to their dimension. Therefore, the number of curves we use to estimate the replicas of Σ_j is *less* than their number of significant eigenvalues.

Figure 3.12 shows how in this case, the permutation test of Cabassi et al. [2017] outperforms the transport-based ANOVA. This effect is less visible when for the test carried out with respect to the operator norm, which, similarly to the one of Cabassi et al. [2017], is based on a maximum distance.

3.1.4 Data Analysis

In order to perform ANOVA on the phoneme dataset, we extract the log-periodograms corresponding to the phonemes "aa", "ao", "iy" which are similar and, hence, hard to distinguish. We performed the test in other situations as well: Table 3.4 shows the results for the phonemes

---

[1] Here "rank" is not meant in the formal sense, rather in an empirical one, i.e. the number of eigenvalues detectably different from 0.

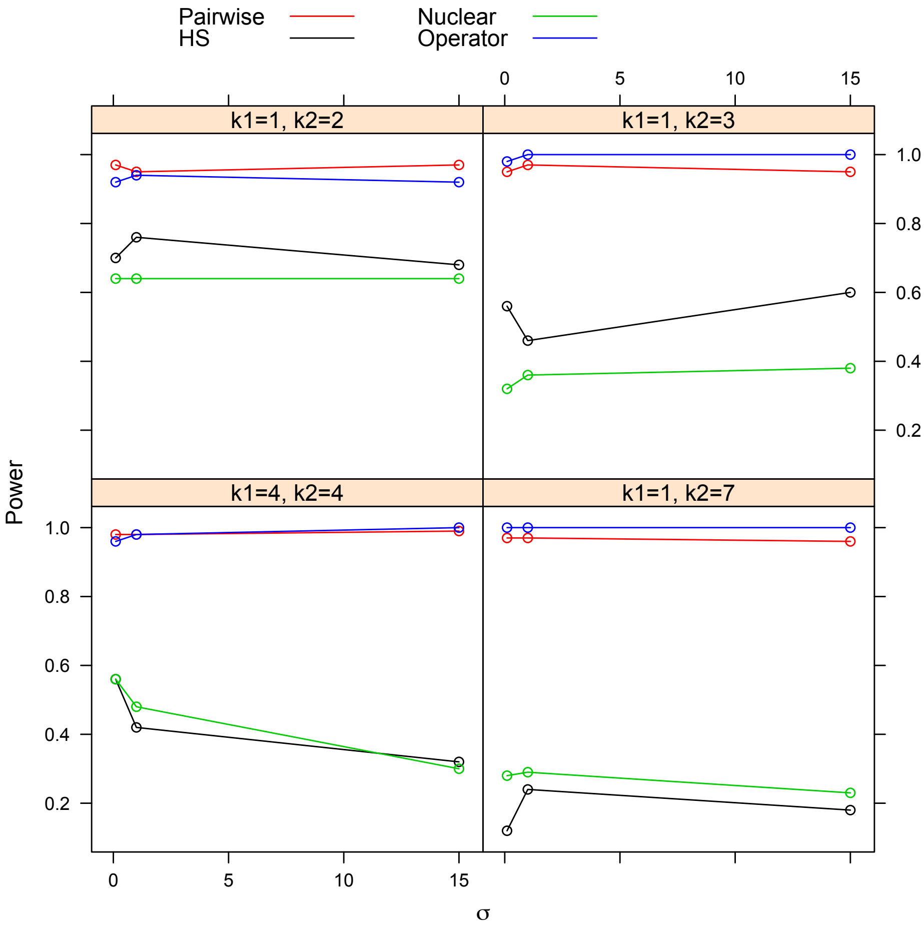


Figure 3.12: Empirical power of our and Cabassi et al. [2017]'s tests as a function of the dispersion parameter σ of the Von Mises distribution in the generative model of Equation (3.8) (3.8). Our test is performed under three different norms: trace-class, Hilbert–Schmidt and operator norm. The origin for the generative model is taken to be a Wishart distribution with 70 degrees of freedom.

	n	Pairwise	Concentration	Transport Maps
H_0	25	0.086	0.000	0.068
	50	0.050	0.000	0.046
H_1	25	0.271	0.300	0.470
	50	0.670	0.944	0.994

Table 3.3: Comparison of the empirical power of the three different testing methods on the `phoneme` dataset, when applied on the phonemes "aa", "ao" and "iy".

	n	Pairwise	Concentration	Transport Maps
H_0	25	0.100	0.000	0.062
	50	0.040	0.00	0.061
H_1	25	0.870	1.000	1.000
	50	0.990	1.000	1.000

Table 3.4: Comparison of the empirical power of the three different testing methods on the `phoneme` dataset, when applied on the phonemes

"sh","dcl","iy", but as we see, when we consider phonemes which are very different among them (such as vowels and consonants) all tests have very high power. Considering other combinations of phonemes gave similar results. Therefore, we decided to limit ourselves to the phonemes "aa", "ao", "iy", as to be able to discriminate better among the different test procedures.

To sample under H_0, we sample $3n$ log-periodograms for the "iy" phoneme which are then randomly assigned to three groups, each of size n. To sample under the alternative H_1, we sample n log-periodograms for each phoneme. We repeat the test for $n = 25$ and $n = 50$, and for 500 replications and 200 permutations. Again we compare both with Cabassi et al. [2017] and Kashlak et al. [2016][2].

Results of power computations are given in Table 3.3. Since the different phonemes have different mean functions (see Figure 3.2), observations must be centered around the sample mean of the groups, before computing the p-value using the permutation approach. Therefore, in this case, it is not guaranteed that Cabassi et al. [2017]'s pairwise test and ours respect the right type I error probability under H_0. Regardless, the transport test delivers a level very close to the nominal 0.05, especially when $n = 50$ (which is still relatively low compared to the 256 points where the curves are sampled). Our test results also more powerful under the alternative hypothesis.

Table 3.4 shows the results for the phonemes "sh","dcl","iy", but as we see, when we consider phonemes which are very different among them (such as vowels and consonants) all tests have very high powers. Considering other combinations of phonemes gave similar results.

---

[2] When interpreting the results, it is important to treat the outputs carefully, since the procedure of Kashlak et al. [2016] was unable to produce a result in a small number of cases, as the computation of the distance using SVD failed.

3.2 Transportation-based functional PCA of covariances

If the null hypothesis of equality among covariances is rejected, a second order analysis would entail explaining and understanding the variability of the sample of covariances around its Fréchet mean, and possibly interpreting the main directions of this variation.
As in multivariate analysis in Euclidean spaces (Jolliffe [2002]), Principal Component Analysis (PCA) is a good candidate for such tasks. For functional data, this has been extended to functional PCA (fPCA) by Grenander [1950a], Karhunen [1947], Rao [1958], Ibazizen and Dauxois [2003]. See also Panaretos and Tavakoli [2013]). A first coarse measure of variance is readily provided by the minimal value of the Fréchet functional, but PCA provides both a dimensionality reduction tool and a means to explain the principal modes of variation of a random vector.
One way of carrying out fPCA in Hilbert spaces is based on the eigenstructure on the covariance operator, by analogy to PCA in finite dimensions. More specifically, it relies on the representation of the data curves given by their Karhunen–Loève expansion [Ramsay and Silverman, 2005a]. When the space is non-Euclidean, as in the case of a sample of covariances, one way to carry out PCA is by lifting the analysis on the tangent space. In finite dimension, this is known as tangent space PCA (see Huckemann et al. [2010], Fletcher et al. [2004], Dryden et al. [2009]).
To the best of our knowledge, there are no results concerning PCA on covariance operators that account for the non-linear nature of the space of these operators.

The tangent space provides a local linear approximation to the curved space of covariances. Once the covariances are mapped onto the tangent space through the log map (Section 1.4.2), they can be uniquely identified with a tangent vector that belongs to a linear space, and therefore, a standard linear (functional) PCA can be carried out. The principle behind tangent space PCA works both in finite and in infinite dimension. However the dimensionality of the space raises different issues, depending on whether it is finite or not. A finite-dimensional tangent space approximation is mostly local, in the sense that every small enough neighbour of a point p on a manifold M of dimension n can be approximated by a copy of $\mathbb{R}^n$ and the log-map is defined locally (see e.g. Lang [2012]). This means that the tangent space definition depends on the choice of the point where is computed at, and a natural choice is the Fréchet mean. Therefore the feasibility and the quality of the approximation depend on how much the observations are spread around their mean.

In infinite dimensions on the other hand, the log-map might not even be defined. However, if the log-map *is* well-defined, then PCA can be performed globally on the space of covariance operators with the Procrustes–Wasserstein distance without any neighbouring restriction, since, as described in Section 1.4.2, the exponential map is surjective. In Section 1.4.2 it was established that for a collection of covariance operators $\Sigma_1, \dots, \Sigma_n$ with Fréchet mean $\bar{\Sigma}$, the log-maps $\log_{\bar{\Sigma}}(\Sigma_i)$ are well-defined only if $\bar{\Sigma}$ is "more injective" than the observations, that is, if $\ker(\bar{\Sigma}) \subseteq \ker(\Sigma_i)$.
Injectivity of the $\bar{\Sigma}$ is not guaranteed and it is conjectured in Conjecture 17 of Masarotto et al.

[2018] (see also Chapter 4). However, thanks to the recent result by Masarotto et al. [2019] and Theorem 23, we can entirely bypass the injectivity issue, and recover well-defined log maps at the Fréchet mean.

Indeed we know from Section 1.4.2 the log-map which lifts the observation $\Sigma_1, \dots, \Sigma_N$ at the tangent space at $\bar{\Sigma}$ is given by

$$\log_{\bar{\Sigma}}(\Sigma_i) = \mathbf{t}_{\overline{\Sigma}}^{\Sigma_i} - \mathfrak{I} = \overline{\Sigma}^{-1/2}[\overline{\Sigma}^{1/2}\Sigma_i\overline{\Sigma}^{1/2}]^{1/2}\overline{\Sigma}^{-1/2} - \mathfrak{I} = \mathbf{t}_j - \mathfrak{I},$$

that is, is achieved exactly as the differences $\Delta_j = \mathbf{t_i} - \mathfrak{I}$, which we know by Theorem 23 to exist and be bounded linear operators.

Now if $\langle , \rangle_{\bar{\Sigma}}$ denotes the tangent space inner product at Σ (Section 1.4.2)

$$\langle A, B \rangle_\Sigma = \text{trace}(A \Sigma B),$$

the span of $\{\Delta_1, \dots, \Delta_n\}$ equipped with $\langle , \rangle_{\bar{\Sigma}}$ has a Hilbert-space structure as shown in Masarotto et al. [2019], since

$$\text{trace}(\Delta_i \Sigma \Delta_j) \leq |||\Sigma^{1/2}\Delta_i|||_2 |||\Sigma^{1/2}\Delta_j|||_2 = \Pi(\Sigma_i, \Sigma)\Pi(\Sigma_j, \Sigma) < \infty.$$

All this implies that we can carry out a tangent space PCA in the following way:

1. Compute the Fréchet mean using the algorithm in Section 2.3

   $$\overline{\Sigma} = \arg\min_{\Sigma} \sum_{j=1}^{n} \Pi^2(\Sigma, \Sigma_j)$$

2. Use the log maps to lift $\Sigma_1, \dots, \Sigma_N$ to their respective correspondents

   $$\Delta_j = \log_{\bar{\Sigma}}(\Sigma_i) = \mathbf{t}_{\overline{\Sigma}}^{\Sigma_i} - \mathfrak{I}$$

   on the tangent space at the Fréchet mean.

3. Perform linear PCA of the tangent vectors Δ_j.

4. Retract the resulting components V_j onto the manifold via the exponential map

   $$\exp_{\bar{\Sigma}} = (V_j + \mathfrak{I})\bar{\Sigma}_j(V_j + \mathfrak{I}).$$

As in practice we are only given the empirical version $\widehat{\Sigma}_1, \dots, \widehat{\Sigma}_N$ of $\Sigma_1, \dots, \Sigma_N$, the procedure just described will be applied to $\widehat{\Sigma}_1, \dots, \widehat{\Sigma}_N$. Moreover, we will see in the next section how some extra care is needed when performing PCA on the collection $\widehat{\Delta}_1, \dots, \widehat{\Delta}_N$, due to the fact that the inner product of the tangent space is not the standard Hilbert–Schmidt inner product.

Point 3. above relies on the spectral decomposition of the covariance operator for the tangent space points $\{\Delta_1, \dots, \Delta_N\}$, which is the non-negative operator

$$\mathcal{K} = \frac{1}{N} \sum_{j=1}^{N} \Delta_j \otimes_{\bar{\Sigma}} \Delta_j = \frac{1}{N} \sum_{j=1}^{N} \left(\mathbf{t}_j - \mathcal{I}\right) \otimes_{\bar{\Sigma}} \left(\mathbf{t}_j - \mathcal{I}\right),$$

with $(A \otimes_{\bar{\Sigma}} B)C = \langle B, C\rangle_{\bar{\Sigma}} A$. The Principal Components (PCs) we obtain via the spectral decomposition of $\mathcal{K}$ can be seen as a traditional (Riemannian) tangent space PCA, with respect to the *Procrustean metric tensor*

$$\langle Q_1, Q_2\rangle_\Gamma = \operatorname{trace}(Q_1 \Gamma Q_2), \quad \Gamma \in \mathcal{L}$$

over the barycentric locus $\mathcal{L}$ of the operators $\{\Sigma_j\}_{j=1}^N$, that is the locus of points which are weighted means of N reference points,

$$\mathcal{L} = \left\{ \underset{\Gamma \geq 0}{\arg\min} \sum_{j=1}^{N} \alpha_j \Pi^2(\Sigma_j, \Gamma) : \alpha_j > 0 \,\&\, \sum_{j=1}^{N} \alpha_j = 1 \right\}.$$

The operator $\mathcal{K}$ constitutes exactly the empirical operator of the collection of differences $\{\Delta_i\}_{i=1}^N$, since by (3.3) the transport maps

$$\mathbf{t}_j = \bar{\Sigma}^{-1/2} (\bar{\Sigma}^{1/2} \Sigma_i \bar{\Sigma}^{1/2})^{1/2} \bar{\Sigma}^{-1/2}, \quad i = 1, \dots, N, \tag{3.10}$$

are centered around the identity. In particular

$$\sum_{i=1}^{N} \Delta_i = 0.$$

The retraction of the principal components onto the manifold will give principal geodesics that describe the main directions of variation of the data on the manifold. More precisely, if E_1 is the eigen-operator associated with the largest eigenvalue of $\mathcal{K}$, the retracted curve

$$t \mapsto (\mathcal{I} + tE_1)\Sigma(\mathcal{I} + tE_1), \qquad t \in [-\epsilon, \epsilon],$$

is a geodesic for sufficiently small $\epsilon > 0$. This principal geodesic is the visualisation of the main mode of variation of $\{\Sigma_j\}_{j=1}^N$ near their Fréchet mean Σ. Section 3.2.3 provides a visualisation of the variation along the principal geodesics for the phoneme dataset.

3.2.1 PCA under the tangent space inner product

In the previous section we stated that once the data are lifted to the tangent space, linear PCA can be performed. In reality the analysis of principal component is dependent upon the geometry, thus the inner product, and developing a rigorous tangent space PCA requires some extra care. A different choice of inner product rather than the standard Hilbert–Schmidt one affects both the implementation (through the function smoothing) and the interpretation of the PCs. Specifically, note that when employing the tangent space inner product inducing the Wasserstein distance (cf. Section 1.4.3), the maximisation problem yielding the first principal component takes the form

$$\underset{\|B\|_\Sigma=1}{\arg\max}\langle \mathcal{K}B, B\rangle_\Sigma = \underset{\|\Sigma^{1/2}A\|_2=1}{\arg\max}\ \langle \mathcal{K}\Sigma^{1/2}A, \Sigma^{1/2}A\rangle_2 = \underset{\operatorname{trace}(A\Sigma A)=1}{\arg\max}\ \operatorname{trace}(\mathcal{K}\Sigma^{1/2}A^2\Sigma^{1/2}),$$

which is a non-standard version of the maximisation problem. Nevertheless, it can be shown that PCA with respect to the tangent space inner product is equivalent to the PCA performed with the Hilbert-Schmidt inner product on suitably transformed data. This problem was first treated by Silverman et al. [1996] in the case of Sobolev inner products, and then generalised by Ocaña et al. [1999]. The next paragraph recaps the results from the latter.

Let $\widehat{\Sigma}_1, \dots, \widehat{\Sigma}_N$ be a collection of covariance operators with empirical Fréchet mean $\overline{\Sigma}$, and let $\langle\cdot,\cdot\rangle_{HS}$ be the Hilbert–Schmidt inner product. Assume we want to perform PCA under the Wasserstein tangent space inner product $\langle\cdot,\cdot\rangle_\Sigma$ at $\bar{\Sigma}$, $\langle A, B\rangle_\Sigma = \operatorname{tr}(A\Sigma B)$. By Ocaña et al. [1999], if $\langle\cdot,\cdot\rangle_\Sigma$ is continuous for $\langle\cdot,\cdot\rangle_{HS}$ then there exists a unique operator $\mathcal{T}$ characterised by

$$\langle A, B\rangle_\Sigma = \langle \mathcal{T}(A), B\rangle_{HS} = \operatorname{tr}([\mathcal{T}(A)]^* B).$$

In our case, $\mathcal{T}$ is the multiplication from the right by Σ, so

$$\mathcal{T}(A) = (A\Sigma^{1/2})\Sigma^{1/2}$$

which is trace class and has an adjoint. It follows from Ocaña et al. [1999], that $\mathcal{T}$ is nonnegative, and the PCA of some collection of data $(X_1, \dots, X_n)$ with respect to $\langle\cdot,\cdot\rangle_\Sigma$ is equivalent to the PCA of $[\mathcal{T}^{1/2}(\mathcal{X}_i)]_{i=1}^n$ with respect to the Hilbert–Schmidt norm, in the sense that the eigenvalues (i.e., the variances) remain the same, and the eigenfunctions with respect to $\langle\cdot,\cdot\rangle_\Sigma$ are $\mathcal{T}^{1/2}$ applied to the eigenfunctions with respect to $\langle\cdot,\cdot\rangle_{HS}$.

The PCA algorithm presented in Section 3.2 is accordingly modified as follows:

1. Obtain an estimate $\widehat{\Sigma}_1, \dots, \widehat{\Sigma}_N$ of the operators $\Sigma_1, \dots, \Sigma_N$. Compute their Fréchet mean

   $$\widehat{\Sigma} = \underset{\Sigma}{\arg\min} \sum_{j=1}^{N} \Pi^2(\Sigma, \widehat{\Sigma}_j).$$

2. Use the estimated-log maps to lift $\widehat{\Sigma}_1, \dots, \widehat{\Sigma}_N$ to their respective correspondants $\widehat{\Delta}_j = \log_{\bar{\Sigma}}(\widehat{\Sigma}_i)$ on the tangent space at the Fréchet mean:

   $$\widehat{\Delta}_j = \mathbf{t}_j - \mathcal{I} = \overline{\Sigma}^{-1/2}(\overline{\Sigma}^{1/2}\widehat{\Sigma}_j\overline{\Sigma}^{1/2})^{1/2}\overline{\Sigma}^{-1/2} - \mathcal{I}, \quad j = 1, \dots, N.$$

3. Multiply $\widehat{\Delta}_j = \mathbf{t}_j - \mathcal{I}$ from the right by $\bar{\Sigma}^{1/2}$.

4. Perform linear PCA of the tangent space data using the spectral decomposition of $\tilde{\mathcal{K}} = K^{-1}\sum \Delta_j \Sigma^{1/2} \otimes \Delta_j \Sigma^{1/2}$. Such spectral decomposition is defined on the space of Hilbert–Schmidt operators with respect to the Hilbert–Schmidt norm.
5. Multiply (from the right) the eigenfunctions of $\tilde{\mathcal{K}}$ by $\Sigma^{-1/2}$ to obtain the eigenfunctions of $\mathcal{K}$.
6. Retract the tangent space segments obtained in (3) onto the manifold via the exponential map $\exp_{\bar{\Sigma}}$.

In the next section we explain how the tangent space PCA admits an elegant interpretation in terms of the problem of curve registration (Section 1.1.4).

3.2.2 Tangent space PCA and multicoupling

Recall that a multicoupling of a collection of zero-mean Gaussian measures $\mu_1, \ldots, \mu_N$ is realised as joint distribution of $\mu_1, \ldots, \mu_N$, such that the sum of pairwise squared distances between the covariances is minimal.
As explained in Section 1.3, optimal multicoupling also admits a probabilistic interpretation. If we adopt this point of view, in order to achieve an optimal multicoupling, we look for random vectors on $\mathscr{H}^n$ with pre-assigned marginals and whose coordinates are maximally correlated. More formally, we wish to construct a random vector $(Y_1, \ldots, Y_n)$ on $\mathscr{H}^n$ with marginals distributed as random variables $X_i \sim \mu_i$ such that

$$\mathbb{E}\sum_{i<j} \| Y_i - Y_j \|^2 \qquad \text{is minimal.}$$

The notion of multicoupling is inherently connected with Fréchet means, as seen in Lemmas 14 and 17. Indeed, Lemma 14 shows that an optimal multicoupling yields the Fréchet mean, while Lemma 17 proves that in turn, the Fréchet mean can yield an optimal multicoupling (see Zemel and Panaretos [2017] and Masarotto et al. [2018]). Moreover as explained in the previous Section, an optimal multicoupling can be used to construct a fPCA for the image of a collection of covariances on the tangent space at $\overline{\Sigma}$.

Consider a Gaussian process $X \sim \mathscr{N}(0, \Sigma)$ displaying both amplitude variation and phase variation in the following sense (see Panaretos and Zemel [2016, Section 2] and Section 1.1.4):

1. Amplitude variation (fluctuations around fixed modes): (realisation of) a Gaussian process $X \sim N(0, \Sigma)$ with Karhunen–Loève expansion as

   $$X = \sum_{n=1}^{\infty} \sigma_n^{1/2} \xi_n \varphi_n$$

   fluctuating around fixed (deterministic) modes φ_n. Here $\{\sigma_n, \varphi_n\}$ are eigenvalues and eigenfunctions of Σ, and $\xi_n \overset{iid}{\sim} N(0,1)$ a sequence of real standard normal variables.

2. Phase variation (arising from deformation fluctuations of the modes): the realisation X are warped into $\widetilde{X}$ via bounded non-negative operator T (usually uncorrelated with X),

$$\widetilde{X} = TX = \sum_{n=1}^{\infty} \sigma_n^{1/2} \xi_n T \varphi_n$$

   such that $\| T\Sigma T \|_1 < \infty$, in order for the resulting $\widetilde{X}$ to have finite variance.

In practice we want to recover the original unwarped covariance Σ and the warping map T. By doing so we can then separate the amplitude from the phase variation, as unaccounted-for phase variation might distort statistical inference.
Let us translate the registration problem to the setting of covariances, to see how tangent space PCA provides a means of registration of the curves.

In terms of covariances, the warped process $\widetilde{X}$ has covariance $T\Sigma T$ conditional on T (provided that T is uncorrelated with X). That is, the covariance of $\widetilde{X}$ is obtained as a perturbation of Σ that corresponds to a tangent perturbation, which is linear on the tangent space by mean of the optimal map T, in the sense of Section 2.4. More precisely, the covariance of $\widetilde{X}$ corresponds to the retraction via the exponential map of such linear perturbation. Now, in view of the intimate link between the exponential map defined in Section 1.4.2 and the geodesics [3], we can claim that $T\Sigma T$ is a geodesic perturbation of Σ. If $\mathbb{E}[T] = \mathscr{I}$ then the perturbations at the tangent space level have "zero mean", and theorem 20 tells us that Σ is a Fréchet mean of the random operator $T\Sigma T$. But now, retracting the PCs onto the manifold via the exponential map, will give us exactly a (smoothed) estimate of T. So in practice, if we observe a collection of perturbed operators $\Sigma_k = T_k \Sigma T_k$, tangent space PCA provides a means to approximately recover Σ and $\{T_k\}_{k=1}^N$.

3.2.3 Data Analysis

In this section, we illustrate the use of tangent space PCA by applying it to the phoneme data set described in Section 3.1.1. The five empirical covariances corresponding to the five phonemes are lifted to the tangent space via the log map at their Fréchet mean $\bar{\Sigma}$, and successively pre-multiplied by $\bar{\Sigma}^{1/2}$ as explained in Section 3.2.1. Standard PCA is then run on these quantities.
Figure 3.13 gives the coordinates of the observations in the principal components space. From it, we see clearly that the tangent space PCA captures very well the difference among the phonemes, as each phoneme group is isolated in at least one plot. More precisely:

1. The first PC captures (part of) the difference between "aa, ao and iy" and "dcl and sh".
2. The second PC captures (part of) the difference between "dcl" and "sh".

---

[3] In general, let v be a tangent vector to a manifold M at the point p. Then there is a unique geodesic $\gamma_v(t)$ such that $\gamma_v(0) = p$ and with v as initial tangent vector. In this case, the corresponding retraction via the exponential map of v is $\exp_p(v) = \gamma_v(1)$.

	PC1	PC2	PC3	PC4	PC5
Standard deviation	5.9365	3.5813	2.4753	1.7213	0.0000
Proportion of Variance	0.6166	0.2244	0.1072	0.0518	0.0000
Cumulative Proportion	0.6166	0.8410	0.9482	1.0000	1.0000

Table 3.5: Importance of each PC for the phoneme dataset

3. The third PC captures (part of) the difference between "aa and ao" and "iy".
4. The fourth PC captures (part of) the difference between "aa" and "ao".

The screeplot of the eigenvalues is given in Figure 3.14, and, the corresponding numerical values are reported in Table 3.5. As there are only five different covariances, the screeplot shows that four PCs explain the full variance of the data, which is obvious as we only have five centered data points. The fourth PC is still relevant and explains 5% of the variance.

Since the PCs are ordered according to the magnitude of the eigenvalues of the combined covariance matrix, the analysis suggests also how important are the differences among the operators. For example, the main component of variation, corresponding to the first PC, separates consonant and vowels, that is, quite different sounds, and indeed it captures about 60% of the total variance. On the other hand, the fourth PC discriminates between "aa" and "ao", which are very similar sounds, and indeed it only explains a small percentage of the total variance.

We now move into a more detailed analysis of the differences captured by the eigenvectors. The retraction of the PCs from the tangent space onto the manifold identifies the principal geodesics. Traversing these principal geodesics provides a visualisation of the main modes of variation of Σ_j near their Fréchet mean $\bar{\Sigma}$. We show that this is actually the case in Figure 3.15. The first PC captures (part of) the differences between "aa, ao and iy" and "dcl and sh". Let $\gamma_t^{(1)}$ be the first principal geodesic, i.e. the retraction of the first principal component via the exponential map. We remark that $\gamma_t^{(1)}$ is traversing the manifold of covariances, so the evaluation of $\gamma_t^{(1)}$ at each time instance t^* yields a covariance operator. We expect that if we move along $\gamma_t^{(1)}$ starting from the Fréchet mean, we find, on one end, operators which are "similar" to the covariance operators corresponding to the phonemes "aa", "ao" and "iy", while if we move in the other direction, we expect to find operators similar to the covariances of "sh" and "dcl". This would mean that $\gamma_t^{(1)}$ captures exactly the main mode of variation, which is indeed the difference between vowels and consonants.

Figure 3.15 shows on the left-hand side the arithmetic difference $\bar{\Sigma}_{aa,ao,iy} - \bar{\Sigma}_{dcl,sh}$ between the "true" barycenters $\bar{\Sigma}_{aa,ao,iy}$ of "aa, ao and iy" and $\bar{\Sigma}_{dcl,sh}$ of "dcl and sh", and on the right-hand side the arithmetic difference between two covariance operators computed along $\gamma_t^{(1)}$ if we move into the two opposite directions. As we can move along $\gamma_t^{(1)}$ for an arbitrary amount of time t, we chose two "representatives" corresponding to $t = 1$ and $t = -1$. Moving along $\gamma_t^{(1)}$ seems to show the expected differences.

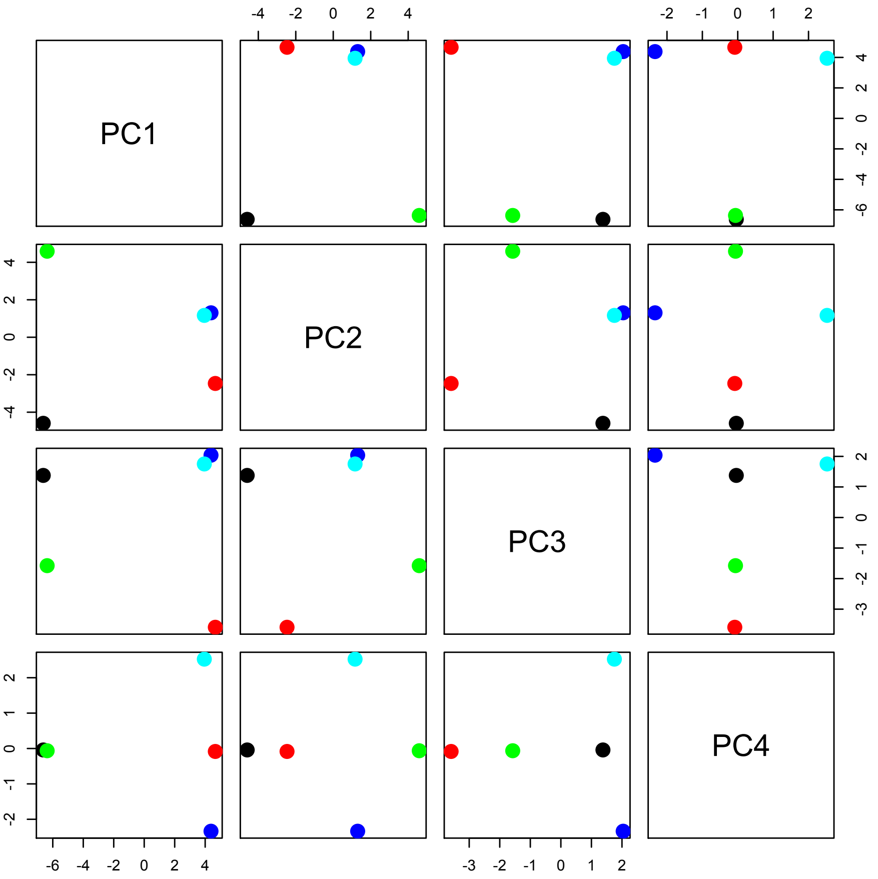


Figure 3.13: PCA scores, as computed from the phoneme dataset. The colours are as follows: “sh” black, “iy” red, “dcl “green”, “aa” blue, “ao” cyan

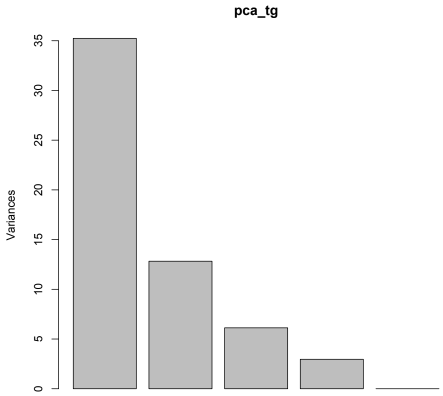
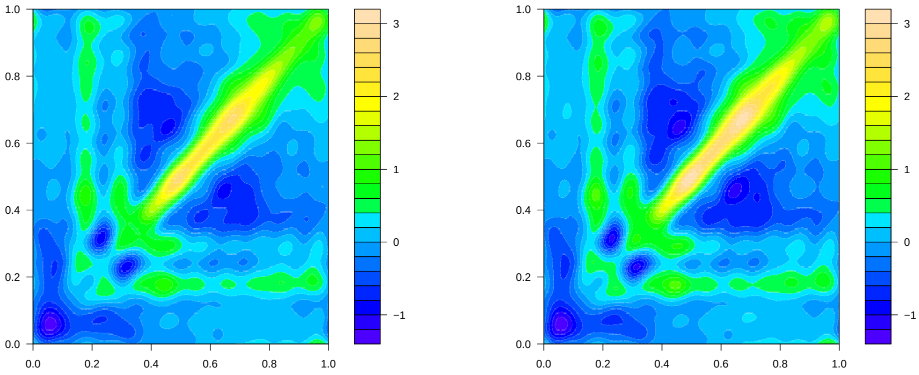


Figure 3.14: Screeplot of eigenvalues, phoneme dataset



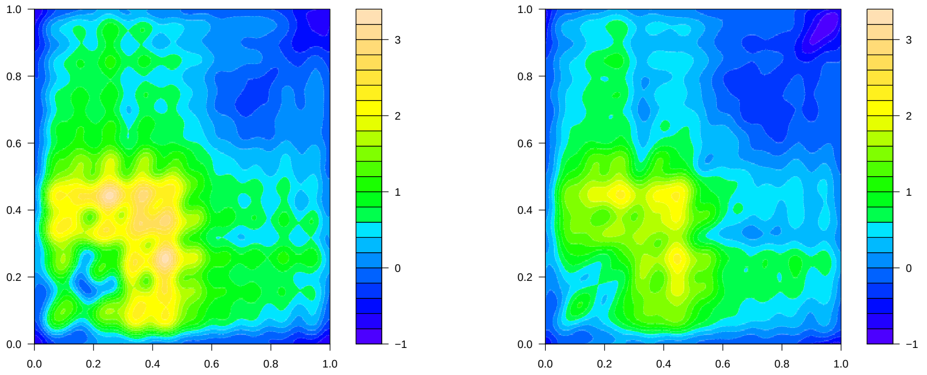
(a)

(b)

Figure 3.15: Variation along the first principal geodesic. Contour plot of the difference $\bar{\Sigma}_{aa,ao,iy} - \bar{\Sigma}_{dcl,sh}$ between the barycenters of the covariances of "aa,ao and iy" and "dcl and sh" on the left-hand side, and the arithmetic difference between two operators computed by moving in opposite directions along $\gamma_t^{(1)}$ on the right.

We do a similar analysis for the second PC. The second PC captures part of the differences "dcl" vs "sh". Figure 3.16 shows on the left-hand side the difference between the covariance operators of "dcl" and "sh", while on the right-hand side the difference between two representative covariance operators computed along $\gamma_t^{(2)}$, if we move into the two opposite directions. Again moving along $\gamma_t^{(2)}$ seems to capture the expected differences.
The third PC captures part of the differences between "iy" and "aa and ao". Figure 3.17 shows on the left-hand side the difference between the covariance operators of "iy" and the barycenter of "aa" and "ao", while on the right-hand side the difference between two representative covariance operators computed along $\gamma_t^{(3)}$ if we move into the two positive directions.
The fifth PC captures part of the differences between "aa" and "ao". Figure 3.18 shows on the left-hand side the difference covariance operators of "aa" and "ao", while on the right-hand side the difference between two representative covariance operators computed along $\gamma_t^{(5)}$ if we move into the two positive directions.

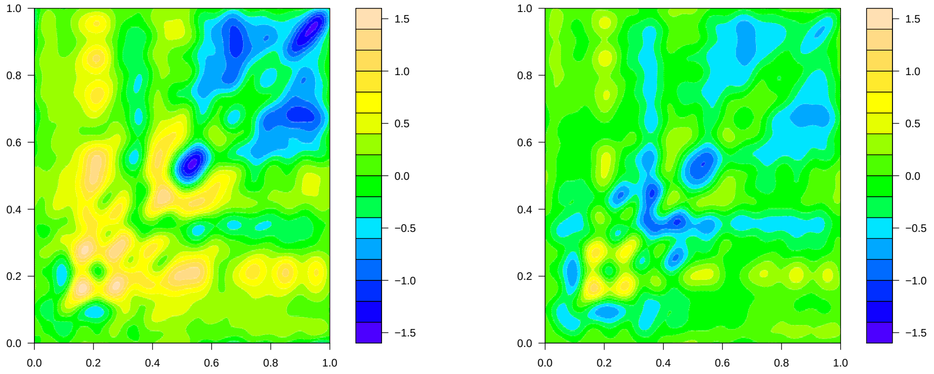
We now repeat the analysis on the more realistic extended dataset, as described in Section 3.1.1. In this case, the phoneme dataset is artificially expanded in order to produce 12 replicas of each of the 5 phoneme covariances. Each replica is computed from a subsample of 50 log-periodograms, for a total count of 60 collections of 50 curves each. PCA is carried out on the resulting 60 covariances.
We plot the first five PC scores in Figure 3.19, while the 3D scatterplot of the first three scores is in Figure 3.20. The screeplot is given in Figure 3.21 while Table 3.6 contains the values of explained variance by the first eight principal components.
Although 5PCs only explain 61% of the variance, it is enough to distinguish each phoneme group. Indeed from Figure 3.19 we see clearly that each group is isolated in at least one plot.



(a)

(b)

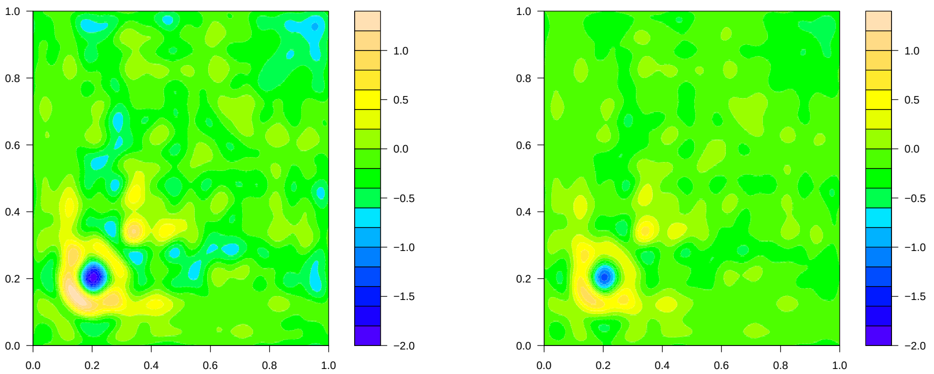
Figure 3.16: Variation along the second principal geodesic. Contour plot of the difference $\Sigma_{dcl} - \Sigma_{sh}$ between the covariances of "dcl" and "sh" on the left-hand side, and the arithmetic difference between two operators computed by moving in opposite directions along $\gamma_t^{(2)}$ on the right.



(a)

(b)

Figure 3.17: Variation along the third principal geodesic. Contour plot of the difference $\Sigma_{iy} - \bar{\Sigma}_{aa,ap}$ between the covariances of "iy" and the barycenter of "aa and ao" on the left, and the arithmetic difference between two operators computed by moving in opposite directions along $\gamma_t^{(3)}$ on the right.



(a)

(b)

Figure 3.18: Variation along the fifth principal geodesic. Variation along the third principal geodesic. Contour plot of the difference $\Sigma_{aa} - \Sigma_{ao}$ between the covariances of "aa" and "ao" on the left, and the arithmetic difference between two operators computed by moving in opposite directions along $\gamma_t^{(5)}$ on the right.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard dev	5.3837	3.5409	2.2804	2.1172	1.5755	1.4561	1.3259	1.3219
Prop of Variance	0.3299	0.1427	0.0592	0.0510	0.0283	0.0241	0.0200	0.0199
Cumulative Prop	0.3299	0.4726	0.5318	0.5828	0.6111	0.6352	0.6552	0.6751

Table 3.6: Importance of each PCs in the expanded phoneme dataset

More precisely:

1. The first PC captures (part of) the difference between "aa, ao and iy" and "dcl and sh".
2. The second PC captures (part of) the difference between "dcl" and "sh".
3. The third PC captures (part of) the difference between "aa and ao" and "iy".
4. The fifth PC captures (part of) the difference between "aa" and "ao".

It is interesting to remark that if we ignore the geometry of the space, and perform a PCA according to the standard Euclidean product, we are much less successful in identifying the different groups of phonemes. See the PC scores shown in Figure 3.22, and, in particular, compare them with those reported in Figure 3.19. Although the use of the Euclidean inner products still seems to separate clusters of covariances on the PC space, we cannot identify the reasons for variation.

We now move into the more detailed analysis of the difference captured by the eigenvectors of the joint covariance.

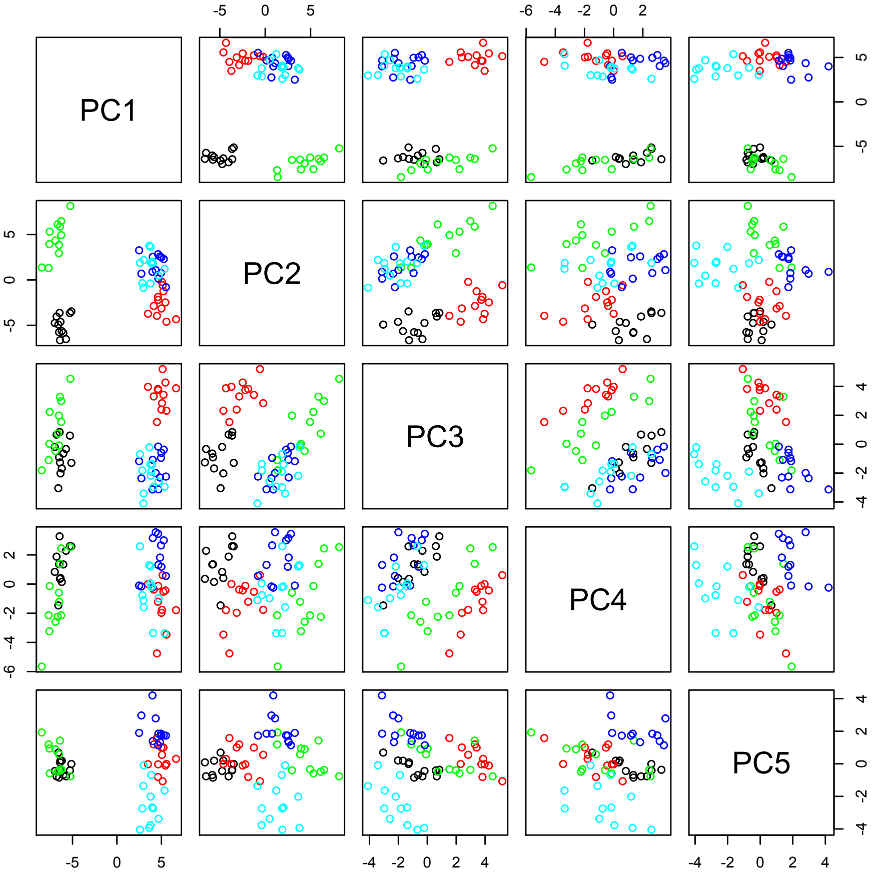


Figure 3.19: PCA scores of the expanded phoneme dataset. The colours are as follows: “sh” black, “iy” red, “dcl “green”, “aa” blue, “ao” cyan

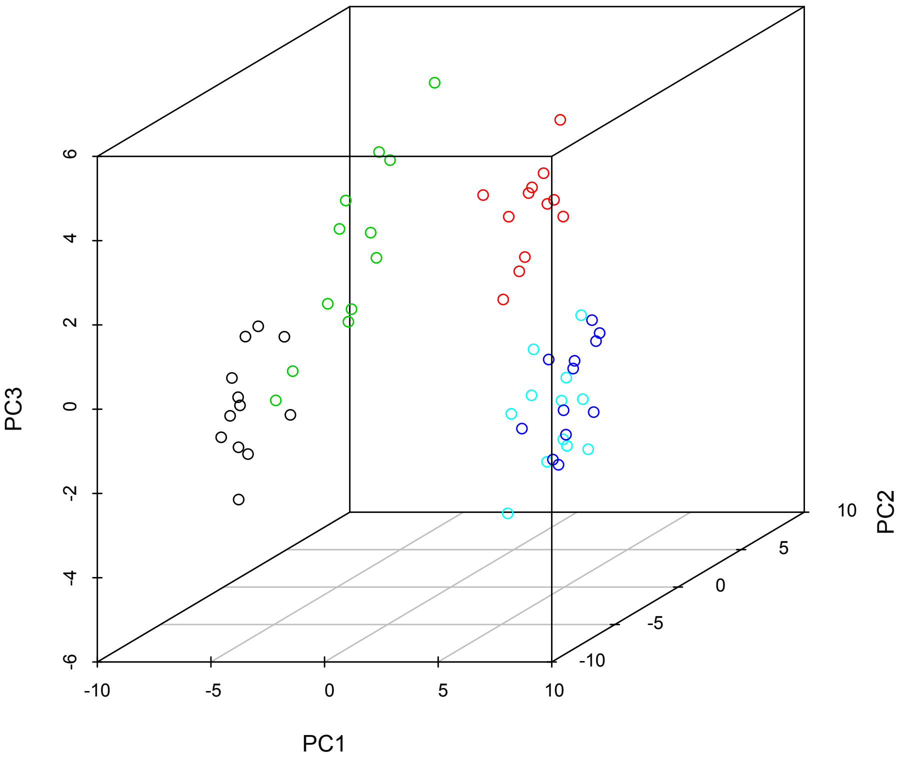


Figure 3.20: PCA scores of the expanded phoneme dataset in a 3D scatterplot. The colours are as follows: “sh” black, “iy” red, “dcl “green”, “aa” blue, “ao” cyan

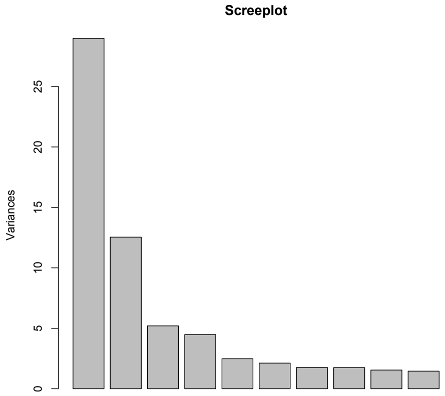


Figure 3.21: Screeplot of eigenvalues, expanded phoneme dataset

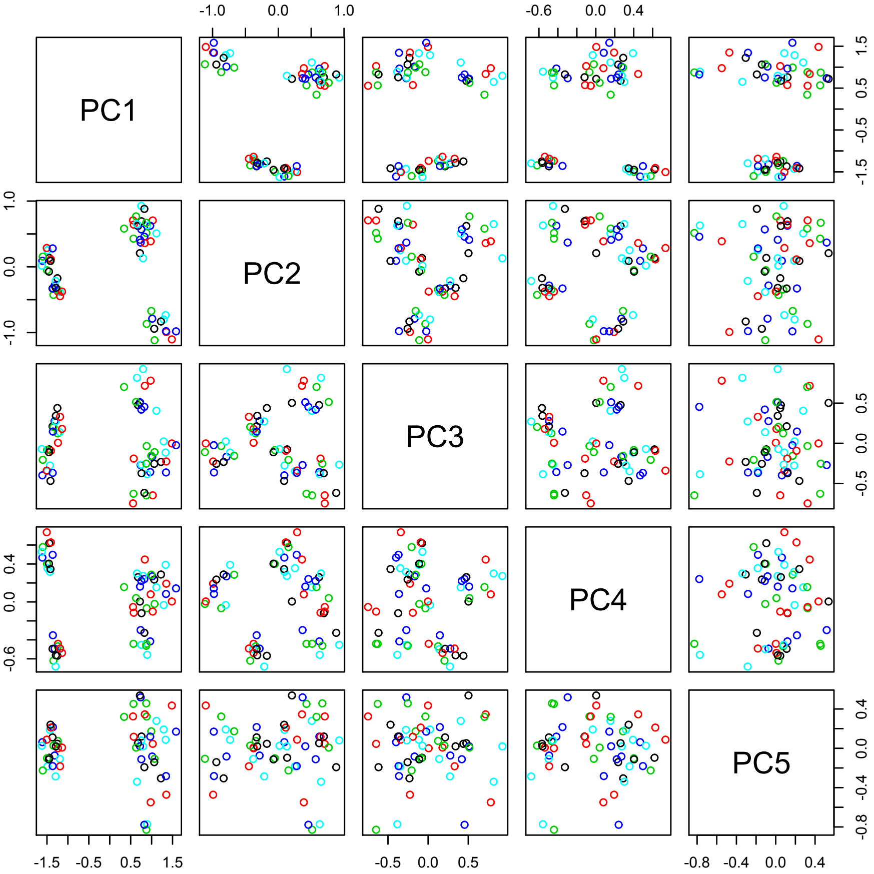
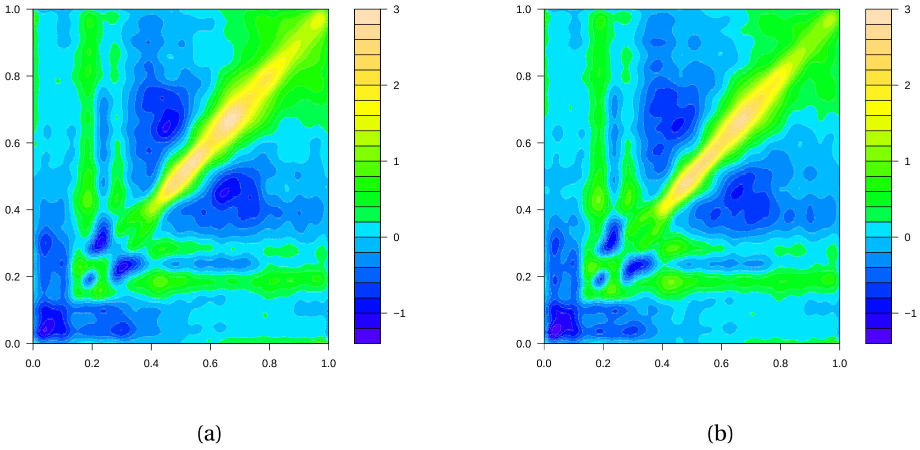


Figure 3.22: PCA scores of the 60 covariance operators based on the Hilbert–Schmidt distance.



(a) (b)

Figure 3.23: Variation along the first principal geodesic. Contour plot of the difference $\bar{\Sigma}_{aa,ao,iy} - \bar{\Sigma}_{dcl,sh}$ between the barycenters of the covariances of "aa,ao and iy" and "dcl and sh" on the left-hand side, and the arithmetic difference between two operators computed by moving in opposite directions along $\gamma_t^{(1)}$ on the right.

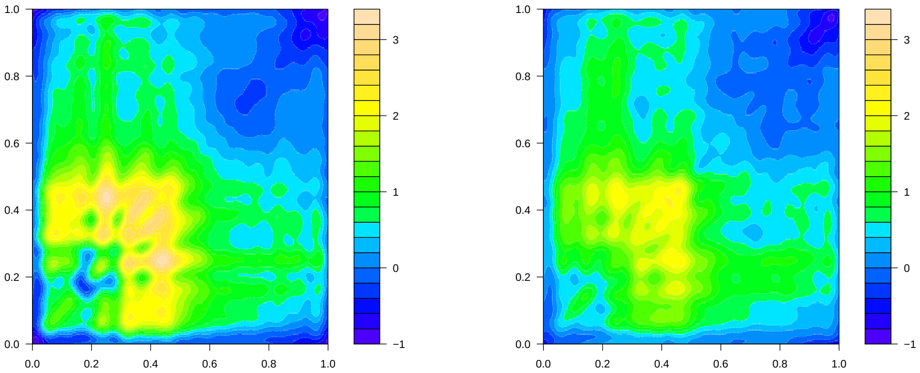
The first PC captures part of the differences between "aa, ao and iy" and "dcl and sh". If $\gamma_t^{(1)}$ again indicates the first principal geodesic, we expect that if we move along $\gamma_t^{(1)}$ starting from the Fréchet mean, we find on one end operators which are "similar" to the covariance operators corresponding to the phonemes "aa", "ao" and "iy", while if we move in the other direction, we expect to find operators similar to the covariances of "sh" and "dcl". This would mean that $\gamma_t^{(1)}$ captures exactly the main mode of variation, which is again the difference between vowels and consonants.
Figure 3.23 shows on the left-hand side the difference between true barycenters of "aa, ao and iy" and the one of "dcl and sh", while on the right-hand side the difference between two representative[4] covariance operators computed along $\gamma_t^{(1)}$, if we move into the two opposite directions. Moving along $\gamma_t^{(1)}$ seems to show the expected differences.

We do a similar analysis for the second PC. The second PC captures part of the differences "dcl" vs "sh". Figure 3.24 shows on the left-hand side the difference between the covariance operators of "dcl" and "sh", while on the right-hand side the difference between two representative covariance operators computed along $\gamma_t^{(2)}$ if we move into the two opposite directions. Again moving along $\gamma_t^{(2)}$ seems to capture the expected differences.

The third PC captures part of the differences between "iy" and "aa and ao". Figure 3.25 shows on the left-hand side the difference between the covariance operators of "iy" and the barycenter of "aa" and "ao", while on the right-hand side the difference between two representative covariance operators computed along $\gamma_t^{(3)}$ if we move into the two opposite directions.

---

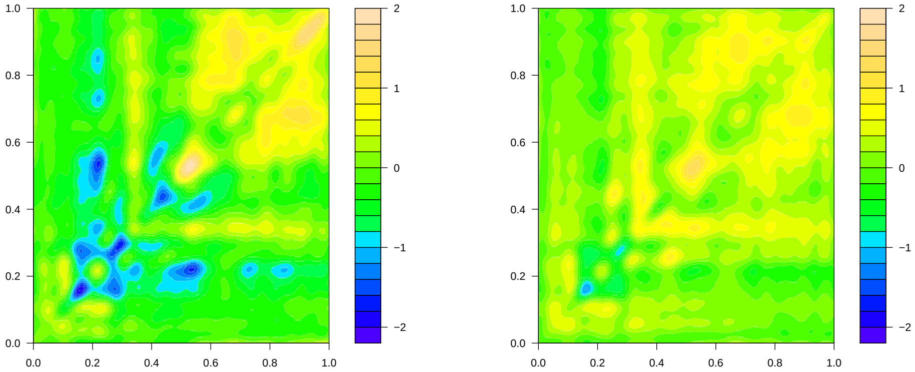
[4] Here "representative" is used in the sense of the previous paragraph, where the same analysis was performed on the original phoneme dataset.



(a)

(b)

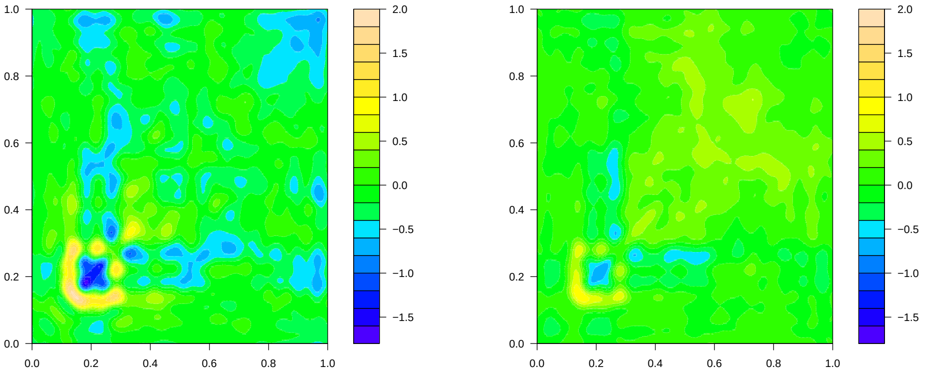
Figure 3.24: Variation along the second principal geodesic. Contour plot of the difference $\Sigma_{dcl} - \Sigma_{sh}$ between the covariances of "dcl" and "sh" on the left-hand side, and the arithmetic difference between two operators computed by moving in opposite directions along $\gamma_t^{(2)}$ on the right.



(a)

(b)

Figure 3.25: Variation along the third principal geodesic. Contour plot of the difference $\bar{\Sigma}_{aa,ao} - \Sigma_{iy}$ between the barycenters of the covariances of "aa and ao" and the covariance of "iy" on the left-hand side, and the arithmetic difference between two operators computed by moving in opposite directions along $\gamma_t^{(3)}$ on the right.



(a)

(b)

Figure 3.26: Variation along the fifth principal geodesic. Contour plot of the difference $\Sigma_{aa} - \Sigma_{ao}$ between the covariances of "aa" and "ao" on the left-hand side, difference between two operators computed by moving in opposite directions along $\gamma_t^{(5)}$ on the right.

The fifth PC captures part of the differences between "aa" and "ao". Figure 3.26 shows on the left-hand side the difference covariance operators of "aa" and "ao", while on the right-hand side the difference between two representative covariance operators computed along $\gamma_t^{(5)}$ if we move into the two opposite directions.

3.2.4 Numerical simulation

We can perform PCA as well on operators obtained by the generative model described in Section 2.4. We generate synthetic datasets which are inspired by the theoretical generative model and which yields N covariances well separated in K groups. We aim to see whether PCA is able to differentiate between the groups. Here, the method is validated over two different simulation experiments. The results seem however reproducible.

In the first experiment, the operators $\Sigma_1, \ldots, \Sigma_K$ are obtained as a conjugation perturbation of some given Fréchet mean by the generated "optimal" maps $T_1, \ldots, T_K$. Recall that the model for the optimal maps T_j is the following:

$$T_i = \sum_n \delta_n^{(i)} \sin(2n\pi t - \theta^{(i)}) \sin(2n\pi t - \theta^{(i)})$$

where the $\delta_n^{(j)}$ are drawn from a χ^2 distribution and $\theta^{(i)}$ are sampled from a von Mises distribution of mean 0 and measure of concentration $1/\sigma$ (see Subsection 3.1.1). The Fréchet mean is chosen to be $\bar{\Sigma} = U\Lambda U^*$ as in Kashlak et al. [2016], with U being a randomly generated unitary operator, and Λ a $d \times d$ diagonal matrix with eigenvalue decay of $O(d^{-4})$, d being the dimension of the matrices.

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	0.5810	0.1545	0.0966	0.0371	0.0276	0.0216
Proportion of Variance	0.8989	0.0635	0.0248	0.0037	0.0020	0.0012
Cumulative Proportion	0.8989	0.9624	0.9873	0.9909	0.9930	0.9942

Table 3.7: Importance of each PC, first experiment with the generative model.

	PC1	PC2	PC3	PC4	PC5	PC6
Standard dev	0.7376	0.5827	0.1277	0.0461	0.0397	0.0281
Prop of Variance	0.5990	0.3739	0.0180	0.0023	0.0017	0.0009
Cumulative Prop	0.5990	0.9730	0.9909	0.9933	0.9950	0.9959

Table 3.8: Importance of each PC, second experiment with the generative model.

The generative model yields optimal maps which are small perturbations of the identity. Hence, the dimension of the matrices used to approximate the operators needs to be large, otherwise the estimation errors would overwhelm the intrinsic variability of the sample. We picked the dimension to be 200, the measure of concentration to be 1 and $K = 3$. For each of the Σ_j, $j = 1, 2, 3$, we generate 100 samples of 50 Gaussian curves each. We then estimate the empirical covariance of these curves, obtaining a sample of $N = 300$ covariances. Results of the PCA are shown in Table 3.7 and Figures 3.27 and 3.28. We see that in this case as well, the different groups are clearly identified.

In the second experiment, we generate another dataset inspired by the generative model, but where the "warping" function depends on different parameters. In order for the Σ_j to be different enough, we chose K values for σ and generate only one T for each of these values. To underline the dependency on the parameter we indicate the maps as T^σ.
For this simulation we have chosen $K = 3$ and $\sigma \in \{0.1, 1, 5\}$. We then generated the $d \times d = 20 \times 20$ covariance matrices $\Sigma_1, \Sigma_2, \Sigma_3$ as $T^\sigma \bar{\Sigma} T^\sigma$, where $\bar{\Sigma}$ is chosen as above. We remark again that we are considering a extension of the generative model, since we only have one sample for each σ. The dataset of covariances obtained this way is then enlarged. For each of the Σ_i, we resample $n = 30$ Gaussian curves and estimate their empirical covariance matrices. We do this 12 times, so that in total we have 36 matrices divided in 3 groups. We then perform PCA on these 36 matrices. Results are shown in Table 3.8 and Figures 3.29 and 3.30. We see that the first two PCs explain nearly the totality of the variance, and that three groups are evident in the plot of PC1 vs PC2.

3.3 Clustering of covariance operators

Clustering deals with the problem of classifying observations belonging to K different functional populations into their respective groups. The number of groups can be known or estimated, and the algorithm aims to assign observations into subsets in such way that similar objects are in the same group, and dissimilar ones are in different groups.

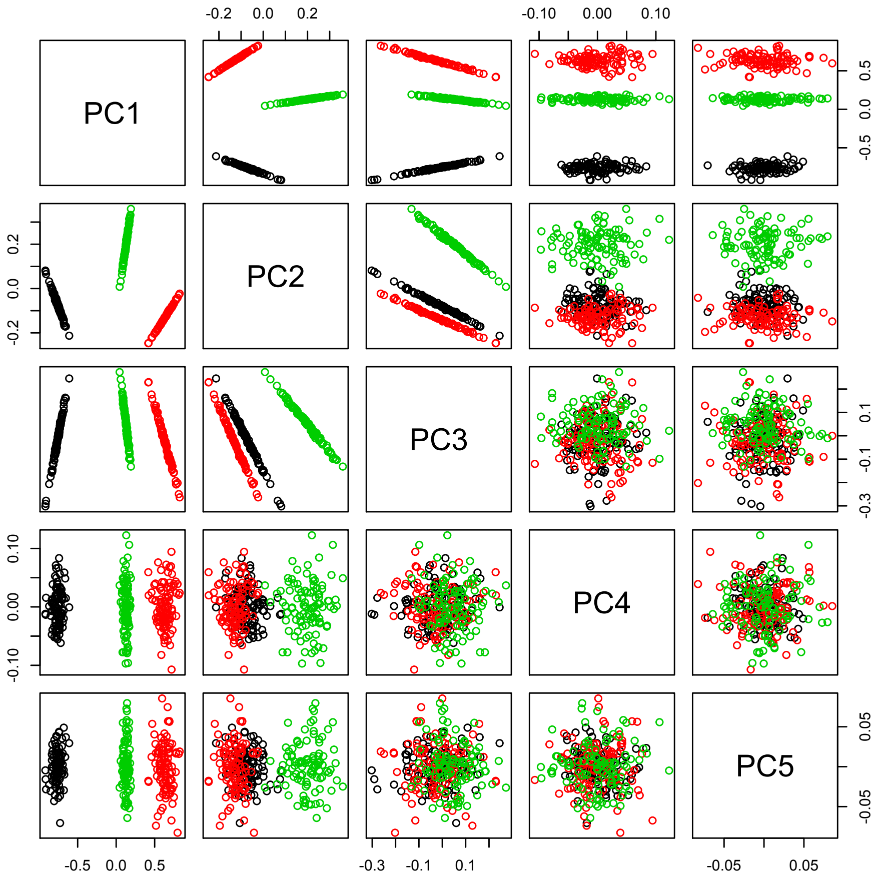


Figure 3.27: PCA scores, first experiment with the generative model. Colours correspond to the three maps generated from the model.

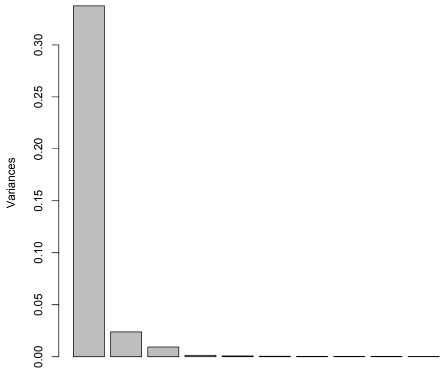


Figure 3.28: Eigenvalues screeplot, first experiment with the generative model

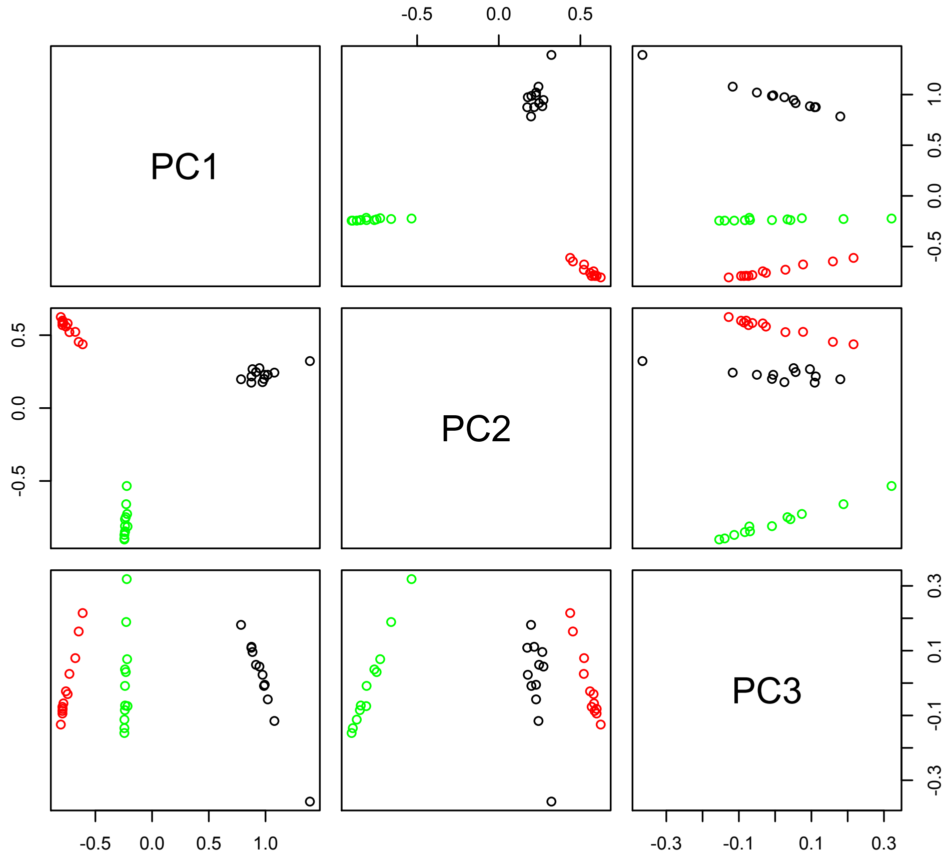


Figure 3.29: PCA scores, second experiment with the generative model. The colours are as follows: black corresponds to $\sigma = 0.1$, green to $\sigma = 1$ and red to $\sigma = 5$

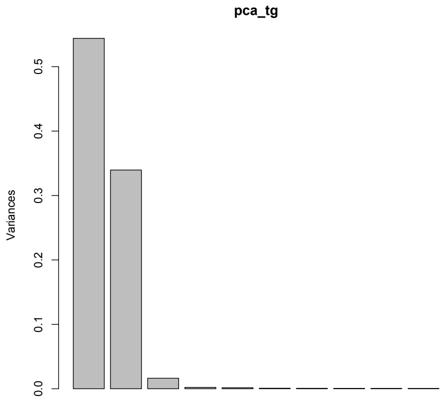


Figure 3.30: Screeplot of eigenvalues, second experiment with the generative model.

The literature on the subject offers several functional clustering approaches. Most of the clustering literature however deals with curve clustering (see, e.g. Abraham et al. [2003] for B-splines based clustering, Chiou and Li [2007] for a use of truncated Karhunen–Loève expansion, Jacques and Preda [2014] for a model-based clustering for Gaussian multivariate functional data, Tokushige et al. [2007] for fuzzy clustering algorithm and Sangalli et al. [2010] for a clustering method for warped curves). A common way to proceed for most of these approaches, is to first transform the data into some finite dimensional approximation thereof, and successively cluster them via methods designed for finite dimension.

As for covariances, in finite dimension, literature on clustering of covariance matrices can be found especially in statistical shape analysis and diffusion tensor imaging (see, among others, Lee et al. [2015], Srivastava and Klassen [2016], Srivastava et al. [2005]). Recently, methods for clustering of functional covariance operators have attracted some attention (see, e.g., Ieva et al. [2016] for an approach based on the Hilbert–Schmidt distance and limited to the case of two classes of equal size, and, Kashlak et al. [2016] for a more general approach based on concentration inequalities). In this section, we investigate two different clustering methods that are naturally linked to the theory developed in Chapters 1 and 2. The first one, described in Subsection 3.3.1, is a version of k-means partitioning performed on the tangent space. k-means partitioning is a widely-used clustering method which relies on the dissimilarity between the clusters' centroids, and which was introduced by Hartigan and Wong [1979].

Similarly to Section 3.2 on PCA, this approach shows that lifting the covariances to the tangent space makes it possible, and indeed interpretable, to use traditional methods based on the Hilbert–Schmidt distance. The second clustering method that we propose is a penalised clustering algorithm coined "soft"-clustering, and can be found in Section 3.3.2. In this approach, which makes direct use of the Wasserstein distance, the in-between cluster variability is penalised by a term proportional to the entropy of the partition matrix. In this way, each covariance operators can be partially classified into more than one group.

3.3.1 Tangent space K-means clustering.

The K-means clustering algorithm is an iterative procedure that takes as input the number K of clusters, and finds the clusters and dataset labels for the particular pre-chosen K. The algorithm is known to depend on the initiation procedure. However, it is guaranteed to converge to a local optimum [Hartigan and Wong, 1979], and *if* the data come from K different enough groups, such groups are very often identified. Inspired by the discussion in Section 3.2, we move the analysis onto the tangent space. Given N covariance matrices $\Sigma_1,\dots,\Sigma_N$ with Fréchet mean $\bar{\Sigma}$, the algorithm works as follows:

1. Obtain an estimate $\widehat{\Sigma}_1,\dots,\widehat{\Sigma}_N$ of the operators $\Sigma_1,\dots,\Sigma_N$. Compute their Fréchet mean

$$\hat{\Sigma} = \arg\min_{\Sigma} \sum_{j=1}^{N} \Pi^2(\Sigma, \widehat{\Sigma}_j)$$

2. Use the estimated-log maps to lift $\widehat{\Sigma}_1, \dots, \widehat{\Sigma}_N$ to their corresponding $\widehat{\Delta}_j = \log_{\bar{\Sigma}}(\widehat{\Sigma}_i)$ on the tangent space at the Fréchet mean:

$$\widehat{\Delta}_j = \mathbf{t}_j - \mathcal{I} = \overline{\Sigma}^{-1/2}(\overline{\Sigma}^{1/2}\widehat{\Sigma}_j\overline{\Sigma}^{1/2})^{1/2}\overline{\Sigma}^{-1/2} - \mathcal{I}, \quad j = 1, \dots, N.$$

3. Classify the N covariance operators in K groups using the standard K-means algorithm based on the Hilbert–Schmidt distance applied to $\widetilde{\Delta}_j = \widehat{\Delta}_j \hat{\Sigma}^{1/2}$.

Numerical simulations

We run the K-means algorithm on simulated data created via the generative model described in Section 3.1.1, according to the following steps:

- generate three “true” Σ_i of size 200×200 as conjugation perturbations via transport maps T_i (cfr. Equation (3.8)) of

  $$\bar{\Sigma} = U \left[\sum_{n=1}^{\infty} n^{-4} \sin(2n\pi t) \otimes \sin(2n\pi t) \right] U^*$$

  where U is a randomly generated orthogonal operator.

- Obtain 300 estimated covariances, divided in three groups, repeating the following step 100 times:

  – for every Σ_i, $i = 1, 2, 3$, generate 50 Gaussian curves with covariance Σ_i and compute the empirical covariance of these collections of curves.

- Classify the 300 covariances into 3 separate groups using the tangent space K-means algorithm.

The phase shifts θ_i in Equation (3.8) are assumed to be sampled from a von Mises distribution with dispersion parameter $1/\sigma$. The previous analysis was repeated 50 times for each of the following values of σ: $\sigma \in \{0.1, 1, 5\}$, Recall that the lowest the value of the parameter σ, the higher is the variance of the distribution.

Tables 3.9 and 3.10 give the percentage of misclassifications and exact classification, obtained in 50 replications of the described simulation experiment. We say that we classify the operators exactly when the 300 covariances are divided into 3 groups of 100 covariances corresponding exactly to the 3 “true” operators generated via the generative model. We can see that the percentage of perfect classification decreases when the dispersion of the warp maps decreases, since, in that case, the transport operators T_i’s, and therefore the true covariances of the three groups, are closer together.

Figure 3.31 illustrates that, as expected, the algorithm works when the three groups are clearly distinct. Indeed, if the minimum distance between the true covariance operators is small, the algorithm might not be able to differentiate between them. However, when the minimum

σ		
0.1	1	1.5
0.055	0.059	0.0766

Table 3.9: Mean of the percentage of wrong classification for different values of the von Mises dispersion σ.

σ		
0.1	1	1.5
0.825	0.780	0.725

Table 3.10: Percentage of cases with perfect classification for different values of the von Mises dispersion σ.

distance between two covariance operators is greater of a threshold, the classification is exact (at least in the considered cases).

Data analysis

We run the tangent space K-means algorithm on the extended phoneme data set (see Section 3.1.1) using $K = 2, \dots, 10$. For different number of classes, Figure 3.32 shows the value of the AIC_c-type criterion (Burnham and Anderson [2002])

$$AIC_c(K) = np\log(SS_K) + 2Kp + \frac{2(Kp)^2 + 2Kp}{(n-K)p-1}$$

where (i) n is the number of estimated covariance operators (60 in our case); (ii) SS_K denotes the within-cluster sum of squares (in our case is the sum of the within-cluster Hilbert-Schmidt distances of the operators lifted to the tangent space); and (iii) p is the length of each cluster centroids (the number of entries in $\widetilde{\Delta}_j$ in our case). The criterion suggests to classify the 60 covariances into $K = 5$ clusters. As shown in Table 3.11, when this number of clusters is used, the algorithm perfectly identifies the true phonemes. Replications of the simulation experiment, not reported here, show that the results presented are reproducible.

	Phonema				
Cluster	aa	ao	dcl	iy	sh
1	0	12	0	0	0
2	12	0	0	0	0
3	0	0	12	0	0
4	0	0	0	0	12
5	0	0	0	12	0

Table 3.11: Distribution of the true phonemes in the cluster identified by the tangent space K-means algorithm.

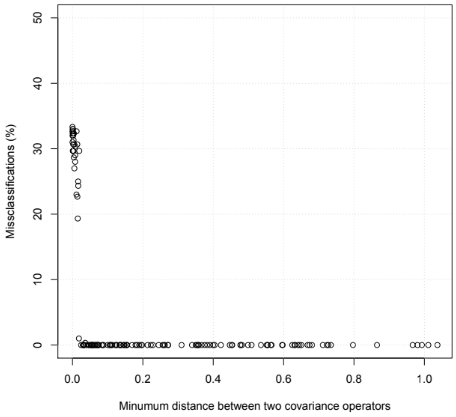


Figure 3.31: Misclassification as a function of intra-cluster distance.

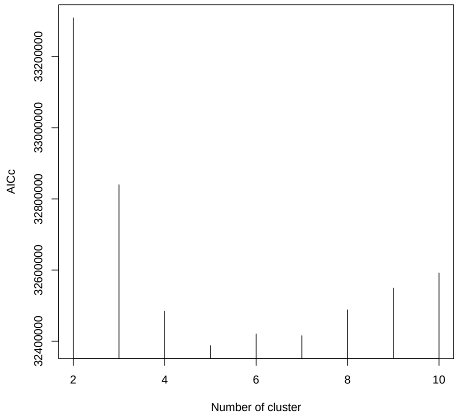


Figure 3.32: AIC_c criterion according to the number of clusters

3.3.2 Soft clustering

In this Section we propose a new clustering method based on penalisation. Although it is not interpretable in the sense of optimal transport, we found that in some cases it is more discriminating than tangent space k-means, so it appears to be a valid alternative.

Assume, as in the previous Section, that we are given N estimated covariance operators $\widehat{\Sigma}_1,\dots,\widehat{\Sigma}_N$, and, that we wish to determine both K prototype covariance operators $\overline{\Sigma}_1,\dots,\overline{\Sigma}_K$ representative of K classes and to compute a $N\times K$ partition matrix

$$P=[\pi_{i,j}] \text{ such that } \pi_{i,j}\geq 0 \text{ and } \sum_{j=1}^{K}\pi_{i,j}=1$$

where each element $\pi_{i,j}$ describes the confidence with which the covariance $\widehat{\Sigma}_i$ can be assigned to the jth class. We can imagine that $\overline{\Sigma}_1,\dots,\overline{\Sigma}_K$ are the Fréchet means of the clusters. The clustering method presented here relies on computing $\overline{\Sigma}_1,\dots,\overline{\Sigma}_K$ and P as the solution of the following optimisation problem

$$\min_{\overline{\Sigma}_1,\dots,\overline{\Sigma}_K,P}\sum_{i=1}^{N}\sum_{j=1}^{K}\pi_{i,j}\Pi^2\big(\widehat{\Sigma}_i,\overline{\Sigma}_j\big)+\eta\left(\sum_{i=1}^{N}\sum_{j=1}^{K}\pi_{i,j}\log(\pi_{i,j})+n\log(k)\right). \tag{3.11}$$

The first term gives the sum of the Wasserstein distances within the classes, while the second penalizes partition matrices with zero entropy. In particular, observe that the second term is zero when the partition matrix is uniform, i.e., $\pi_{i,j}=1/K$ for each i and j, and it reaches its maximum value $\eta N\log(K)$ for degenerate partition matrices of the type

$$\pi_{i,j}=\begin{cases}0, & j\neq r_i\\ 1, & j=r_i\end{cases} \tag{3.12}$$

for some $r_i\in\{1,\dots,k\}$. In (3.11), η is a positive tuning parameter. A zero-entropy solution of type (3.12) is obtained when $\eta=0$, while η going to infinity yields a uniform partition matrix. Observe that

1. given the partition matrix P, the desired covariance matrices/operators $\overline{\Sigma}_j$, $j=1,\dots,K$, are the (weighted) Frechet means

$$\overline{\Sigma}_j=\arg\min_{\Omega}\sum_{i=1}^{N}\pi_{i,j}\Pi^2(\widehat{\Sigma}_i,\Omega);$$

2. given $\overline{\Sigma}_j$, $j=1,\dots,K$, the partition matrix which minimizes (3.11) is

$$P=[\pi_{i,j}]=\left[\frac{e^{-\Pi^2(\widehat{\Sigma}_i,\overline{\Sigma}_j)/\eta}}{\sum_{s=1}^{K}e^{-\Pi^2(\widehat{\Sigma}_i,\overline{\Sigma}_s)/\eta}}\right]. \tag{3.13}$$

Indeed, (3.13) can be easily obtained introducing the Lagrange multipliers

$$\mathscr{L} = \sum_{i=1}^{N}\sum_{j=1}^{K} \pi_{i,j}\Pi^2\big(\widehat{\Sigma}_i, \overline{\Sigma}_j\big) + \eta\left(\sum_{i=1}^{N}\sum_{j=1}^{K} \pi_{i,j}\log(\pi_{i,j}) + N\log(K)\right) + \sum_{i=1}^{N}\lambda_i\left(\sum_{j=1}^{K}\pi_{i,j} - 1\right)$$

and solving the first-order conditions

$$\begin{aligned} \frac{d\mathscr{L}}{d\pi_{i,j}} &= \Pi^2\big(\widehat{\Sigma}_i, \overline{\Sigma}_j\big) + \log(\pi_{i,j}) + 1 + \lambda_i = 0, && (i = 1,\dots,N;\, j = 1,\dots,K),\\ \frac{d\mathscr{L}}{d\lambda_i} &= \sum_{j=1}^{K}\pi_{i,j} - 1 = 0, && (i = 1,\dots,N). \end{aligned}$$

This suggests the use of a *block coordinate descent* algorithm (Xu and Yin [2013]). The algorithm starts from K prototype covariance matrices/operators $\overline{\Sigma}_1^{(0)},\dots,\overline{\Sigma}_K^{(0)}$ and the corresponding partition matrix $P^{(0)} = [\pi_{i,j}^{(0)}]$ obtained from (3.13), and then for $r = 1, 2, \dots$:

1. compute

   $$\overline{\Sigma}_j^{(r)} = \arg\min_{\mathbf{\Omega}} \sum_{i=1}^{n} \pi_{i,j}^{(r-1)}\Pi^2(\widehat{\Sigma}_i, \mathbf{\Omega})$$

   using the gradient descent algorithm for the Fréchet mean (Section 2.3).

2. Set

   $$P^{(r)} = [\pi_{i,j}^{(r)}] = \left[\frac{e^{-\Pi^2(\widehat{\Sigma}_i, \overline{\Sigma}_j^{(r)})/\eta}}{\sum_{s=1}^{K} e^{-\Pi^2(\widehat{\Sigma}_i, \overline{\Sigma}_s)/\eta}}\right].$$

Iterations can be stopped when $\max_{i,j}\left|\pi_{i,j}^{(r)} - \pi_{i,j}^{(r-1)}\right|$ is sufficiently small.
Equation (3.13) shows that η essentially determines the "cluster sizes" and we have chosen it to be

$$\exp\left(-\frac{1}{\eta}\mathrm{median}_{i\neq j}\Pi^2(\widehat{\Sigma}_i, \widehat{\Sigma}_j)\right) \approx 10^{-6},$$

as this value seems to provide a reasonable performance in a variety of scenarios.

While no probabilistic interpretation for the partition matrix P is possible, we stress the similarity of the proposed algorithm with the classical EM approach for fitting mixture models (and performing model-based clustering).

The described algorithm only finds a *local* minimum of the objective function. Hence, it is important to choose $\overline{\Sigma}_1^{(0)},\dots,\overline{\Sigma}_K^{(0)}$ so that this local minimum corresponds to a "good" solution. For this reason, we developed a stochastic algorithm inspired by the initialisation phase of the `kmeans++` (Arthur and Vassilvitskii [2007]) and `PAM` (Kaufman and Rousseeuw [2009])

algorithms.
The initialisation attempts to find an approximate solution of the optimisation problem (3.11) constraining

$$\overline{\Sigma}_1 = \widehat{\Sigma}_{i_1}, \ldots, \overline{\Sigma}_K = \widehat{\Sigma}_{i_K}$$

for some $i_j \in \{1, \ldots, N\}$. The reason behind is to avoid computing the Fréchet mean of the data at each initialisation step , as this is the most time-consuming of all steps.
The idea is that $\Pi^2(\widehat{\Sigma}_{i_a}, \widehat{\Sigma}_{i_b})$ should be large when $a \neq b$ so we randomly sample $(i_1, \ldots, i_K)$ trying to impose this condition.
For a given `nstart` repetition of the initialisation procedure and `nrefine` refinement steps, the algorithm can be described as follows:

> Repeat `nstart` times the following steps and keep the best subset $(i_1, \ldots, i_K)$ generated.
>
> - Choose i_1 uniformly at random in $\{1, \ldots, n\}$.
> - For $j = 2, \ldots, K$, sequentially choose i_j from $\{1, \ldots, N\}$ with probability proportional to $\min(\Pi^2(\widehat{\Sigma}_i, \widehat{\Sigma}_{i_1}), \ldots, \Pi^2(\widehat{\Sigma}_i, \widehat{\Sigma}_{i_{j-1}}))$
> - Repeat `nrefine` times the following step.
>
>   For each $j = 1, \ldots, K$, sample without replacement from $\{1, \ldots, N\}$ `ntry` possible substitutions of i_j with probability proportional to $\min_{s \neq j} \Pi^2(\widehat{\Sigma}_i, \widehat{\Sigma}_{i_s})$. Keep the best found value for i_j.

We tested that the initialisation indeed tends to select covariances $\widehat{\Sigma}_{i_1}, \ldots, \widehat{\Sigma}_{i_K}$ belonging to different groups (*if* groups really exist).

Data analysis

We test the previous procedure on the extended phoneme dataset described in Section 3.1.1. To obtain an even more challenging and realistic simulation scenario, from each phoneme we extract 20 subsamples of different sizes to estimate 20 covariance matrices, as to obtain a complete dataset of 100 covariances (20 for each of the 5 phonemes). In particular, we use subsample size of either 25 or 50 curves to estimate the covariance operators. To analyse the stability of the algorithm, we repeat the classification procedure for 5 times.
Figures 3.33 and 3.34 show the medians of the column of P for the rows corresponding to the various phonemes, when the subsamples used for the analysis where of size 25 and 50, respectively. Reflecting the distances between the phoneme covariances, (i) the algorithm classify quite well phonemes “dcl”, “iy” and “sh”, while (ii) the partition matrices show some degree of uncertainty between the “aa” and “ao” covariances; this uncertainty correctly decreases as the sample size increases.

The degree of uncertainty is intrinsic in the results produced by the algorithm, as it is not required by the user to actually know the true classes. In particular, the magnitude of the

	Class				
Class	1	2	3	4	5
1	16.3	1.7	0.1	0.1	0.0
2	1.7	19.8	0.1	0.1	0.1
3	0.1	0.1	9.9	8.7	1.8
4	0.1	0.1	8.7	9.9	1.8
5	0.0	0.1	1.8	1.8	15.4

Table 3.12: Uncertainty matrix $P^* P$ (experiment n. 3 with sample size equal to 25). Observe, from Figure 3.33, that the classes identified by the algorithm correspond to: 1 ↔ "dcl"; 2 ↔ "sh"; 3 ↔ "aa" (or "ao"); 4 ↔ "ao" (or "aa"); 5 ↔ "iy".

non-diagonal entries of the matrix $P^* P$ gives a measure of the degree of "confusion" between the corresponding groups. For example, Table 3.12 reports this matrix for the third experiment conducted with a subsample of size 25. The Table shows that the algorithm produces reasonably definite classification results corresponding to the classes of "iy", "sh" and "dcl" and less about "aa" and "ao", which, being very similar sounds, are harder to distinguish.

Once we obtain an estimate of the prototypes $\overline{\Sigma}_1, \dots, \overline{\Sigma}_K$, further analyses can be done to understand the differences between the identified classes, e.g., by lifting the estimates to the tangent space, we can perform a tangent space PCA. As an example, Figure 3.35 shows the plot of the PCA scores computed from $\overline{\Sigma}_1, \dots, \overline{\Sigma}_k$ in the case of the third experiment with a subsample of size 25. Regardless of the small sample size and the partial classification errors, and without considering the sign and the colors which are completely arbitrary, the scores have a spatial distribution similar to those displayed in Figure 3.13, which was based on the covariances estimates using the entire dataset available for each phoneme. Indeed, an analysis of the eigenfunctions (not reported here) reveals that the differences between the estimated class prototypes $\overline{\Sigma}_1, \dots, \overline{\Sigma}_5$ captures the differences between the "true" covariances.

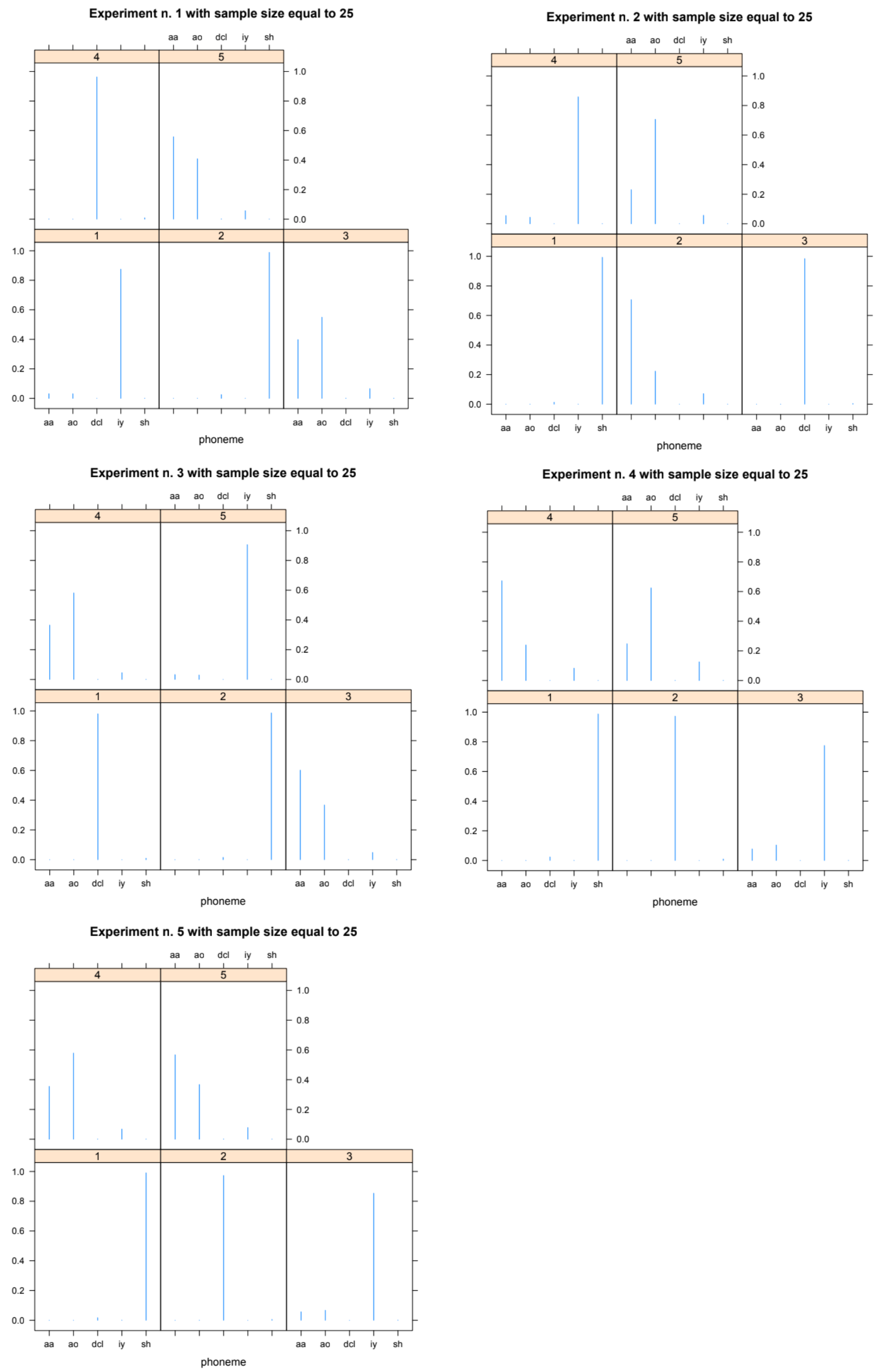


Figure 3.33: Classifications into clusters based on a subsample of 25. The picture show the medians of the column of the partition matrix P for the rows corresponding to the various phonemes.

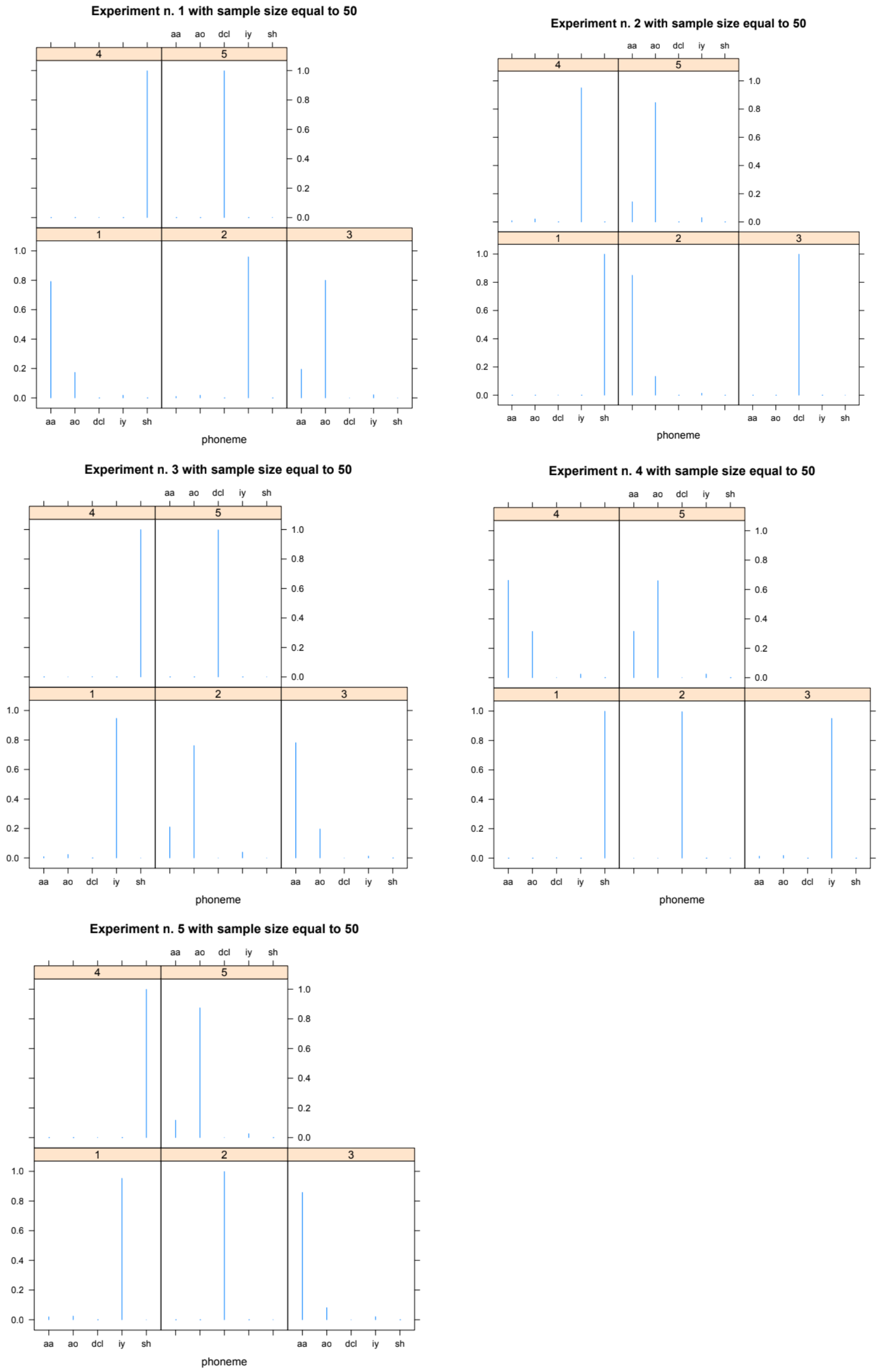


Figure 3.34: Classifications into clusters based on a subsample of 50. The picture show the medians of the column of the partition matrix P for the rows corresponding to the various phonemes.

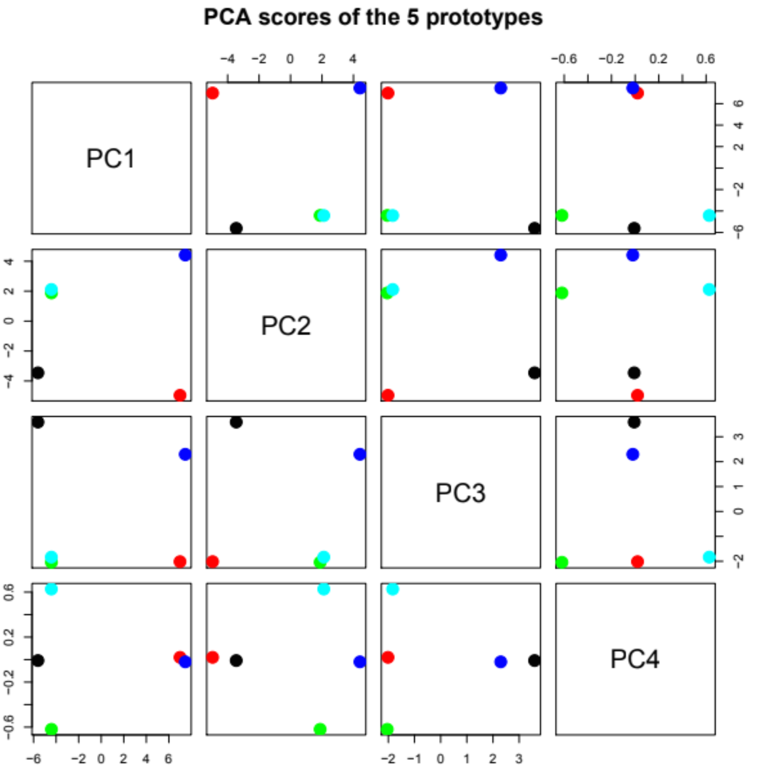


Figure 3.35: Tangent space PCA of estimated prototypes (experiment n. 3 with sample size equal to 25).

4 Conclusions and outlook

After giving a summary of the results presented in this manuscript, we wish to conclude with an outlook over possibile future research directions.

Covariance operators are crucial elements in FDA, primely due to their connection with the Karhunen-Loéve expansion. However situations might arise where covariances operators are highlighted as independently relevant statistical objects. The line of research in this direction mostly viewed covariances as Hilbert–Schmidt operators, thus ignoring their trace–class nature and the fundamental non-linearity of the space they lie in. The first steps towards a non-linear analysis of covariance operators were moved by Pigoli et al. [2014a]. We made a connection between the Procrustes distance promoted by Pigoli et al. [2014a] and the Wasserstein metric for Gaussian processes (Section 2.1). This connection allowed us to exploit the wealth of geometrical and analytical properties of optimal transportation. We illustrated key geometrical and topological aspects of the space of covariance operators endowed with the Procrustes metric (Sections 2.1.1 and 1.4.2). Through the notion of multicoupling of Gaussian measures, we establish existence, uniqueness and stability for the Fréchet mean of covariances with respect to the Procrustes metric (Section 2.2). Such a Fréchet mean is computable via a version of the generalized Procrustes algorithm, that provably converges in finite dimension (Section 2.3). Moreover, we gave generative statistical models for covariances which are: linear on the tangent space, compatible with the Procrustes metric, and connected with the problem of curves registration (Section 2.4). This novel perspective on optimal transport allowed us to introduce a new ANOVA test, which in the functional case seems to dominate in power the state-of-the-art results of other testing procedures (Section 3.1). The understanding of the geometry leads to a PCA that respects the nature of the covariance operators (Section 3.2). Finally we gave another example of applications of the Wasserstein framework, showing two algorithms for clustering of covariance operators (Section 3.3).

With this work, we also generated new questions which are so far unanswered and that can offer interesting leads for future research.

The most important issue would be establishing the injectivity (regularity) of the Fréchet mean

$\overline{\Sigma}$, as it would automatically yield the solution to the multicoupling problem. As shown by Masarotto et al. [2018, Section 12], injectivity holds for commuting operators. We conjecture the result to hold true in the general case, as it is in finite dimensions.

Conjecture 24 (Regularity of the Fréchet Mean)**.** *Let* $\Sigma_1, \dots, \Sigma_n$ *be covariances on* $\mathscr{H}$ *with* Σ_1 *injective. Then their Fréchet mean* $\overline{\Sigma}$ *with respect to the Procrustes metric* Π *is also injective.*

Another relevant question would be to establish the stability of the Procrustes algorithm presented in Section 2.3 to increasing projection dimensions.
Finally, one more interesting question would be whether the (empirical) Fréchet mean of $\Sigma_1, \dots, \Sigma_n$ is consistent with respect to its population counterpart, as the sample size grows to infinity. The Fréchet mean can be seen as a M-estimator [van der Vaart and Wellner, 1996]. When studying properties of these estimators, the theory of empirical properties comes in naturally. Specifically, to assess consistency of the Fréchet mean, we can employ argmin theorems as, for example, in van der Vaart and Wellner [1996, Chapter 3.2]). In finite dimension this has been done by Bigot et al. [2013]. Unfortunately applications of argmin theorems require convergence or continuity hypotheses not verified for infinite measures.
Under mild conditions on the law of Σ, this population mean is guaranteed to be unique (see Proposition 15). Thanks to a result from Ziezold [1977], if (uniqueness holds and) a population mean exists and the sequence of empirical means converge, then the limit must be the population mean.The problem is that we cannot show that in general a Fréchet mean exists. Le Gouic and Loubes [2016] showed existence, but their result do not apply in our case because $\mathscr{H}$ is not locally compact.

Bibliography

Christophe Abraham, Pierre-André Cornillon, Eric Matzner-Løber, and Nicolas Molinari. Unsupervised curve clustering using b-splines. *Scandinavian Journal of Statistics*, 30(3): 581–595, 2003.

Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

Andrew L Alexander, Jee Eun Lee, Mariana Lazar, and Aaron S Field. Diffusion tensor imaging of the brain. *Neurotherapeutics*, 4(3):316–329, 2007.

PC Álvarez-Esteban, E Del Barrio, JA Cuesta-Albertos, C Matrán, et al. Uniqueness and approximate computation of optimal incomplete transportation plans. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 47(2):358–375, 2011.

PC Álvarez-Esteban, E del Barrio, JA Cuesta-Albertos, and C Matrán. A fixed-point approach to barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications*, 441 (2):744–762, 2016.

Luigi Ambrosio and Nicola Gigli. A user's guide to optimal transport. In *Modelling and optimisation of flows on networks*, pages 1–155. Springer, 2013.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures.* Springer Science & Business Media, 2008.

Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM journal on matrix analysis and applications*, 29(1):328–347, 2007.

David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

John AD Aston, Jeng-Min Chiou, and Jonathan P Evans. Linguistic pitch analysis using functional principal component mixed effect models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(2):297–317, 2010.

Joao Lucas Barbosa and M do Carmo. On the size of a stable minimal surface in r3. *American Journal of Mathematics*, pages 515–528, 1976.

Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

Michal Benko, Wolfgang Härdle, and Alois Kneip. Common functional principal components. *Ann. Stat.*, 37(1):1–34, 2009. doi: 10.1214/07-AOS516.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the Bures-Wasserstein distance between positive definite matrices. *Expositiones Mathematicae (in press, https://doi.org/10.1016/j.exmath.2018.01.002)*, 2018. ISSN 0723-0869. doi: https://doi.org/10.1016/j.exmath.2018.01.002. URL http://www.sciencedirect.com/science/article/pii/S0723086918300021.

Rabi Bhattacharya and Vic Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifolds: i. *Annals of statistics*, pages 1–29, 2003.

Rabi Bhattacharya and Vic Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifolds: ii. *Annals of statistics*, pages 1225–1259, 2005.

Jérémie Bigot and Thierry Klein. Characterization of barycenters in the Wasserstein space by averaging optimal transport maps. *arXiv preprint arXiv:1212.2562*, 2012.

Jérémie Bigot, Xavier Gendre, et al. Minimax properties of fréchet means of discretely sampled curves. *The Annals of Statistics*, 41(2):923–956, 2013.

Graciela Boente, Daniela Rodriguez, and Mariela Sued. Testing equality between several populations covariance operators. *Annals of the Institute of Statistical Mathematics*, 70(4):919–950, 2018.

Fred L Bookstein et al. Size and shape spaces for landmark data in two dimensions. *Statistical science*, 1(2):181–222, 1986.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

Donald Bures. An extension of kakutani's theorem on infinite product measures to the tensor product of semifinite w*-algebras. *Transactions of the American Mathematical Society*, 135:199–212, 1969.

K. P. Burnham and D. R. Anderson. *Model Selection and Multimodel Inference: a Practical-Theoretic Approach*. Springer, New York, NY, 2002.

Alessandra Cabassi, Davide Pigoli, Piercesare Secchi, Patrick A Carter, et al. Permutation tests for the equality of covariance operators of functional data with applications to evolutionary biology. *Electronic Journal of Statistics*, 11(2):3815–3840, 2017.

Jeng-Min Chiou and Pai-Ling Li. Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):679–699, 2007.

N Coffey, AJ Harrison, OA Donoghue, and K Hayes. Common functional principal components analysis: A new approach to analyzing human movement data. *Human movement science*, 30(6):1144–1166, 2011.

JA Cuesta-Albertos, C Matrán-Bea, and A Tuero-Diaz. On lower bounds for the l_2-Wasserstein metric in a Hilbert space. *Journal of Theoretical Probability*, 9(2):263–283, 1996.

Juan Antonio Cuesta-Albertos and Carlos Matrán. Notes on the Wasserstein metric in Hilbert spaces. *The Annals of Probability*, 17(3):1264–1276, 1989.

Michael J Daniels and Mohsen Pourahmadi. Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika*, 89(3):553–566, 2002.

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.*, 12:136–154, 1982. doi: 10.1016/0047-259X(82)90088-4.

Marie-Hélène Descary. *Functional data analysis by matrix completion.* PhD thesis, Ecole Polytechnique Fédérale de Lausanne, 2017.

DC Dowson and BV Landau. The Fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.

I. L. Dryden and K. V. Mardia. *Statistical Shape Analysis, with Applications in R. Second Edition.* John Wiley and Sons, Chichester, 2016.

Ian L Dryden, Alexey Koloydenko, and Diwei Zhou. Non-euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *The Annals of Applied Statistics*, pages 1102–1123, 2009.

IL Dryden and KV Mardia. *Statistical analysis of shape.* Wiley, 1998.

Richard M Dudley. *Real Analysis and Probability.* Chapman and Hall/CRC, 2018.

Sam Efromovich. *Nonparametric curve estimation: methods, theory, and applications.* Springer Science & Business Media, 2008.

LC Evans. Measure theory and fine properties of functions studies in advances mathematics studies in advanced mathematics, 1992.

Jianqing Fan, Irène Gijbels, Tien-Chung Hu, and Li-Shan Huang. A study of variable bandwidth selection for local polynomial regression. *Statistica Sinica*, pages 113–127, 1996.

F. Ferraty and P. Vieu. Nonparametric models for functional data, with application in regression, time series prediction and curve discrimination. *Journal of Nonparametric Statistics*, 16 (1-2):111–125, 2004. doi: 10.1080/10485250310001622686.

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

Maurice Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. H. Poincaré*, 10(3):215–310, 1948.

Stefan Fremdt, Josef G Steinebach, Lajos Horváth, and Piotr Kokoszka. Testing the equality of covariance operators in functional samples. *Scandinavian Journal of Statistics*, 40(1): 138–152, 2013.

Robertas Gabrys, Lajos Horváth, and Piotr Kokoszka. Tests for error correlation in the functional linear model. *Journal of the American Statistical Association*, 105(491):1113–1125, 2010.

Wilfrid Gangbo and Andrzej Święch. Optimal maps for the multidimensional Monge–Kantorovich problem. *Communications on pure and applied mathematics*, 51(1):23–45, 1998.

John C Gower. Generalized Procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

Ulf Grenander. Stochastic processes and statistical inference. *Arkiv för matematik*, 1(3): 195–277, 1950a.

Ulf Grenander. Stochastic processes and statistical inference. *Arkiv för Matematik*, 1:195–277, 1950b. ISSN 0004-2080. URL http://dx.doi.org/10.1007/BF02590638.

John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

Trevor Hastie, Andreas Buja, and Robert Tibshirani. Penalized discriminant analysis. *The Annals of Statistics*, pages 73–102, 1995.

L. Horvath and P. Kokoszka. *Inference for Functional Data with Applications.* Springer Series in Statistics, 2012.

Lajos Horváth and Piotr Kokoszka. *Inference for functional data with applications*, volume 200. Springer Science & Business Media, 2012.

Lajos Horváth, Marie Hušková, and Gregory Rice. Test of independence for functional data. *Journal of Multivariate Analysis*, 117:100–119, 2013.

Tailen Hsing and Randall Eubank. *Theoretical foundations of functional data analysis, with an introduction to linear operators.* John Wiley & Sons, 2015.

Stephan Huckemann, Thomas Hotz, and Axel Munk. Intrinsic shape analysis: Geodesic pca for Riemannian manifolds modulo isometric Lie group actions. *Statistica Sinica*, pages 1–58, 2010.

Mohamed Ibazizen and Jacques Dauxois. A robust principal component analysis. *Statistics*, 37:73–83, 2003. doi: 10.1080/0233188031000065442. URL http://dx.doi.org/10.1080/0233188031000065442.

Francesca Ieva, Anna Maria Paganoni, and Nicholas Tarabelloni. Covariance-based clustering in multivariate and functional data analysis. *Journal of Machine Learning Research*, 17(143):1–21, 2016. URL http://jmlr.org/papers/v17/15-568.html.

Julien Jacques and Cristian Preda. Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis*, 71:92–106, 2014.

Daniela Jarušková. Testing for a change in covariance operator. *Journal of Statistical Planning and Inference*, 143(9):1500–1511, 2013.

Ian T Jolliffe. Principal component analysis and factor analysis. *Principal component analysis*, pages 150–166, 2002.

Harold E Jones and Nancy Bayley. The berkeley growth study. *Child development*, pages 167–173, 1941.

Emmanuel K Kalunga, Sylvain Chevallier, Quentin Barthélemy, Karim Djouani, Yskandar Hamam, and Eric Monacelli. From euclidean to riemannian means: Information geometry for ssvep classification. In *International Conference on Networked Geometric Science of Information*, pages 595–604. Springer, 2015.

Leonid V Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201, 1942.

Stepan Karamardian. *Fixed Points.* Elsevier, 2014.

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

Kari Karhunen. Über lineare Methoden in der Wahrscheinlichkeitsrechnung. *Ann. Acad. Sci. Fenn., Ser. A I, No.*, 37:79 p., 1947.

Adam B Kashlak, John AD Aston, and Richard Nickl. Inference on covariance operators via concentration inequalities: k-sample tests, classification, and clustering via rademacher complexities. *arXiv preprint arXiv:1604.06310*, 2016.

Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

David G Kendall. A survey of the statistical theory of shape. *Statistical Science*, pages 87–99, 1989.

David George Kendall, Dennis Barden, Thomas K Carne, and Huiling Le. *Shape and shape theory*, volume 500. John Wiley & Sons, 2009.

Alois Kneip and Theo Gasser. Statistical tools to analyze data representing a sample of curves. *The Annals of Statistics*, pages 1266–1305, 1992.

Martin Knott and Cyril S Smith. On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43(1):39–49, 1984.

David Kraus. Components and completion of partially observed functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2014.

David Kraus and Victor M Panaretos. Dispersion operators and resistant second-order functional data analysis. *Biometrika*, 99(4):813–832, 2012.

Serge Lang. *Fundamentals of differential geometry*, volume 191. Springer Science & Business Media, 2012.

Jimmie Lawson and Yongdo Lim. Weighted means and karcher equations of positive operators. *Proceedings of the National Academy of Sciences*, page 201313640, 2013.

Thibaut Le Gouic and Jean-Michel Loubes. Existence and consistency of Wasserstein barycenters. *Probability Theory and Related Fields*, pages 1–17, 2016.

Haesung Lee, Hyun-Jung Ahn, Kwang-Rae Kim, Peter T Kim, and Ja-Yong Koo. Geodesic clustering for covariance matrices. *Communications for Statistical Applications and Methods*, 22(4):321–331, 2015.

Xueli Liu and Hans-Georg Müller. Functional convex averaging and synchronization for time-warped random curves. *Journal of the American Statistical Association*, 99(467):687–699, 2004.

Tomas Masak. Iteratively reweighted least squares algorithm for sparse principal component analysis with application to voting records. *Statistika*, 97:88–106, 01 2017.

Valentina Masarotto, Victor M Panaretos, and Yoav Zemel. Procrustes metrics on covariance operators and optimal transportation of gaussian processes. *Sankhya A*, pages 1–42, 2018.

Valentina Masarotto, Victor M Panaretos, and Yoav Zemel. Introduction to statistics in the wasserstein space. *In preparation*, 2019.

Robert J McCann. A convexity principle for interacting gases. *Advances in mathematics*, 128 (1):153–179, 1997.

Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

Francisco Antonio Ocaña, Ana M Aguilera, and Mariano J Valderrama. Functional principal components analysis by choice of norm. *Journal of Multivariate Analysis*, 71(2):262–276, 1999.

Ingram Olkin and Friedrich Pukelsheim. The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 48:257–263, 1982.

V. M. Panaretos and Y. Zemel. *Introduction to Statistics in the Wasserstein Space*. Springer Briefs in Probability and Mathematical Statistics, To appear.

V. M. Panaretos, D. Kraus, and J. H. Maddocks. Second-order comparison of gaussian random functions and the geometry of dna minicircles. *JASA Theory & Methods*, 105(490):670–682, June 2010a.

Victor M Panaretos and Shahin Tavakoli. Cramér–Karhunen–Loève representation and harmonic principal component analysis of functional time series. *Stochastic Processes and their Applications*, 123(7):2779–2807, 2013.

Victor M Panaretos and Yoav Zemel. Amplitude and phase variation of point processes. *The Annals of Statistics*, 44(2):771–812, 2016.

Victor M Panaretos, David Kraus, and John H Maddocks. Second-order comparison of gaussian random functions and the geometry of dna minicircles. *Journal of the American Statistical Association*, 105(490):670–682, 2010b.

Efstathios Paparoditis and Theofanis Sapatinas. Bootstrap-based testing for functional data. *arXiv preprint arXiv:1409.4317*, 2014.

Fortunato Pesarin and Luigi Salmaso. *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons, 2010.

Davide Pigoli and Laura M Sangalli. Wavelets in functional data analysis: estimation of multidimensional curves and their derivatives. *Computational Statistics & Data Analysis*, 56 (6):1482–1498, 2012.

Davide Pigoli, John AD Aston, Ian L Dryden, and Piercesare Secchi. Distances and inference for covariance operators. *Biometrika*, 101(2):409–422, 2014a.

Davide Pigoli, John D. Aston, Ian L. Dryden, and Piercesare Secchi. Distances and inference for covariance functions. *Biometrika*, 101(2):409–422, 2014b.

J. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, New York, 2005a.

Jim O Ramsay. Functional components of variation in handwriting. *Journal of the American Statistical Association*, 95(449):9–15, 2000.

JO Ramsay and BW Silverman. Springer series in statistics. In *Functional data analysis*. Springer, 2005b.

C. Radhakrishna Rao. Some statistical methods for comparison of growth curves. *Biometrics*, 14(1):pp. 1–17, 1958. ISSN 0006341X. URL http://www.jstor.org/stable/2527726.

Lüdger Rüschendorf and Svetlozar T Rachev. A characterization of random variables with minimum L^2-distance. *Journal of Multivariate Analysis*, 32(1):48–54, 1990.

Ludger Rüschendorf and Ludger Uckelmann. On the n-coupling problem. *Journal of multivariate analysis*, 81(2):242–258, 2002.

Laura M Sangalli, Piercesare Secchi, Simone Vantini, and Valeria Vitelli. K-mean alignment for curve clustering. *Computational Statistics & Data Analysis*, 54(5):1219–1233, 2010.

Armin Schwartzman. *Random ellipsoids and false discovery rates: Statistics for diffusion tensor imaging data*. PhD thesis, Stanford University, 2006.

Bernard W Silverman et al. Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics*, 24(1):1–24, 1996.

B.W. Silverman. Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics*, 24(1), 1996.

Anuj Srivastava and Eric P Klassen. *Functional and shape data analysis*. Springer, 2016.

Anuj Srivastava, Shantanu H Joshi, Washington Mio, and Xiuwen Liu. Statistical shape analysis: Clustering, learning, and testing. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 27(4):590–602, 2005.

Joan G Staniswalis and J Jack Lee. Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*, 93(444):1403–1418, 1998.

Elias M Stein and Rami Shakarchi. *Real analysis: measure theory, integration, and Hilbert spaces*. Princeton University Press, 2009.

Rong Tang and Hans-Georg Müller. Pairwise curve synchronization for functional data. *Biometrika*, 95(4):875–889, 2008.

Shahin Tavakoli and Victor M Panaretos. Detecting and localizing differences in functional time series dynamics: a case study in molecular biophysics. *Journal of the American Statistical Association*, 111(515):1020–1035, 2016.

Shuichi Tokushige, Hiroshi Yadohisa, and Koichi Inada. Crisp and fuzzy k-means clustering algorithms for multivariate functional data. *Computational Statistics*, 22(1):1–16, 2007.

Aad W van der Vaart and Jon A Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

Cédric Villani. *Topics in Optimal Transportation*, volume 58. American Mathematical Society, 2003.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Max-K von Renesse and Karl-Theodor Sturm. Entropic measure and Wasserstein diffusion. *The Annals of Probability*, 37(3):1114–1191, 2009.

Matt P Wand and M. Chris Jones. *Kernel smoothing*, volume 60 of *Monographs on statistics and applied probability*. Chapman & Hall, London, 1995. ISBN 0-412-55270-1.

Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg Müller. Functional data analysis. *Annual Review of Statistics and Its Application*, 3:257–295, 2016.

Zhizhou Wang, Baba C Vemuri, Yunmei Chen, and Thomas H Mareci. A constrained variational principle for direct estimation and smoothing of the diffusion tensor field from complex dwi. *IEEE transactions on Medical Imaging*, 23(8):930–939, 2004.

Yangyang Xu and Wotao Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on imaging sciences*, 6(3):1758–1789, 2013.

Fang Yao and Thomas C. M. Lee. Penalized spline models for functional principal component analysis. *J. Roy. Statist. Soc. Ser. B*, 68:3–25, 2006.

Fang Yao, Hans-Georg Müller, Andrew J Clifford, Steven R Dueker, Jennifer Follett, Yumei Lin, Bruce A Buchholz, and John S Vogel. Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate. *Biometrics*, 59(3):676–685, 2003.

Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590, 2005a.

Fang Yao, Hans-Georg Müller, Jane-Ling Wang, et al. Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, 33(6):2873–2903, 2005b.

Yoav Zemel. *Fréchet means in Wasserstein space: theory and algorithms.* PhD thesis, Ecole Polytechnique Fédérale de Lausanne, 2017.

Yoav Zemel and Victor M Panaretos. Fréchet means and Procrustes analysis in Wasserstein space. *Bernoulli (to appear), available on arXiv:1701.06876*, 2017.

Jin-Ting Zhang. *Analysis of variance for functional data.* Chapman and Hall/CRC, 2013.

Herbert Ziezold. On expected figures and a strong law of large numbers for random elements in quasi-metric spaces. In *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes and of the 1974 European Meeting of Statisticians*, pages 591–602. Springer, 1977.

VALENTINA MASAROTTO

PERSONAL DETAILS

Work Address
EPFL SB MATH SMAT
CH-1015 Lausanne
Telephone (work): +41(0)216932566
Telephone (private): +41(0)762498273

DOB: 10th May 1985
Nationality: Italian
Gender: Female
Email: valentina.masarotto@gmail.com
Family: Mother of Lorenzo, 3.5.

CURRENT EMPLOYMENT

01/2014–present	École polytechnique fédérale de Lausanne - Ph.D. student in Mathematical Statistics.

PREVIOUS EMPLOYMENT

01/2013–12/2013	Signal Processing Lab, École polytechnique fédérale de Lausanne - Visiting student.
01/2012–12/2012	Università di Padova - Ph.D. in Statistics.
09/2010–11/2011	ESA/ESTEC - YGT in the Automation and Robotics section.
09/2009-07/2010	TU Delft University - Researcher.

EDUCATION

07.2009	Master degree ALGANT in Mathematics, Leiden University.
03.2009	Master degree ALGANT in Mathematics, University of Padua. Final grade: 110/110 cum laude.
09.2006	B.Sc. in Mathematics, University of Padua. Final grade: 110/110 cum laude.

TEACHING ASSISTANCE

Statistics (Spring 2014, 2016, 2018)
Statistics for Data Science (Fall 2017)
Probability and Statistics (Spring 2017)
Linear Model (Fall 2016)
Linear Algebra (Fall 2015)
Probability Theory (Fall 2014).

PUBLICATIONS, CONFERENCES, DISTINCTIONS

Publications

K. Dajani, D. Hensley, C. Kraaikamp, V. Masarotto
“Arithmetic and ergodic properties of ’flipped’-continued fraction algorithms”,
Acta Arithmetica, 153(1), 51-79.

V. Masarotto, V.M. Panaretos, Y. Zemel
“Procrustes Metrics on Covariance Operators
and Optimal Transportation of Gaussian Processes”
Sankhya A (2018): 1-42.

V. Masarotto, V.M. Panaretos, Y. Zemel
“Optimal transport based applications of Gaussian Processes”
(In preparation)

Conferences

ASTRA 2011
(11th Symposium on Advanced Space Technologies in Robotics and Automation)
Conference paper: “Planetary Terrain Analysis For Robotics Missions”,
Masarotto, V. ; Joudrier, L. ; Hidalgo Carrio, J. ; Lorenzoni, L. .

International BASP Frontiers workshop 2015
”Bayesian Average SParsity” (contributed poster).

Other talks:
Heriot-Watt University Edinburgh (2013, “Bayesian Compressed Sensing”),
Università di Padova (2014, “Bayesian Compressed Sensing”),
Leiden University
(2018, “Procrustes metrics on covariance operators and optimal transportation of Gaussian processes”).

Scholarships

Scholarship winner (competitive contest) by Fond. Ing. A.Gini, Padova (2009-2010).

SKILLS PROFILE

Languages
Italian (native speaker), English (fluent), French (conversational skills), Spanish and German (once fluent, now rusty).

Computer skills
Programming languages: R, Matlab, C. Also experience with Python, Mathematica, C++, Fortrand90, Scilab, IDL.
Platforms: OS, Unix, Linux, Windows.

ADDITIONAL

One month cooperation (October 2009) with TU Delft as invited guest.
Mathematics student union representative while undergraduate.

Several years of voluntary work (support to mentally/physically handicapped people; waiter at Cucine Economiche Popolari, Padova; salesman in fair-trade organisation Manitese, Padova).

Scout leader in the AGESCI association.
Among the many things involved, it comprised raising money for international projects (Sarajevo, Ethiopia, Israel) and keeping relationships with local authorities.

Ex Cat.A Basketball player (national level as teenager). Travelled all over Europe to dance and teach (private and groups) Argentinian tango. Travelled across Europe by bike. Racing cyclist. I find restoring old books a form of meditation.

REFERENCES

References available upon request

Ce document a été imprimé au Centre d'impression EPFL,
imprimerie climatiquement neutre, certifiée myClimate.