
Statistical Methods for Sparse Functional Object Data: Elastic Curves, Shapes and Densities

Lisa Maike Steyer

München 2023

Statistical Methods for Sparse Functional Object Data: Elastic Curves, Shapes and Densities

Lisa Maike Steyer

Dissertation
at the Faculty of Mathematics, Informatics and Statistics
of the Ludwig-Maximilians-Universität München

handed in by
Lisa Maike Steyer

Munich, September 26th 2023

First Referée: Prof. Dr. Sonja Greven

Second Referée: PD Dr. Fabian Scheipl

Third Referée: Prof. Simone Vantini

Date of the disputation: November 20th 2023

**Statistische Methoden für spärlich
beobachtete funktionale Objektdaten:
Elastische Kurven, Formen und Dichten.**

Lisa Maike Steyer

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

eingereicht von
Lisa Maike Steyer

München, den 26. September 2023

Erstgutachterin: Prof. Dr. Sonja Greven

Zweitgutachter: PD Dr. Fabian Scheipl

Drittgutachter: Prof. Simone Vantini

Tag der mündlichen Prüfung: 20. November 2023

Acknowledgments

First of all, I would like to thank my Ph.D. advisor, Sonja Greven, for her invaluable guidance in scientific research. She has not only provided expert mentorship, but also served as an inspiring role model as a female scientist and university educator. I would also like to express my gratitude to PD Dr. Fabian Scheipl and Prof. Simone Vantini for their time and consideration in reviewing this thesis.

I would like to express my deepest gratitude to my fellow Ph.D. students, Almond, Alex, Eva, Marco, and Manuel, who provided camaraderie and constructive feedback throughout this academic journey. Special thanks go to Almond for our inspiring and close collaboration on numerous projects, and to both Almond and Alex for their detailed proofreading, which has greatly improved the clarity and quality of this thesis. Furthermore, I would like to thank the numerous people who accompanied me during my Ph.D. studies at the LMU Munich and the HU Berlin, whose support was crucial, although it is impossible to name them all.

I am deeply grateful for the support I have received from my mother, who encouraged me to pursue an academic education, and my partner, Patti, who supports me in every aspect of my life. Thanks also to my rugby team for being the best mental support group I could ask for.

I would also like to acknowledge the support of the German research foundation (DFG), which partially funded this research through the grant GR 3793/3-1, entitled “Flexible regression methods for curve and shape data”.

Summary

Many applications naturally yield data that can be viewed as elements in non-linear spaces. Consequently, there is a need for non-standard statistical methods capable of handling such data. The work presented here deals with the analysis of data in complex spaces derived from functional L^2 -spaces as quotient spaces (or subsets of such spaces). These data types include elastic curves represented as d -dimensional functions modulo re-parametrization, planar shapes represented as 2-dimensional functions modulo rotation, scaling and translation, and elastic planar shapes combining all of these invariances. Moreover, also probability densities can be thought of as non-negative functions modulo scaling.

Since these functional object data spaces lack a natural Hilbert space structure, this work proposes specialized methods that integrate techniques from functional data analysis with those for metric and manifold data. In particular, but not exclusively, novel regression methods for specific metric quotient spaces are discussed. Special attention is given to handling discrete observations, since in practice curves and shapes are typically observed only as a discrete (often sparse or irregular) set of points. Similarly, density functions are usually not directly observed, but a (small) sample from the corresponding probability distribution is available. Overall, this work comprises six contributions that propose new methods for sparse functional object data and apply them to relevant real-world datasets, predominantly in a biomedical context.

Based on the square-root velocity (SRV) framework, Paper I develops methods for modeling Fréchet means of irregularly/sparsely observed curves modulo re-parametrization using splines. It also provides identifiability statements for individual spline representations under re-parametrization, which are also relevant to subsequent contributions. In Paper II, this approach is extended to elastic plane shapes, that is 2-dimensional curves modulo translation, rotation, scaling, and re-parametrization. This extension is achieved by identifying the real plane with the complex numbers and establishing a connection between full Procrustes mean estimation to covariance estimation in irregular/sparse functional data analysis.

Paper III extends unconditional mean estimation for curves modulo re-parametrization to regression, that is to conditional mean estimation given covariates. It also discusses the generalization of 'linear' regression to quotient metric spaces arising from actions by isometries in a broader context. In complement to this, Paper IV focuses on inelastic plane curves as response objects in a generalized additive regression model, taking into account the Riemannian manifold structure of the shape space. In this case, the conditional mean shape is modeled by a geodesic response function, while

residuals and distances are determined by the shape geometry. To estimate the model, a Riemannian L2-Boosting algorithm is proposed. Combining both regression approaches, Paper V presents a regression method for elastic shapes that respects all invariances: rotation, translation, scaling, and re-parametrization.

Paper VI focuses on 1-dimensional probability density functions, which are defined as equivalence classes modulo scaling and can be identified with a subspace of the separable Hilbert space \mathbb{L}^2 by the centered log-ratio transformation. Building on this correspondence, this contribution proposes Functional Principal Component Analysis (FPCA) for densities based on discrete samples drawn from each density. To achieve this, the underlying functional densities are treated as latent variables within a maximum likelihood framework, and model estimation is carried out using a Monte Carlo Expectation Maximization (MCEM) algorithm.

Zusammenfassung

Viele Anwendungen liefern Daten, die natürlicherweise als Elemente in nichtlinearen Räumen aufgefasst werden können. Daher besteht ein Bedarf an speziellen statistischen Methoden, die mit solchen Daten umgehen können. Die hier vorgestellte Arbeit befasst sich mit der Analyse von Daten in komplexen Räumen, die von funktionalen \mathbb{L}^2 -Räumen als Quotientenräumen (oder Teilmengen solcher Räume) abgeleitet sind. Zu diesen Datentypen gehören elastische Kurven, die als d -dimensionale Funktionen modulo Reparametrisierung dargestellt werden können, 2-dimensionale Formen, die als Funktionen modulo Rotation, Skalierung und Translation darstellbar sind, und elastische 2-dimensionale Formen, die alle diese Invarianzen kombinieren. Darüber hinaus können auch Wahrscheinlichkeitsdichten als nichtnegative Funktionen modulo Skalierung aufgefasst werden.

Da die oben genannten funktionalen Objektdatenräume keine natürliche Hilbert-Raumstruktur haben, werden in dieser Arbeit spezielle Methoden vorgeschlagen, die Techniken der funktionalen Datenanalyse mit denen für metrische Daten und für Daten auf Riemannschen Mannigfaltigkeiten integrieren. Insbesondere, aber nicht ausschließlich, werden neue Regressionsmethoden für spezifische metrische Quotientenräume diskutiert. Besonderes Augenmerk wird auf den sinnvollen Umgang mit diskreten Beobachtungen gelegt, da in der Praxis Kurven und Formen typischerweise nur als diskrete (oft spärliche oder unregelmäßige) Punktmengen beobachtet werden. Ebenso werden Dichtefunktionen in der Regel nicht direkt beobachtet, sondern es steht eine (kleine) Stichprobe aus der entsprechenden Wahrscheinlichkeitsverteilung zur Verfügung. Insgesamt umfasst diese Arbeit sechs Beiträge, die neue Methoden für spärliche funktionale Objektdaten vorschlagen und sie auf relevante reale Datensätze vorwiegend in einem biomedizinischen Kontext anwenden.

Basierend auf dem Square-Root-Velocity (SRV) Ansatz werden in Paper I Methoden zur Modellierung von Fréchet-Mitteln von unregelmäßig/spärlich beobachteten Kurven modulo Re-Parametrisierung mit Splines entwickelt. Außerdem werden Identifizierbarkeitsaussagen für einzelne Splinedarstellungen unter Re-Parametrisierung gezeigt, die auch für nachfolgende Beiträge relevant sind. In Paper II wird dieser Ansatz auf elastische 2-dimensionale Formen, d.h. 2-dimensionale Kurven modulo Translation, Rotation, Skalierung und Re-Parametrisierung erweitert. Diese Erweiterung wird durch die Identifikation von \mathbb{R}^2 mit den komplexen Zahlen und durch die Verbindung der Procrustes-Mittelwertschätzung mit der Kovarianzschätzung, die zur Analyse von unregelmäßigen/spärlich beobachteten funktionalen Daten verwendet wird, erreicht.

Paper III erweitert die Mittelwertschätzung für Kurven modulo Reparametrisierung

auf die Regression, d.h. auf die bedingte Mittelwertschätzung für gegebene Kovariaten. Darüber hinaus wird in einem allgemeineren Kontext die Verallgemeinerung der „linearen“ Regression auf metrische Quotientenräume diskutiert, die sich aus Aktionen durch Isometrien ergeben. Ergänzend dazu konzentriert sich Paper IV auf unelastische ebene Kurven als Response in einem verallgemeinerten additiven Regressionsmodell, wobei die Riemannsche Mannigfaltigkeitsstruktur des Formraums berücksichtigt wird. In diesem Fall wird die bedingte mittlere Form durch eine geodätische Antwortfunktion modelliert, während die Residuen und Abstände durch die Formgeometrie bestimmt werden. Zur Schätzung des Modells wird ein Riemannscher L2-Boosting-Algorithmus vorgeschlagen. Die Kombination beider Regressionsansätze führt in Paper V zu einer Regressionsmethode für elastische Formen, die alle Invarianten berücksichtigt: Rotation, Translation, Skalierung und Reparametrisierung.

Paper VI konzentriert sich auf eindimensionale Wahrscheinlichkeitsdichtefunktionen, die als Äquivalenzklassen modulo Skalierung definiert sind und durch die zentrierte Log-Ratio-Transformation mit einem Unterraum des separablen Hilbert-Raums \mathbb{L}^2 identifiziert werden können. Aufbauend auf dieser Identifikation wird eine funktionale Hauptkomponentenanalyse (FPCA) für Dichten vorgeschlagen, die auf diskreten Stichproben basiert, von denen jede als unabhängige Stichprobe aus der Verteilung mit der jeweiligen Dichte betrachtet wird. Dazu werden die zugrundeliegenden funktionalen Dichten als latente Variablen in einem Maximum-Likelihood Ansatz behandelt und die Modellschätzung mit einem Monte-Carlo-Expectation-Maximization (MCEM) Algorithmus durchgeführt.

Contents

1	Introduction	1
1.1	Functional Data Analysis	1
1.1.1	Smoothing	2
1.1.2	Mean and Covariance Function	4
1.1.3	Functional Principal Component Analysis	6
1.1.4	Regression for Functional Response	9
1.1.5	Registration	10
1.2	Selected Methods from Object Data Analysis	13
1.2.1	Fréchet Means	14
1.2.2	Regression in Metric Spaces	15
1.2.3	Regression on Riemannian Manifolds	16
1.3	Functional Object Data	17
1.3.1	Functional Shapes	17
1.3.2	Elastic Curves and Shapes	19
1.3.3	Densities	20
1.4	Overview of Thesis Contributions in the Context of Functional Object Data Analysis	22
	Bibliography	22

Contributing manuscripts 31

2	Paper I: Elastic Analysis of Irregularly or Sparsely Sampled Curves <i>Lisa Steyer, Almond Stöcker and Sonja Greven</i>	31
3	Paper II: Elastic Full Procrustes Analysis of Plane Curves via Hermitian Co- variance Smoothing <i>Almond Stöcker, Manuel Pfeuffer, Lisa Steyer and Sonja Greven</i>	45
4	Paper III: Regression in Quotient Metric Spaces with a Focus On Elastic Curves <i>Lisa Steyer, Almond Stöcker and Sonja Greven</i>	67
5	Paper IV: Functional Additive Models on Manifolds of Planar Shapes and Forms <i>Almond Stöcker, Lisa Steyer and Sonja Greven</i>	109

Contents

6	Paper V: Elastic Shape Regression for Plane Curves	125
	<i>Almond Stöcker, Lisa Steyer and Sonja Greven</i>	
7	Paper VI: Principal Component Analysis in Bayes Spaces for Sparsely Sampled Density Functions	145
	<i>Lisa Steyer and Sonja Greven</i>	
	Appendix	175
A	Online Supplement for Paper I	175

1. Introduction

This thesis is dedicated to the statistical analysis of functional object data. Functional object data, also referred to as next-generation functional data (Wang et al., 2016), include objects that are elements of a complex data space derived from a functional \mathbb{L}^2 space, the space of square integrable functions. Specifically, this work considers functional object data spaces that can be obtained as quotient spaces of subsets of \mathbb{L}^2 with respect to certain equivalence relations.

To enable a meaningful analysis of this type of data, the contributions of this thesis are guided by two fundamental methodological approaches: functional data analysis (FDA), in particular for handling sparsely observed functions, and object (oriented) data analysis (ODA), a framework that summarizes methods tailored for observations in more complex and potentially nonlinear spaces. For functional data, methods have been developed to deal with discrete, error-prone observations of the functions. This work is intended to contribute to the development of such methods for functional object data as well, and therefore we present in the following relevant methods from both FDA (Section 1.1) and ODA (Section 1.2). A more detailed description of the functional object data spaces considered in this thesis is then given in Section 1.3, before an overview of the contributions of this thesis in the context of functional object data analysis is given in Section 1.4.

1.1. Functional Data Analysis

Functional data analysis (FDA) focuses on the analysis and theoretical study of random variables and their corresponding observations that are defined by their functional behavior. Unlike traditional statistical analysis of scalar or vector variables, FDA considers the units of observation themselves to be functions. These functions can take various forms, such as curves, surfaces, or images, depending on the dimension of their domain and image. Examples of such functional data in various fields include mortality and fertility rates over time (Hyndman and Ullah, 2007), temperature and precipitation measurements over time (Ferraty and Vieu, 2006) or over a 2-dimensional spatial domain (Cressie and Wikle, 2011) and 3-dimensional hand trajectories of participants in a neurological experiment (Gallivan and Chapman, 2014).

The origins of functional data can be traced back at least to the work of Ramsay (1982), while Ramsay and Silverman (2005, first published in 1997) made the subject of FDA available to a wide audience. This book introduced basic concepts such as functional data smoothing and registration, functional principal components, and

functional linear regression models, all while focusing on independent and identically distributed (i.i.d.) samples of curves measured on dense, common grids. For a more theoretical perspective on FDA and methods applicable to functional data observed on sparse and irregular grids, researchers can refer to the work of Hsing and Eubank (2015). Besides, books of Ferraty and Vieu (2006) and of Horváth and Kokoszka (2012) should be mentioned as comprehensive literature on FDA.

This section highlights fundamental techniques in FDA that are relevant to the contributions presented in this thesis. In doing so, the discussion focuses on one-dimensional functional data $f : I \rightarrow \mathbb{R}$ defined on a real interval I , where the parameter $t \in I$ is denoted as “time” unless otherwise specified for better readability. Formally, an observed function f is thereby interpreted as a realization of a continuous stochastic process $\{F(t)\}_{t \in I}$ in the Hilbert space \mathbb{L}^2 , the space of square-integrable functions.

1.1.1. Smoothing

In practice, functional data is usually not observed completely, but only at a finite number of discrete time points $t_1 < \dots < t_m \in I$, $m \in \mathbb{N}$, possibly with additional noise. This means that instead of observing a function $f(t)$ for all $t \in I$, usually only a discrete sample y_j from $Y_j = f(t_j) + \epsilon_j$ with independent random error ϵ_j and $\mathbb{E}(\epsilon_j) = 0$ for $j = 1, \dots, m$ is available. A common approach is then to apply a suitable smoothing technique to obtain a continuous representation of the underlying functional data while preserving the essential features of the observed curves.

There are several strategies for effectively smoothing functional data. One widely used method is kernel smoothing, which is based on the concept of kernel density estimation (Rosenblatt, 1956; Parzen, 1962). The kernel density estimator for a data set $y_1, \dots, y_m \in \mathbb{R}$, which is sampled from a probability distribution with density $g : \mathbb{R} \rightarrow \mathbb{R}^+$, is constructed as

$$\hat{g}(y) = \frac{1}{mh} \sum_{j=1}^m K\left(\frac{y - y_j}{h}\right),$$

where $h > 0$ is the bandwidth and $K : \mathbb{R} \rightarrow \mathbb{R}^+$ is the kernel function. Although any continuous probability density functions can be chosen for the kernel K , symmetric functions like uniform, quadratic, or Gaussian kernels are typically preferred. From this, one can derive the Nadaraya-Watson estimator (Nadaraya, 1964; Watson, 1964) for functional data

$$\hat{f}(t) = \frac{\sum_{j=1}^m y_j K\left(\frac{t-t_j}{h}\right)}{\sum_{j=1}^m K\left(\frac{t-t_j}{h}\right)}. \quad (1.1)$$

This estimator is also used to estimate nonparametric regression models. Please refer to the textbook of Härdle et al. (2012) for a discussion of common nonparametric and semiparametric modeling techniques, including the choice of the bandwidth parameter h which controls the smoothness of the estimated function. In general, a larger bandwidth results in a smoother estimate, but may fail to capture fine details. Conversely, a smaller bandwidth allows the estimate to closely follow the fluctuations in the data, thereby capturing intricate structures. However, choosing a bandwidth that is too small can result in a noisy estimate that follows the data points too closely and does not provide a reliable representation of the underlying function (bias-variance tradeoff). In practice, the optimal bandwidth is usually determined by cross-validation.

Another smoothing technique is to represent functional data by a linear combination of a finite number of basis functions $b_k : I \rightarrow \mathbb{R}$, where $k = 1, \dots, N$, $N \in \mathbb{N}$ (Ramsay and Silverman, 2005). The basis coefficients ξ_k , $k = 1, \dots, N$ of the underlying function f can then be estimated from the equation $y_j = f(t_j) + \epsilon_j = \sum_{k=1}^N \xi_k b_k(t_j) + \epsilon_j$ for all $j = 1, \dots, m$ via least-squares, that is

$$\hat{\xi}_1, \dots, \hat{\xi}_N = \operatorname{argmin}_{\xi_1, \dots, \xi_N} \sum_{j=1}^m \left(y_j - \sum_{k=1}^N \xi_k b_k(t_j) \right)^2 \quad (1.2)$$

which is a linear regression problem with coefficients ξ_1, \dots, ξ_N . Appropriate basis functions should be selected to enable a smooth representation of the data, while the number of basis functions, similar as the bandwidth for kernel smoothing, controls how closely the estimated curve fits the observed points. Prominent choices of basis functions include polynomial, Fourier, wavelet, and spline basis functions. The emphasis here is on bases that span a locally polynomial space, the spline space. A spline is composed of polynomials of fixed degree n_{deg} between a given set of knots and is $n_{deg} - 1$ times differentiable on the entire interval. For a comprehensive understanding of splines, including a precise definition and insightful discussion, see the work of de Boor (2001). This source also provides an overview of appropriate bases for spline spaces.

In particular, the B-spline basis introduced by Schoenberg (1946) has proven to be particularly convenient, since it consists of splines with local support defined by a recursion formula that can also be used to provide analytically available derivatives and integrals. In addition, the work of Eilers and Marx (1996) has contributed to the popularity of B-splines by implementing a penalty mechanism to address the delicate issue of knot selection. Their proposal involves the use of a large number of equidistant knots along with a differential penalty applied to adjacent B-spline coefficients. In this framework, the smoothness of the resulting fitted curve is controlled by a roughness penalty parameter (also called smoothing parameter), with a large penalty resulting in a nearly constant function and a penalty equal to zero yielding the unpenalized estimate.

The optimal roughness penalty parameter can again be chosen by cross-validation. Alternatively, Wand (2003) points out the strong analogy between smoothing and linear mixed model estimation. This correspondence leads to an estimate for the roughness penalty parameter via restricted maximum likelihood (REML) estimation.

Although a penalized approach diminishes the emphasis on precise knot selection, it remains an active research area. For example, a recent paper by Basna et al. (2022) presents a method that assumes flexible knot locations and estimates them along with the spline coefficients.

1.1.2. Mean and Covariance Function

Just as in univariate or multivariate data analysis, a central goal of FDA is to describe the location and variability of the observed functions. Formally, this corresponds to estimating the mean and the covariance function of the underlying stochastic process $\{F(t)\}_t$ which are given as the pointwise mean and covariance function

$$\mu(t) = \mathbb{E}(F(t)), \quad C(t_1, t_2) = \text{Cov}(F(t_1), F(t_2))$$

for all $t, t_1, t_2 \in I$. For a set of fully observed functional observations f_1, \dots, f_n sampled independently from the process $\{F(t)\}_t$ these characteristics μ and C of the process can be estimated as

$$\begin{aligned} \hat{\mu}(t) &= \frac{1}{n} \sum_{i=1}^n f_i(t) \quad \text{for all } t \in I \\ \hat{C}(t_1, t_2) &= \frac{1}{n} \sum_{i=1}^n (f_i(t_1) - \hat{\mu}(t_1))(f_i(t_2) - \hat{\mu}(t_2)) \quad \text{for all } t_1, t_2 \in I. \end{aligned} \quad (1.3)$$

Similar to covariance estimators for multivariate data, the divisor $n - 1$ can be used instead of n to obtain an unbiased estimator.

However, since the functions are usually observed only at discrete time points and with possible additive error, only these discrete, erroneous values can be used to estimate the mean and covariance. One option is to estimate the underlying curves using the methods described in Subsection 1.1.1, and then use these estimates in the formulas given in (1.3).

Particularly simple forms for mean and covariance are obtained if one uses basis representations $\hat{f}_i(t) = \sum_{k=1}^N \hat{\xi}_{ik} b_k(t)$ for all $i = 1, \dots, n, t \in I$ with same basis functions for each $i = 1, \dots, n$ for the estimation of the underlying functions. Then the estimate for the mean becomes

$$\hat{\mu}(t) = \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^N \hat{\xi}_{ik} b_k(t) = \sum_{k=1}^N \bar{\xi}_k b_k(t)$$

for all $t \in I$, with $\bar{\xi}_k = \frac{1}{n} \sum_{i=1}^n \hat{\xi}_{ik}$ being the mean of the estimated basis coefficients and, similarly, the estimate for the covariance becomes

$$\hat{C}(t_1, t_2) = \sum_{k=1}^N \sum_{l=1}^N b_k(t_1) b_l(t_2) \hat{\Sigma}_{kl} \quad \text{where } \hat{\Sigma}_{kl} = \frac{1}{n} \sum_{i=1}^n (\hat{\xi}_{ik} - \bar{\xi}_k)(\hat{\xi}_{il} - \bar{\xi}_l)$$

for all $k, l = 1, \dots, N$ is the covariance of the estimated coefficients.

However, this two-step approach is only reasonable if a sufficient number of observations y_{ij} at time points $t_{ij} \in I$, $j = 1, \dots, m_i$ are available for each curve f_i , $i = 1, \dots, n$. In particular, one needs more observations than basis coefficients for each curve, or one must rely on a penalized estimate for each underlying function. An alternative approach exploits the fact that in the Hilbert space \mathbb{L}^2 the mean can be represented as the minimizer of the sum of squared distances to the mean, i.e. $\hat{\mu} = \operatorname{argmin}_{\mu \in \mathbb{L}^2} \sum_{i=1}^n \|f_i - \mu\|_{\mathbb{L}^2}^2$. Choosing a basis representation b_1, \dots, b_N and approximating the square of the \mathbb{L}^2 distance by the squared sum of distances at the observed time points provides the estimate $\hat{\mu} = \sum_{k=1}^N \hat{\xi}_k b_k$ where

$$\hat{\xi}_1, \dots, \hat{\xi}_N = \operatorname{argmin}_{\xi_1, \dots, \xi_N} \sum_{i=1}^n \sum_{j=1}^{m_i} \left(y_{ij} - \sum_{k=1}^N \xi_k b_k(t_{ij}) \right)^2.$$

Consequently, the mean $\hat{\mu}$ is estimated in the same way as the basis representation for each of the underlying functions in 1.2, except that the pooled data from all of the underlying functions are used simultaneously to estimate the mean. As a result, one can use any of the various basis functions described in Subsection 1.1.1 for estimation. In particular, a B-spline basis can be employed in combination with an additional roughness penalty.

Similarly, for sparsely observed functional data, it is unreliable to estimate the covariance function based on pre-smoothed individual functional observations \hat{f}_i , $i = 1, \dots, n$, as pointed out by Yao et al. (2005). They suggest estimating the covariance function $C : I \times I \rightarrow \mathbb{R}$ by local linear smoothing of the empirical covariances $(y_{ij_1} - \hat{\mu}(t_{ij_1}))(y_{ij_2} - \hat{\mu}(t_{ij_2}))$ at observed pairs of time points t_{ij_1}, t_{ij_2} with $j_1, j_2 \in I$, $j_1 \neq j_2$ and $i = 1, \dots, n$. Note that the observations on the diagonal, i.e. the observed variances, are not included in the estimation of the covariance. This is because the measurement error ϵ_j , $j = 1, \dots, m_i$ appears in the covariance $\operatorname{Cov}(Y_{ij}, Y_{ij}) = \operatorname{Var}(F_i(t_{ij})) + \operatorname{Var}(\epsilon_{ji})$ but not in the covariance of different time points, since the measurement error is assumed to be independent.

Thus, the optimization problem for this covariance smoothing procedure is very similar to that of the smooth mean estimation. The difference is that the domain of the covariance function C is $I \times I$, i.e. two-dimensional. Consequently, a method suitable for this domain is required. One approach is to select a univariate basis representation

and then use the products of these basis functions to obtain a basis with the desired domain. Using a B-spline basis as the univariate basis results in the so-called tensor product splines (e.g. Wood, 2017). These, in turn, can be accompanied by a roughness penalty to achieve additional smoothing, with the penalty parameter chosen either via REML (Goldsmith et al., 2013) or (an approximation of) leave-one-subject-out cross-validation (Xiao et al., 2018). Extensions of this covariance smoothing approach (e.g. Di et al., 2014; Cederbaum et al., 2018) have also been considered for more complex data structures.

While Xiao et al. (2018) and Cederbaum et al. (2018) account for the symmetry constraint of the covariance matrix, covariance estimates obtained by the covariance smoothing techniques described above are not guaranteed to be positive semidefinite in general. An alternative approach that incorporates this constraint is to estimate the covariance by estimating a reduced-rank mixed-effects model, which results in a reduced-rank covariance (James et al., 2001). This model has been estimated more efficiently using the Newton-Raphson procedure on the Stiefel manifold (Peng and Paul, 2009). Recently, a novel low-rank covariance estimator which is guaranteed to be positive semidefinite has been developed by Wang et al. (2022), utilizing the reproducing kernel Hilbert space framework.

1.1.3. Functional Principal Component Analysis

Covariance functions capture the variability present in observed functions. However, analyzing and understanding this variability using these covariance functions, which are a two-dimensional surfaces, is not very intuitive. Another key tool within FDA that provides a clearer visual representation and meaningful decomposition of the variability present in functional data is functional principal component analysis (FPCA).

FPCA extends principal component analysis (PCA), a technique widely used in multivariate data analysis, and adapts it to the realm of functional data. Analogous to PCA, FPCA transforms the coordinate system to an orthonormal coordinate system such that the first new basis function, the first functional principal component, is the direction along which the functional data f_1, \dots, f_n varies the most. Subsequent functional principal components are orthogonal to the previous ones and capture the remaining variance in the data, ranked in decreasing order of magnitude of the variance. Hence, in total there are n functional principal components.

As Ramsay and Silverman (2005) observe, computing the functional principal components that successively maximize the explained variance is equivalent to solving the

eigenequation

$$\left\langle \frac{1}{n} \sum_{i=1}^n f_i(\cdot) f_i(t_2), \phi \right\rangle_{\mathbb{L}^2} = \int_I \frac{1}{n} \sum_{i=1}^n f_i(t_1) f_i(t_2) \phi(t_1) dt_1 = \lambda \phi(t_2) \quad (1.4)$$

for a functional principal component $\phi \in \mathbb{L}^2$ with $\|\phi\|_{\mathbb{L}^2} = 1$ and corresponding eigenvalue $\lambda \in \mathbb{R}$. Thus, like multivariate PCA, FPCA can be framed as an eigenvalue problem, which is also true in general Hilbert spaces.

Assuming the mean $\frac{1}{n} \sum_{i=1}^n f_i$ is as usually subtracted from the observations before computing the FPCA, we have $\frac{1}{n} \sum_{i=1}^n f_i = 0$ and the bivariate function $(t_1, t_2) \rightarrow \frac{1}{n} \sum_{i=1}^n f_i(t_1) f_i(t_2)$ gives an estimate for the covariance function in the case of completely observed functional data f_1, \dots, f_n . One way to account for the fact that functional data are typically observed discretely is again to smooth the data before computing the principal components (Besse and Ramsay, 1986; Ramsay and Dalzell, 1991), where an additional penalty term can also be used to obtain further regularized principal components (Silverman, 1996; Huang et al., 2008).

However, as Yao et al. (2005) point out, relying on pre-smoothed functional observations for FPCA does not lead to robust results for sparsely observed functions. In this case, an alternative characterization of FPCA using the spectral decomposition of the covariance operator $\mathcal{C} : \mathbb{L}^2 \rightarrow \mathbb{L}^2$, $f \mapsto \int_I C(t_1, \cdot) f(t_1) dt_1$ of the underlying stochastic process $\{F(t)\}_{t \in I}$ as proposed in Dauxois et al. (1982) is beneficial. Indeed, the Karhunen-Loève theorem (Karhunen, 1946; Loève, 1946) gives the decomposition of the process $\{F(t)\}_{t \in I}$ as

$$F(t) = \mu(t) + \sum_{k=1}^{\infty} Z_k \phi_k(t) \quad (1.5)$$

where ϕ_k , $k \in \mathbb{N}$ are the orthonormal eigenfunctions of the covariance operator \mathcal{C} for the eigenvalues $\lambda_1 \geq \lambda_2, \dots \geq 0$, respectively, and Z_k are the uncorrelated principal component scores with $\mathbb{E}(Z_k) = 0$. For more details on this widely used representation of second-order stochastic processes (i.e with mean and covariance functions) refer to Hsing and Eubank (2015).

It is evident that both perspectives of FPCA, maximizing the explained variance and decomposition of the covariance of the underlying stochastic process, lead to the same functional principal components if the covariance function C for the underlying stochastic process $\{F(t)\}_{t \in I}$ is estimated as the covariance of the fully observed (or previously smoothed) functions, given in (1.3). However, an eigenvalue analysis of the covariance operator also defines FPCA for sparse observations, since one can use any estimator for the covariance function. In particular, one can use an estimator that is better suited for sparse observations, i.e. covariance smoothing, as discussed in

Subsection 1.1.2. Moreover, the stochastic process perspective has also led to further developments, such as FPCA for multivariate functions observed on different domains (Happ and Greven, 2018) and principal components for nested functional observations (Di et al., 2014).

Similar to estimating the mean and the covariance function, particularly neat forms of the principal component functions can be obtained by using the same basis to represent all functions $\hat{f}_i(t) = \sum_{k=1}^N \hat{\xi}_{ik} b_k(t)$ with $t \in I$ for all $i = 1, \dots, n$. In this case the functional principal component decomposition can be computed by solving the matrix eigenequation $\hat{\Sigma} \mathbf{G} \boldsymbol{\psi} = \lambda \boldsymbol{\psi}$ for $\boldsymbol{\psi} \in \mathbb{R}^N$ and $\lambda \in \mathbb{R}$, where $\hat{\Sigma}$ with entries $\hat{\Sigma}_{kl}$, $k, l = 1, \dots, N$ is the estimated covariance of the basis coefficients as given in Equation (1.3) and \mathbf{G} is the Gram matrix corresponding to the basis $b_1, \dots, b_N \in \mathbb{L}^2$. That is, \mathbf{G} has entries $G_{kl} = \int_I b_k(t) b_l(t) dt$ for $k, l = 1, \dots, N$. Denote by $\boldsymbol{\psi}_i = (\psi_{i1}, \dots, \psi_{iN})^T$, $i = 1, \dots, N$ the solutions to the above eigenequation. Then, the functional principal components are obtained as $\phi_i = \frac{\tilde{\phi}_i}{\|\tilde{\phi}_i\|_{\mathbb{L}^2}}$ with $\tilde{\phi}_i(t) = \sum_{k=1}^N \psi_{ik} b_k(t)$ and with corresponding eigenvalues λ_i for all $i = 1, \dots, N$ (cf. Ramsay and Silverman, 2005; Reiss and Xu, 2020). Note that in this case the number of principal components is not only bounded by the number of observations n but also by N , the number of basis functions with $N \leq n$.

When tensor product splines are used to smooth the covariance function, the computation of the functional eigenvalues can be reformulated as a matrix eigenvalue problem, similar to the case if one uses a fixed basis for pre-smoothing as described above. Thus, using a tensor product basis for smoothing yields an explicit spline representation of the estimated eigenfunctions (Reiss and Xu, 2020).

Similar to PCA, the calculation of scores, i.e. projections of the data onto the functional principal components, is crucial. These scores can serve multiple purposes. They can be used to obtain a low-dimensional representation of the original functions via truncating the sum in Equation (1.5), hence only considering a small number of principal component functions. These low-dimensional approximations of the original functions can then be used for further statistical analyses such as regression, clustering, classification or visualization. Furthermore, the scores themselves can also be used for similar analysis purposes.

If the functions f_i , $i = 1, \dots, n$ are densely observed or pre-smoothed versions are used, the principal component scores $\langle \phi_k, f_i \rangle_{\mathbb{L}^2}$, $k \in \mathbb{N}$ can be computed using numerical integration routines. However, this can be inaccurate if the curves are sparsely observed. Therefore, Yao et al. (2005) propose PACE (Principal Component Analysis through Conditional Expectation), where they predict the scores as expectations conditional on the observed data, similar to the prediction of random effects in a linear mixed model.

1.1.4. Regression for Functional Response

The objective of regression is to establish a relationship between one or more independent variables, often referred to as covariates, and a dependent variable, known as the response. While classical regression analysis (e.g. Fahrmeir et al., 2013) assumes that all variables are scalar or multivariate vectors, FDA introduces three different scenarios (see e.g. Morris, 2015; Greven and Scheipl, 2017): functional covariates and scalar or multivariate response (scalar-on-function), functional response and scalar covariates (function-on-scalar), and functional covariates and response (function-on-function). As parts of this thesis are concerned with (flexible) regression for functional object data as the response given scalar covariates, here function-on-scalar regression is most relevant among those and will be reviewed in more detail in the following. Nonetheless, it is important to note that methods developed for function-on-function regression are also useful, since function-on-scalar regression can be seen as a special case of function-on-function regression with constant functions as covariates.

The simplest case of function-on-scalar regression, the linear model for the functional observations f_i given scalar covariates x_{i1}, \dots, x_{iK} for all $i = 1, \dots, n$, is defined as

$$f_i(t) = \beta_0(t) + \sum_{k=1}^K x_{ik}\beta_k(t) + \epsilon_i(t) \quad (1.6)$$

for all $t \in I$. Here $\beta_0 \in \mathbb{L}^2$ is a functional intercept, $\beta_1, \dots, \beta_K \in \mathbb{L}^2$ denote the effect functions and $\epsilon_i \in \mathbb{L}^2$ are i.i.d. error functions for all $i = 1, \dots, n$ with $\mathbb{E}(\epsilon_i(t)) = 0$ for all $t \in I$ pointwise. This function-on-scalar model can also be seen as a special type of varying coefficient model (Hastie and Tibshirani, 1993).

Estimating model (1.6) based on discrete observations y_{ij} at time points t_{ij} , $j = 1, \dots, m_i$ for all $i = 1, \dots, n$ can, like mean estimation, be done using a basis expansion for the coefficient functions β_k , $k = 1, \dots, K$. If the remaining errors $\epsilon_i(t_{ij})$, $j = 1, \dots, m_i$, $i = 1, \dots, n$ are further assumed to be independent, this results in a linear regression problem for estimating the basis coefficients, which can be approached using penalized least squares estimation (Reiss et al., 2010). However, errors $\epsilon_i(t_{ij})$ for observations of the same function, i.e. for same $i = 1, \dots, n$, are typically correlated. To account for this Reiss et al. (2010) iterate between estimating the covariance of the error function and penalized generalized least squares estimation of the basis coefficients while Guo (2002) introduce random effects for each function to account for within-function correlation. The latter has also been used to account for the correlation between dimensions of multivariate functions (Volkman et al., 2023).

It is noteworthy that, as with unconditional mean estimation, there are several different modeling and fitting approaches besides spline expansions for the (often assumed

smooth) intercept and effect functions. If the observations on each function occur at the same time points, Fan and Zhang (2002) suggest fitting separate linear models at each time point and then smoothing the resulting coefficient functions. Hoover et al. (1998) and Wu and Chiang (2000) examine the use of kernel methods to estimate the coefficient functions, while Xie and Kong (2023) and Reimherr et al. (2023) use a reproducing kernel Hilbert space representation of the regression function, the latter even on certain manifolds, which are more complex domains than the interval $I \subset \mathbb{R}$.

Note that the linear function-on-scalar model (1.6) discussed up to this point can be extended in multiple ways. For the more general function-on-function model, a comprehensive framework for these extensions is presented in Greven and Scheipl (2017). In this paper, the authors discuss how established models for scalar data, including generalized, mixed and additive models, can be extended to accommodate functional data. In this context, one possible extension of the linear function-on-scalar model is to include non-linear effects of the scalar covariates. Besides additive structures (Scheipl et al., 2015), i.e. regression functions that are sums of non-linear functions in each covariate, interactions of covariates (Liu et al., 2023) and fully flexible regression functions based on neural network architecture (Luo and Qi, 2023) have been considered.

Another topic of recent interest is how to select the variables that actually affect the functional response, in the case of many potentially relevant scalar covariates, i.e., for large K . This variable selection problem has been addressed using various penalization approaches (Chen et al., 2016; Barber et al., 2017; Mirshani and Reimherr, 2021; Cai et al., 2022) including LASSO and elastic net type penalization. In contrast, by fitting their models with gradient boosting (e.g. Hofner et al., 2014), Brockhaus et al. (2017) and Stöcker et al. (2021) incorporate variable selection via early stopping.

1.1.5. Registration

A central aspect of the analysis of functional data discussed so far is that observations of different functions at the same time point t can be compared, e.g., to estimate a pointwise mean. However, this is not always the case for observed functional data. Often the beginning of the observed time interval varies (time shift), or the time is stretched or compressed. Thus, methods are required to align the time patterns of the observed curves. For an overview of existing approaches refer to Marron et al. (2015). An essential part of each of these approaches is to define a measure of the (dis)similarity of two functions and to give a set of possible time transformations, called re-parametrizations or warping functions.

Since the dissimilarity of two functions $f_1, f_2 \in \mathbb{L}^2$ is usually measured by their \mathbb{L}^2 distance (with smaller distances corresponding to greater similarity), it seems natural

to formulate the alignment problem as an optimization problem, where one seeks to minimize this dissimilarity, i.e. find $\operatorname{argmin}_{\gamma \in \Gamma} \|f_1 - f_2 \circ \gamma\|_{L_2}$ for a set of suitable re-parametrization/warping functions $\Gamma = \{\gamma : I \rightarrow I\}$. A common assumption here is that $\gamma \in \Gamma$ is monotonically increasing and surjective (onto). However, this choice of similarity measure and unrestricted warping is problematic, since the resulting alignment is not symmetric, which means aligning f_2 to f_1 will not be equivalent to aligning f_1 to f_2 . Furthermore, the so-called ‘‘pinching’’ effect (Marron et al., 2015) can occur, that is $\inf_{\gamma \in \Gamma} \|f_1 - f_2 \circ \gamma\|_{L_2}$ can be zero even if f_2 is not a warped version of f_1 .

To overcome this ‘‘pinching’’ problem, one can limit the slope of the warping functions. This can be done by restricting the set of possible warping functions (Ramsay and Silverman, 2005; Vitelli et al., 2010) or adding a regularization term to the dissimilarity measure, which is done in various dynamic time warping algorithms, where the first derivative of the warping function is penalized (Sakoe and Chiba, 1978; Keogh and Ratanamahatana, 2005). Ramsay and Li (1998) combine both approaches and use a penalized basis representation for the warping functions, which is similar to Wrobel et al. (2019) for a different similarity measure suited to discrete observations on exponential family functions. Alternatively, a distribution can be assumed for the warping functions in a Bayesian model (Cheng et al., 2016; Lu et al., 2017). These penalized or Bayesian approaches can also be extended to incompletely observed functions (Bauer et al., 2021; Matuk et al., 2021).

But since all of these approaches restrict the amount of warping, they do not impose a proper metric on the space of functions modulo re-parametrization, and any further analysis is not independent of the initial parametrizations of the observed functions. In contrast, Srivastava and Klassen (2016) manage to define a proper metric on the set of absolutely continuous curves \mathcal{A} (which is a subspace of \mathbb{L}^2) modulo translation (adding a constant) and re-parametrization by considering the Fisher-Rao Riemannian metric as a dissimilarity measure. They show that for two absolutely continuous functions f_1 and f_2 , the Fisher-Rao metric simplifies to the \mathbb{L}^2 -distance after applying the square-root-velocity (SRV) transformation to both functions f_1, f_2 defined as

$$q_i(t) = \begin{cases} \frac{\dot{f}_i(t)}{\sqrt{|\dot{f}_i(t)|}} & \text{if } \dot{f}_i(t) \neq 0 \\ 0 & \text{if } \dot{f}_i(t) = 0 \end{cases} \quad \text{for } i = 1, 2. \quad (1.7)$$

Here, \dot{f}_i denotes the first derivative of f_i with respect to the time $t \in I$.

Since this distance is invariant under simultaneous warping of both functions, minimizing it over all admissible warping functions yields a proper distance on the quotient \mathcal{A}/Γ modulo translation, where Γ is the set of monotonically increasing, onto and dif-

ferentiable warping functions. More precisely, this elastic distance d is given as

$$d([f_1], [f_2]) = \inf_{\gamma \in \Gamma} \|q_1 - (q_2 \circ \gamma)\sqrt{\dot{\gamma}}\|_{\mathbb{L}^2}, \quad (1.8)$$

where $[f_1], [f_2] \in \mathcal{A}/\Gamma$ denote equivalence classes with respect to warping. For more details on the SRV framework see Srivastava and Klassen (2016).

In this framework, to obtain an alignment for a set of observed functions f_1, \dots, f_n , $n \in \mathbb{N}$ one first computes the Fréchet mean $[\hat{\mu}] \in \mathcal{A}/\Gamma$ with respect to the elastic distance, i.e., the minimizer of the sum of squared distances to the observed functions (see Subsection 1.2.1 for a precise definition). Then, the functions aligned to this mean $[\hat{\mu}]$ are considered to be “aligned” to each other.

Although the mean $[\hat{\mu}]$ is completely invariant under the initial parametrization of the observed functions, one must be careful with the interpretation of it and the aligned functions. One point to note is that often, for example for visualization or further analysis, a representative of the equivalence class $[\hat{\mu}]$ needs to be chosen. Srivastava and Klassen (2016) propose to choose the representative such that the average warping of the observed functions is zero. However, this means that this representative, and thus the functions optimally aligned to that representative, are not independent of the initial parametrization of the observed functions. This must be taken into account in any further analysis such as FPCA or regression. A second point is that the alignment with respect to the elastic distance (1.8) is not transitive. This means that although the functions aligned to the mean are considered to be an aligned set of functions, any two functions in this set are in general not aligned to each other with respect to the elastic distance (1.8).

Therefore, although the SRV framework allows for a fully invariant analysis with respect to the parametrization of the observed functions, it is typically used more for separating the variation of the observed functions into amplitude (the orbit of a function modulo warping) and phase (warping) variation. In contrast, Pegoraro and Secchi (2023) propose a fully parametrization invariant analysis by representing functions as merge trees, which are identical if and only if the functions are in the same orbit modulo warping.

In summary, there are various methods available to analyze functions observed with different temporal patterns. For 1-dimensional functions, a completely parametrization invariant analysis seems to be often not desired, though, since the remaining information contains only the height of the extrema and their order. In contrast, for functions in 2 or 3 dimensions, the equivalence classes modulo re-parametrization are just the images of the functions (plus the orientation, i.e. in which direction the image of a function (the curve) is traversed). In this case, analysis of the image is often desired, for example

if they represent the outline of an object. See Subsection 1.3.2 for more details on how the SRV framework can be used for an elastic analysis of such curves.

1.2. Selected Methods from Object Data Analysis

The data objects in this thesis can naturally be understood as observations in quotient spaces of \mathbb{L}^2 function spaces. These object spaces are inherently more complex than the Hilbert space \mathbb{L}^2 discussed in the previous section, which serves as a model space for functional data. In particular, quotient spaces typically lack a linear structure, meaning that addition and scalar multiplication are not defined in these spaces. As a result, concepts developed for functional data such as FPCA (Subsection 1.1.3) and regression (Subsection 1.1.4) are not readily applicable, requiring the development of methods tailored to handle complex data structures.

To summarize the statistical analysis of complex data objects, the term “object oriented data analysis” was introduced by Wang and Marron (2007). Thus, object-oriented data analysis, or shorter object data analysis (e.g. Patrangenaru et al., 2018), is concerned with the analysis of non-Euclidean data, i.e., data that cannot be represented by unconstrained scalar numbers or multivariate vectors. In this sense, FDA is a special case of ODA, where the functions in \mathbb{L}^2 are the data objects. However, \mathbb{L}^2 has a linear structure, allowing many standard Euclidean concepts, such as mean and covariance, to be extended to functional data by pointwise analogies, as demonstrated in the previous section.

The focus in this section is on methods for object data that have no natural representation in a Hilbert space, but can be conceptualized as elements of an object space with at least a metric structure. Examples of such metric object data are the infinite dimensional covariance functions along the frequency spectrum for different languages (Pigoli et al., 2014) and phylogenetic trees, which provide graphical representations of evolutionary relationships (Holmes, 2003). However, some object data can also be understood as elements of a curved space, i.e. a manifold, which is called a Riemannian manifold if it is equipped with smoothly varying inner products and thus with a metric and a local notion of angles. Examples of object data spaces that have the structure of a Riemannian manifold are the unit sphere which serves as an object space for directional data (e.g. wind direction in Mardia and Jupp, 2000), the space of symmetric positive definite matrices used for functional connectivity analysis of Magnetic Resonance Imaging (MRI) data (You and Park, 2021) and Stiefel and Grassmanian manifolds used in computer vision and pattern recognition (Turaga et al., 2008).

These examples demonstrate that the term “object (oriented) data analysis” covers

a wide range of data problems that are modeled in a variety of mathematical object spaces. Each of these spaces requires specialized analysis methods. For a comprehensive introduction and overview of this topic, the textbook authored by Marron and Dryden (2021) provides valuable insights and discusses potential analysis approaches for several illustrative examples. The goal of this section is to highlight object data analysis methods that are relevant to this thesis, with an emphasis on methods that require only a metric structure. In the following we assume that the data objects v_1, \dots, v_n , $n \in \mathbb{N}$ are elements of an object space Υ which is at least equipped with a metric $d_\Upsilon : \Upsilon \times \Upsilon \rightarrow \mathbb{R}$.

1.2.1. Fréchet Means

The Fréchet mean (Fréchet, 1948), also sometimes known as Karcher mean (Karcher, 1977) in particular on Riemannian manifolds, generalizes the notion of a central or average object of a set of objects from Euclidean to metric spaces. For a set of observed data objects $v_1, \dots, v_n \in \Upsilon$ the Fréchet mean $\hat{\mu}_{FR}$ is defined as an element in Υ that minimizes the sum of squared distances from $\hat{\mu}_{FR}$ to each of the observed objects. That is

$$\hat{\mu}_{FR} = \operatorname{argmin}_{\mu \in \Upsilon} \sum_{i=1}^n d_\Upsilon(v_i, \mu)^2. \quad (1.9)$$

In Euclidean spaces, the Fréchet mean is the same as the conventional arithmetic mean. Other familiar location parameters can be obtained as Fréchet means when alternative distances or powers of distances are used. For example, the median can be obtained as the minimizer of the sum of absolute distances. Note that $\hat{\mu}_{FR} \subseteq \Upsilon$ is a set of objects, which may be the empty set if the minimum in Equation (1.9) is not attained, or may contain multiple elements in Υ if the minimum is not unique. However, in Euclidean spaces as well as in general Hilbert spaces the minimum is unique (see e.g. Panaretos and Zemel, 2020), i.e. the Fréchet mean contains a single element.

For a random variable $U : \Omega \rightarrow \Upsilon$ defined on a probability space (Ω, \mathcal{F}, P) , one can define the population mean, called the expected element, analogous to the Fréchet mean. By replacing the sum of the squared distances with the expectation of the squared distance, the expected element of U becomes $\operatorname{argmin}_{\mu \in \Upsilon} \mathbb{E}_P(d_\Upsilon(U, \mu)^2)$. In this sense Fréchet means are empirical versions of expected elements and a set version of the law of large numbers holds (Ziezold, 1977), i.e. the set of empirical Fréchet means converges to the set of expected elements.

However, expected elements, just like Fréchet means, are also set-valued and can be either empty or can contain more than one element. Consider for example a uniform distribution on the set $\Upsilon = [-1, 1] \setminus \{0\}$. In this case the expected element is the empty

set. Contrarily, for the uniform distribution on a sphere every point on the sphere is a valid Fréchet mean. The study of what conditions can be imposed on the metric space Υ and on the distribution of the random variable U to guarantee the existence and uniqueness of the minimizer is of ongoing interest. For example Le (1998) considers Fréchet means on (discrete) shape spaces, Charlier (2013) discusses distributions on the circle and Panaretos and Zemel (2020) consider Fréchet Means in a Wasserstein Space.

1.2.2. Regression in Metric Spaces

In addition to describing the location of observed objects using the Fréchet mean, it is often also desired to establish a relationship between objects and covariates within a regression model. In this subsection the focus is on objects v_1, \dots, v_n in the metric space Υ as response variables and scalar or multivariate covariates $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$. This regression setting may be called object-on-scalar regression, which mimics the term function-on-scalar regression used in FDA.

However, unlike the response space \mathbb{L}^2 in FDA, Υ has no notion of linearity, therefore one cannot perform addition and scalar multiplication, which is used to define function-on-scalar regression. Moreover, regression for functions in \mathbb{L}_2 yields a pointwise perspective, which allows, for example, to assume a pointwise distribution for the response, but such a pointwise perspective cannot be taken for responses in Υ . If Υ inherits only a metric structure, any intrinsic regression model can only be based on that metric.

One approach is to extend kernel regression by generalizing the Nadaraya-Watson estimator (1.1) from Euclidean data to metric data. This estimator gives the predicted response for a covariate value \mathbf{x} as a weighted average of the observed values, where the weights depend on the proximity of the observed covariates to \mathbf{x} . To obtain an analogous estimator for a metric response, one can replace the weighted average by a weighted Fréchet mean (Hein, 2009; Davis et al., 2007). Thus, a nonparametric object-on-scalar regression estimator can be expressed as

$$\hat{v}(\mathbf{x}) = \operatorname{argmin}_{v \in \Upsilon} \frac{\sum_{i=1}^n d_{\Upsilon}(v, v_i)^2 K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)},$$

where K is an appropriate multivariate kernel function. Typical kernels assign a higher weight to observations \mathbf{x}_i that are closer to \mathbf{x} and a lower weight to those farther away.

In contrast to this local averaging approach, in order to generalize linear regression Petersen and Müller (2019) developed a global model they call Fréchet regression. They

estimate their model also as a weighted Fréchet mean

$$\hat{v}(\mathbf{x}) = \operatorname{argmin}_{v \in \Upsilon} \sum_{i=1}^n \omega_i d_{\Upsilon}(v, v_i)^2,$$

but with the weights $\omega_1, \dots, \omega_n$ obtained from standard linear regression as $\omega_i = 1 + (\mathbf{x}_i - \bar{\mathbf{x}})^{\top} \hat{\Sigma}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})$ with $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$ being the mean of the observed covariates and $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\top}$ their empirical covariance matrix. Petersen and Müller (2019) show that if Υ is a separable Hilbert space (e.g. \mathbb{L}^2) their model coincides with the usual linear regression model.

Note that the weights used for Fréchet regression, unlike the weights used in the Nadaraya-Watson estimator, can be negative, resulting in a global behavior of the regression estimate. However, both approaches only implicitly define the regression function as a weighted Fréchet mean, so no global model parameters such as intercept or slope are estimated. Such interpretable model parameters seem difficult to define in general metric spaces, but there have been proposals for the case where Υ additionally has the structure of a Riemannian manifold, which are briefly presented in the following.

1.2.3. Regression on Riemannian Manifolds

Since (connected) Riemannian manifolds inherit a metric (see Lee, 2018, for an introduction to Riemannian manifolds), the methods discussed in the previous subsection can also be applied if Υ is a Riemannian manifold. But unlike in general metric spaces, on Riemannian manifolds there is the notion of the tangent space $T_{v_0} \Upsilon$ at a point $v_0 \in \Upsilon$ which gives a parametrization of all geodesics passing through v_0 . This yields an interpretation of v_0 as an intercept and the velocity $\beta \in T_{v_0} \Upsilon$ with which v_0 is traversed as a slope parameter. With this observation, several authors (Shi et al., 2009; Fletcher, 2013; Niethammer et al., 2011) propose geodesic regression as a generalization of simple linear regression for one scalar covariate.

For this geodesic model, the conditional distribution of $U \in \Upsilon$ given a covariate $x \in \mathbb{R}$ as

$$U = \operatorname{Exp}_{h(x)}(\epsilon) \quad \text{with } h(x) = \operatorname{Exp}_{v_0}(\beta x), \quad (1.10)$$

where ϵ is a random variable taking values in $T_{h(x)} \Upsilon$. Thereby, the exponential map $\operatorname{Exp}_{v_0}(\beta)$ for a given point $v_0 \in \Upsilon$ takes the tangent vector $\beta \in T_{v_0} \Upsilon$ and maps it to the point on the geodesic starting in v_0 with velocity β after one time step.

For observations $v_1, \dots, v_n \in \Upsilon$ and observed covariates $x_1, \dots, x_n \in \mathbb{R}$ the geodesic regression model can be estimated via least squares, that is

$$(\hat{v}_0, \hat{\beta}) = \operatorname{argmin}_{v_0, \beta} \sum_{i=1}^n d_{\Upsilon}(\operatorname{Exp}_{v_0}(\beta x_i), v_i)^2. \quad (1.11)$$

Fletcher (2013) proposes to use a gradient decent algorithm to find the minimizer of (1.11) which requires the computation of the inverse of the exponential map, called log map as well as taking derivatives of these. This means that for a given manifold, the concrete computability of the estimators depends on the availability of expressions for these functions.

The geodesic regression model (1.10) has been extended to include non-linear effects. For example Hong et al. (2014) still consider geodesic paths but with varying speed, and Hinkle et al. (2014) include polynomials as a generalization of geodesics. Furthermore, the geodesic model has been extended to multiple regression, i.e. including multivariate covariates (Cornea et al., 2017), and also to incorporate additive model terms (Lin et al., 2022).

1.3. Functional Object Data

In this section, a brief description of the quotient spaces derived from the function space \mathbb{L}^2 , which are considered in this thesis, will be provided. This includes a discussion of how these spaces can be equipped with an appropriate metric. For more detailed information, the reader is encouraged to consult the corresponding chapters below.

1.3.1. Functional Shapes

In this thesis, one of the quotient spaces under consideration is introduced for the purpose of modeling the shape of an object. In this context, the term “shape” refers to “what is left when the differences which can be attributed to translations, rotations, and dilatations have been quotiented out”(Kendall, 1984, p. 82). This informal definition originally motivated the field of statistical shape analysis, where a shape is defined as a collection of $m \in \mathbb{N}$ landmark points in a Euclidean space \mathbb{R}^d , $d \in \mathbb{N}$ (see e.g. Dryden and Mardia, 2016).

This shape analysis of point configurations can be extended to continuous shapes defined by continuous multivariate functions (Srivastava and Klassen, 2016), similar to how multivariate data analysis has been adapted to the analysis of functional data (see Section 1.1). More precisely, a functional shape is given by an equivalence class of continuous functions $\mathbf{f} : I \rightarrow \mathbb{R}^d$, $d \in \mathbb{N}$ with respect to the invariances translation, rotation and rescaling (dilatation). Formally, these equivalence relations \sim are defined for $\mathbf{f}_1, \mathbf{f}_2 : I \rightarrow \mathbb{R}^d$ as

- translation: $\mathbf{f}_1 \sim \mathbf{f}_2 \Leftrightarrow \exists \mathbf{a} \in \mathbb{R}^d$ such that $\mathbf{f}_1 = \mathbf{f}_2 + \mathbf{a}$,

- rotation: $\mathbf{f}_1 \sim \mathbf{f}_2 \Leftrightarrow \exists Q \in \text{SO}(d)$ such that $\mathbf{f}_1 = Q\mathbf{f}_2$. Here $\text{SO}(d)$, denotes the special orthogonal group, also called rotation group, which is the group of orthogonal matrices in $\mathbb{R}^{d \times d}$ with determinant 1,
- rescaling: $\mathbf{f}_1 \sim \mathbf{f}_2 \Leftrightarrow \exists \alpha > 0$ such that $\mathbf{f}_1 = \alpha\mathbf{f}_2$.

In order to apply methods from object data analysis (Section 1.2) to the shape equivalence classes defined by one or more of the above invariances, it is necessary to equip the set of equivalence classes with at least a metric. For translation, this is straightforward, since in this case every equivalence class can be identified with an element in \mathbb{L}_0^2 , the space of square-integrable functions with integral zero constraint, which is a Hilbert space.

This is different from rescaling, where the equivalence classes can be identified with elements on the unit sphere $S(\mathbb{L}^2)$ in \mathbb{L}^2 , which is a nonlinear space, i.e. does not have a Hilbert space structure. For this identification one needs to exclude the $\mathbf{0}$ element, which is also done in the following discussion. To equip \mathbb{L}^2 modulo rescaling with a suitable distance, various metrics defined on the unit sphere can be used. For example simply the metric of \mathbb{L}^2 restricted to the unit sphere $S(\mathbb{L}^2)$, or the geodesic distance, induced by the Riemannian manifold structure of $S(\mathbb{L}^2)$. Furthermore, a Procrustes-type distance has been proposed, which is defined as $\inf_{\alpha > 0} \|\mathbf{f}_1 - \alpha\mathbf{f}_2\|_{\mathbb{L}^2}$ for \mathbf{f}_1 and \mathbf{f}_2 on the sphere $S(\mathbb{L}^2)$.

Since the special orthogonal group $\text{SO}(d)$ acts by isometries, for the set of equivalence classes modulo rotation, a metric is canonically given by the quotient metric (Burago et al., 2001) as $\inf_{Q \in \text{SO}(d)} \|\mathbf{f}_1 - Q\mathbf{f}_2\|_{\mathbb{L}^2}$ for $[\mathbf{f}_1], [\mathbf{f}_2] \in \mathbb{L}^2/\text{SO}(d)$, which has the same form as the Procrustes distance for scaling. For this reason the distance combining both has been termed full Procrustes distance in statistical shape analysis (Dryden and Mardia, 2016, e.g.) and this term will be used for continuous shapes analogously here. That means that the full Procrustes distance for continuous shapes, i.e. functions modulo translation, rotation and rescaling is given as $\inf_{Q \in \text{SO}(d), \alpha > 0} \|\mathbf{f}_1 - \alpha Q\mathbf{f}_2\|_{\mathbb{L}^2}$ for $[\mathbf{f}_1], [\mathbf{f}_2] \in S(\mathbb{L}_0^2)/\text{SO}(d)$, where \mathbb{L}_0^2 is the subspace of \mathbb{L}^2 with functions integrating to zero.

One additional point worth noting is that, in the special case of planar shapes (i.e. equivalence classes of functions mapping to \mathbb{R}^2), the identification with the complex plane \mathbb{C} allows us to represent rotation as multiplication with imaginary elements. Consequently, simultaneous rotation and rescaling can be expressed as multiplication by a complex number, which means that the corresponding equivalence classes can be associated with the complex sphere. Thus, in this case, the equivalence classes with respect to both rotation and rescaling can be equipped with a Riemannian manifold

structure, similar to how it has been done for planar landmark shapes in Kendall (1989). However, as pointed out there, in higher dimensions, even the landmark shape space lacks a manifold structure. Therefore, one cannot expect functional shapes in three or more dimensions to possess one.

1.3.2. Elastic Curves and Shapes

For the functional shapes discussed in the previous subsection, usually only the image of the functions is of interest, not the parametrization by coordinate functions. Thus, in addition to invariance with respect to translation, rotation, and rescaling, one might also be interested in invariance with respect to re-parametrization. In fact, there are also numerous scenarios where exactly the image of the function is what represents the objects of interest, such as handwritten symbols or object outlines, where size and orientation might be considered important. In these cases, one might seek an analysis of functions that is invariant with respect to parametrization, but not invariant with respect to traditional shape invariances.

In situations where the focus is on the image of a continuous function $\mathbf{f} : I \rightarrow \mathbb{R}^d$, $d \in \mathbb{N}$, it is referred to as a curve in the following. Note, however, that the term “curve” can be somewhat ambiguous, sometimes referring to either the image or the entire function. Therefore, the term “elastic curves” is used to indicate that equivalence classes of functions modulo re-parametrization are the objects of interest.

To conduct a statistical analysis of these objects, it is necessary to establish a metric on the resulting quotient space modulo re-parametrization. Srivastava et al. (2011) propose to use the elastic distance discussed for 1-dimensional functions in Subsection 1.1.5. Analogous to (1.8), this distance for two absolutely continuous curves $\mathbf{f}_1, \mathbf{f}_2 : I \rightarrow \mathbb{R}^d$ can be calculated as $\inf_{\gamma \in \Gamma} \|\mathbf{q}_1 - (\mathbf{q}_2 \circ \gamma)\sqrt{\dot{\gamma}}\|_{\mathbb{L}^2}$ where $\mathbf{q}_i(t) = \frac{\dot{\mathbf{f}}_i(t)}{\sqrt{\|\dot{\mathbf{f}}_i(t)\|}}$ if $\dot{\mathbf{f}}_i(t) \neq 0$ and zero otherwise for $t \in I$ and $i = 1, 2$. Here $\dot{\mathbf{f}}_i = (\dot{f}_{i1}, \dots, \dot{f}_{id})$ denotes the coordinate-wise derivative with respect to t , $\|\dot{\mathbf{f}}_i(t)\|$ is its usual Euclidean norm in \mathbb{R}^d and Γ is the set of monotonically increasing, onto and differentiable warping functions.

Since the elastic distance is a proper metric on the quotient space of absolutely continuous curves modulo translation and parametrization, it enables completely parametrization invariant analysis. Unlike for one-dimensional functions, it is quite obvious that this is often the desired property. For example, if a two-dimensional function is used to represent the outline of an object, only the image itself is relevant, not the speed at which the parametrization traverses the outline, while for one-dimensional functions a restricted/penalized re-parametrization might be more appropriate.

Figure 1.1 illustrates this contrast between the one-dimensional and the two-dim-

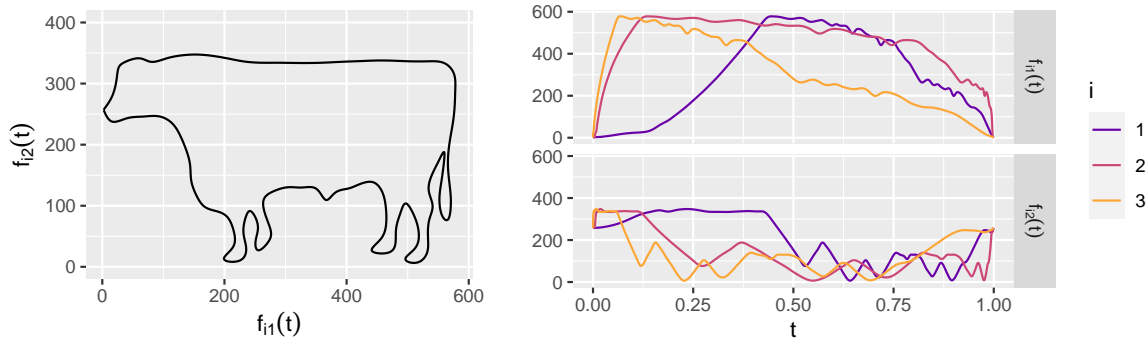


Figure 1.1.: Planar cow shape (left, obtained from the R-package “fdasrvf” Tucker, 2020) parametrized by different 2-dimensional functions $\mathbf{f}_i : [0, 1] \rightarrow \mathbb{R}^2$, $i = 1, 2, 3$, for which both coordinate functions $f_{i1}(t)$ and $f_{i2}(t)$, $t \in [0, 1]$ are displayed (right).

ensional case. In this example, the cow shape (obtained from the R-package “fdasrvf” Tucker, 2020) on the left is the image of all three two-dimensional functions shown on the right. While it is not clear whether a separate analysis of the coordinate functions should consider all three of them identically, it seem natural that, if the cow shape is the object of interest, the analysis should not depend on which of the three functions was used to create the shape.

Combining shape invariances with re-parametrization invariance, Srivastava et al. (2011) originally introduced the elastic distance in the context of functional shape data analysis, where they also proposed combined distances for all invariances rotation, scaling, translation, and parametrization. In other words, they already introduced the elastic analysis of functional shapes, which are referred to as elastic shapes in this thesis.

1.3.3. Densities

In a similar way, probability density functions $f : I \rightarrow \mathbb{R}_+$ can be modeled as equivalence classes of the function space \mathbb{L}^2 modulo rescaling, to account for the constraint that densities must integrate to one. For densities however, directly defining a distance on the quotient, as it was done for (elastic) shapes in the previous subsections, is less convenient, since this quotient must additionally be restricted to non-negative functions.

Different transformations of density functions into a Hilbert space have been proposed to accommodate both constraints, including the log hazard transformation, the log quantile density transformation and the centered log ratio (clr) transformation (see e.g. Petersen and Müller, 2016). In particular, the clr transformation establishes a one-to-one correspondence between squared-log integrable (proper and improper) density functions and the separable Hilbert space \mathbb{L}_0^2 , the space of square-integrable functions that integrate to zero. As a result, the clr transformation induces a Hilbert space

structure on the squared-log integrable (proper and improper) density functions, known as the Bayes Hilbert space (Egozcue et al., 2006; van den Boogaart et al., 2014).

To be more precise, rescaling establishes also an equivalence relation \sim on the set of squared-log integrable (proper and improper) functions, denoted as $B = \{f = \exp(g) | g \in \mathbb{L}^2(I)\}$. The resulting quotient set, denoted as $\mathcal{B} = B/\sim$, comprises equivalence classes $[f]$ for $f \in B$. This set equipped with the operations $[f_1] \oplus [f_2] = [f_1 \cdot f_2]$ for all $[f_1], [f_2] \in \mathcal{B}$ (addition), $\alpha \odot [f] = [f^\alpha]$ for all $[f] \in \mathcal{B}, \alpha \in \mathbb{R}$ (scalar multiplication) and the scalar product defined by $\langle [f_1], [f_2] \rangle_{\mathcal{B}} = \frac{1}{2|I|} \int_I \int_I \log\left(\frac{f_1(x)}{f_1(y)}\right) \log\left(\frac{f_2(x)}{f_2(y)}\right) dx dy$ for all $[f_1], [f_2] \in \mathcal{B}$ is referred to as the Bayes Hilbert space and it is isometrically isomorphic to \mathbb{L}_0^2 . Note that to allow this identification of \mathcal{B} with \mathbb{L}_0^2 , the set of density functions must be restricted to the square-log integrable densities and must be extended to include improper density functions. But every equivalence class in \mathcal{B} can either be uniquely represented by a proper density function or contains only improper densities.

In addition to the Hilbert space structure, alternative metrics have been explored for the space of probability density functions. In particular, the Wasserstein distance (Panaretos and Zemel, 2019) is a popular choice for probability measures, while the Fisher-Rao metric (Srivastava et al., 2007) introduces a manifold structure on the space of density functions. However, identifying density functions with \mathbb{L}_0^2 offers the advantage of performing statistical modeling in this well-known function space, which in principle allows to directly derive methods originally developed for functional data, such as FPCA (Hron et al., 2016) or regression (Scimone et al., 2021; Maier et al., 2022).

In practice, however, density functions are often unobserved and accessible only through discrete samples from the distributions given by each density function. This differs from the setting of discretely observed functions considered in FDA (Section 1.1), since observations from a distribution with a given density are not observations of the density with potential additive error, which would be needed to directly apply methods for sparsely observed functions. To deal with the typically discrete observations in the case of density functions, the common approach has been to estimate the observed densities using pre-processing techniques. These include methods such as aggregating the data using kernel density estimates (Maier et al., 2022), or using compositional spline estimates based on histogram data, as proposed by (Machalová et al., 2021).

1.4. Overview of Thesis Contributions in the Context of Functional Object Data Analysis

As discussed in Section 1.1, there is a wide range of methods available for functional data, including mean estimation, FPCA and flexible regression models capable of handling sparse observations and error-prone data. This work aims to help bridge the gap to the methods from ODA (Section 1.2) available for functional object data spaces discussed in Section 1.3.

To this end, Paper I and Paper II add to Fréchet mean estimation for elastic curves and shapes, respectively, in the square-root-velocity (SRV) framework in the case of irregularly or sparsely observed curves. This is achieved by introducing splines for modeling the mean in Paper I, and showing that certain splines are identifiable modulo re-parametrization. These methods are available in the R-packages “elasdics” (Steyer, 2022) and “elastes” (Pfeuffer et al., 2023).

Paper III, IV and V add to regression methods available for elastic curves, shapes and elastic shapes, respectively. In particular, Paper III proposes a generalization of linear regression not only for these response spaces but also for quotient spaces under an action by isometries, which do not need to inherit a manifold structure.

Paper VI develops FPCA for density functions in the Bayes Hilbert space taking into account that usually there is only a discrete sample from each density available.

Bibliography

- Barber, R. F., Reimherr, M., and Schill, T. (2017). The function-on-scalar LASSO with applications to longitudinal GWAS. *Electronic Journal of Statistics*, 11(1):1351–1389. 10
- Basna, R., Nassar, H., and Podgórski, K. (2022). Data driven orthogonal basis selection for functional data analysis. *Journal of Multivariate Analysis*, 189:Article 104868. 4
- Bauer, A., Scheipl, F., Küchenhoff, H., and Gabriel, A.-A. (2021). Registration for incomplete non-gaussian functional data. arXiv:2108.05634. 11
- Besse, P. and Ramsay, J. (1986). Principal components analysis of sampled functions. *Psychometrika*, 51:285–311. 7
- Brockhaus, S., Melcher, M., Leisch, F., and Greven, S. (2017). Boosting flexible functional regression models with a high number of functional historical effects. *Statistics and Computing*, 27:913–926. 10
- Burago, D., Burago, I., and Ivanov, S. (2001). *A Course in Metric Geometry*. Crm Proceedings & Lecture Notes. American Mathematical Society. 18
- Cai, X., Xue, L., Cao, J., and for the Alzheimer’s Disease Neuroimaging Initiative (2022). Robust estimation and variable selection for function-on-scalar regression. *Canadian Journal of Statistics*, 50(1):162–179. 10

- Cederbaum, J., Scheipl, F., and Greven, S. (2018). Fast symmetric additive covariance smoothing. *Computational Statistics & Data Analysis*, 120:25–41. 6
- Charlier, B. (2013). Necessary and sufficient condition for the existence of a fréchet mean on the circle. *ESAIM: Probability and Statistics*, 17:635–649. 15
- Chen, Y., Goldsmith, J., and Ogden, R. (2016). Variable selection in function-on-scalar regression. *Stat*, 5:88–101. 10
- Cheng, W., Dryden, I. L., and Huang, X. (2016). Bayesian registration of functions and curves. *Bayesian Analysis*, 11(2):447–475. 11
- Cornea, E., Zhu, H., Kim, P., Ibrahim, J. G., and the Alzheimer’s Disease Neuroimaging Initiative (2017). Regression models on Riemannian symmetric spaces. *Journal of the Royal Statistical Society: Series B*, 79(2):463–482. 17
- Cressie, N. and Wikle, C. (2011). *Statistics for Spatio-Temporal Data*. CourseSmart Series. Wiley. 1
- Dauxois, J., Pousse, A., and Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *Journal of Multivariate Analysis*, 12(1):136–154. 7
- Davis, B. C., Fletcher, P. T., Bullitt, E., and Joshi, S. C. (2007). Population shape regression from random design data. *International Journal of Computer Vision*, 90:255–266. 15
- de Boor, C. (2001). *A Practical Guide to Splines*. Applied Mathematical Sciences. Springer New York. 3
- Di, C., Crainiceanu, C. M., and Jank, W. S. (2014). Multilevel sparse functional principal component analysis. *Stat*, 3(1):126–143. 6, 8
- Dryden, I. L. and Mardia, K. V. (2016). *Statistical Shape Analysis: With Applications in R*. John Wiley & Sons. 17, 18
- Egozcue, J. J., Díaz-Barrero, J. L., and Pawlowsky-Glahn, V. (2006). Hilbert space of probability density functions based on Aitchison geometry. *Acta Mathematica Sinica*, 22:1175–1182. 21
- Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–121. 3
- Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). *Regression: Models, Methods and Applications*. Springer Berlin Heidelberg. 9
- Fan, J. and Zhang, J.-T. (2002). Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 62(2):303–322. 10
- Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Series in Statistics. Springer New York. 1, 2
- Fletcher, P. (2013). Geodesic regression and the theory of least squares on riemannian manifolds. *International Journal of Computer Vision*, 105:171—185. 16, 17
- Fréchet, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. In *Annales de l’institut Henri Poincaré*, volume 10(4), pages 215–310. 14
- Gallivan, J. and Chapman, C. (2014). Three-dimensional reach trajectories as a probe of real-time decision-making between multiple competing targets. *Frontiers in neuroscience*,

- 8:Article 215. 1
- Goldsmith, J., Greven, S., and Crainiceanu, C. (2013). Corrected confidence bands for functional data using principal components. *Biometrics*, 69(1):41–51. 6
- Greven, S. and Scheipl, F. (2017). A general framework for functional regression modelling (with discussion and rejoinder). *Statistical Modelling*, 17(1–2):1–35 and 100–115. 9, 10
- Guo, W. (2002). Functional mixed effects models. *Biometrics*, 58(1):121–128. 9
- Happ, C. and Greven, S. (2018). Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association*, 113(522):649–659. 8
- Härdle, W., Müller, M., Sperlich, S., and Werwatz, A. (2012). *Nonparametric and Semiparametric Models*. Springer Series in Statistics. Springer Berlin Heidelberg. 3
- Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(4):757–796. 9
- Hein, M. (2009). Robust nonparametric regression with metric-space valued output. In *Advances in Neural Information Processing Systems*, volume 22, pages 718–726. Curran Associates, Inc. 15
- Hinkle, J., Fletcher, P. T., and Joshi, S. C. (2014). Intrinsic polynomials for regression on riemannian manifolds. *Journal of Mathematical Imaging and Vision*, 50:32–52. 17
- Hofner, B., Mayr, A., Robinzonov, N., and Schmid, M. (2014). Model-based boosting in R: a hands-on tutorial using the R package mboost. *Computational statistics*, 29:3–35. 10
- Holmes, S. (2003). Statistics for phylogenetic trees. *Theoretical population biology*, 63(1):17–32. 13
- Hong, Y., Singh, N., Kwitt, R., and Niethammer, M. (2014). Time-warped geodesic regression. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2014: 17th International Conference, Boston, MA, USA, September 14–18, 2014, Proceedings, Part II 17*, pages 105–112. Springer. 17
- Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85(4):809–822. 10
- Horváth, L. and Kokoszka, P. (2012). *Inference for functional data with applications*, volume 200. Springer Science & Business Media. 2
- Hron, K., Menafoglio, A., Templ, M., Hruzová, K., and Filzmoser, P. (2016). Simplicial principal component analysis for density functions in Bayes spaces. *Comput. Stat. Data Anal.*, 94:330–350. 21
- Hsing, T. and Eubank, R. (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*. John Wiley & Sons. 2, 7
- Huang, J. Z., Shen, H., and Buja, A. (2008). Functional principal components analysis via penalized rank one approximation. *Electronic Journal of Statistics*, 2:678–695. 7
- Hyndman, R. J. and Ullah, M. S. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis*, 51(10):4942–4956. 1
- James, G., Hastie, T., and Sugar, C. (2001). Principal component models for sparse functional

- data. *Biometrika*, 87:587–602. 6
- Karcher, H. (1977). Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics*, 30(5):509–541. 14
- Karhunen, K. (1946). Zur Spektraltheorie stochastischer Prozesse. In *Annales Academiae Scientiarum Fennicae Series A*, volume 1, pages 23–39. 7
- Kendall, D. G. (1984). Shape manifolds, procrustean metrics, and complex projective spaces. *Bulletin of the London Mathematical Society*, 16(2):81–121. 17
- Kendall, D. G. (1989). A Survey of the Statistical Theory of Shape. *Statistical Science*, 4(2):87–99. 19
- Keogh, E. and Ratanamahatana, C. (2005). Exact indexing of dynamic time warping. *Knowledge and Information Systems*, 7:358–386. 11
- Le, H. (1998). On the consistency of procrustean mean shapes. *Advances in Applied Probability*, 30(1):53–63. 15
- Lee, J. M. (2018). *Introduction to Riemannian manifolds*, volume 2. Springer. 16
- Lin, Z., Müller, H. G., and Park, B. U. (2022). Additive models for symmetric positive-definite matrices and Lie groups. *Biometrika*, 110(2):361–379. 17
- Liu, H., You, J., and Cao, J. (2023). A dynamic interaction semiparametric function-on-scalar model. *Journal of the American Statistical Association*, 118(541):360–373. 10
- Loève, M. (1946). Fonctions aléatoires à décomposition orthogonale exponentielle. *La Revue Scientifique*, 84:159–162. 7
- Lu, Y., Herbei, R., and Kurtek, S. (2017). Bayesian registration of functions with a Gaussian process prior. *Journal of Computational and Graphical Statistics*, 26(4):894–904. 11
- Luo, R. and Qi, X. (2023). Nonlinear function-on-scalar regression via functional universal approximation. *Biometrics*. online before print. 10
- Machalová, J., Talská, R., Hron, K., and Gába, A. (2021). Compositional splines for representation of density functions. *Computational Statistics*, 36:1–34. 21
- Maier, E.-M., Stöcker, A., Fitzenberger, B., and Greven, S. (2022). Additive density-on-scalar regression in Bayes hilbert spaces with an application to gender economics. arXiv:2110.11771. 21
- Mardia, K. V. and Jupp, P. E. (2000). *Directional Statistics*. John Wiley and Sons, LTD. 13
- Marron, J. and Dryden, I. (2021). *Object Oriented Data Analysis*. Chapman & Hall/CRC Monographs on Statistics and Applied Probability. CRC Press. 14
- Marron, J. S., Ramsay, J. O., Sangalli, L. M., and Srivastava, A. (2015). Functional data analysis of amplitude and phase variation. *Statistical Science*, pages 468–484. 10, 11
- Matuk, J., Bharath, K., Chkrebti, O., and Kurtek, S. (2021). Bayesian framework for simultaneous registration and estimation of noisy, sparse, and fragmented functional data. *Journal of the American Statistical Association*, In Press:1–17. 11
- Mirshani, A. and Reimherr, M. (2021). Adaptive function-on-scalar regression with a smoothing elastic net. *Journal of Multivariate Analysis*, 185:Articel 104765. 10
- Morris, J. S. (2015). Functional Regression. *Annual Review of Statistics and its Applications*, 2:321–359. 9

- Nadaraya, E. (1964). On estimating regression. *Theory of Probability and Its Applications*, 9:141–142. 2
- Niethammer, M., Huang, Y., and Vialard, F.-X. (2011). Geodesic regression for image time-series. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2011: 14th International Conference, Toronto, Canada, September 18-22, 2011, Proceedings, Part II 14*, pages 655–662. Springer. 16
- Panaretos, V. M. and Zemel, Y. (2019). Statistical aspects of Wasserstein distances. *Annual Review of Statistics and Its Application*, 6(1):405–431. 21
- Panaretos, V. M. and Zemel, Y. (2020). *Fréchet Means in the Wasserstein Space W_2* , pages 59–74. Springer International Publishing, Cham. 14, 15
- Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076. 2
- Patrangenaru, V., Bubenik, P., Paige, R., and Osborne, D. (2018). Challenges in topological object data analysis. *Sankhya A*, 81:244—271. 13
- Pegoraro, M. and Secchi, P. (2023). Functional data representation with merge trees. arXiv:2108.13147. 12
- Peng, J. and Paul, D. (2009). A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data. *Journal of Computational and Graphical Statistics*, 18(4):995–1015. 6
- Petersen, A. and Müller, H.-G. (2016). Functional data analysis for density functions by transformation to a Hilbert space. *Annals of Statistics*, 44:183–218. 20
- Petersen, A. and Müller, H.-G. (2019). Fréchet regression for random objects with euclidean predictors. *Annals of Statistics*, 47:691–719. 15, 16
- Pfeuffer, M., Steyer, L., and Stoecker, A. (2023). *elastes: Elastic Full Procrustes Means for Sparse and Irregular Planar Curves*. R package version 0.1.7. 22
- Pigoli, D., Aston, J. A. D., Dryden, I. L., and Secchi, P. (2014). Distances and inference for covariance operators. *Biometrika*, 101(2):409–422. 13
- Ramsay, J. O. (1982). When the data are functions. *Psychometrika*, 47(4):379–396. 1
- Ramsay, J. O. and Dalzell, C. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(3):539–561. 7
- Ramsay, J. O. and Li, X. (1998). Curve registration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):351–363. 11
- Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer New York. 1, 3, 6, 8, 11
- Reimherr, M., Sriperumbudur, B., and Kang, H. B. (2023). Optimal function-on-scalar regression over complex domains. *Electronic Journal of Statistics*, 17(1):156–197. 10
- Reiss, P., Huang, L., and Mennes, M. (2010). Fast function-on-scalar regression with penalized basis expansions. *The International Journal of Biostatistics*, 6:28–28. 9
- Reiss, P. T. and Xu, M. (2020). Tensor product splines and functional principal components. *Journal of Statistical Planning and Inference*, 208:1–12. 8
- Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837. 2

- Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26:159–165. 11
- Scheipl, F., Staicu, A.-M., and Greven, S. (2015). Functional additive mixed models. *Journal of Computational and Graphical Statistics*, 24(2):477–501. 10
- Schoenberg, I. J. (1946). Contributions to the problem of approximation of equidistant data by analytic functions. part b. on the problem of osculatory interpolation. a second class of analytic approximation formulae. *Quarterly of Applied Mathematics*, 4:112–141. 3
- Scimone, R., Menafoglio, A., Sangalli, L., and Secchi, P. (2021). A look at the spatio-temporal mortality patterns in Italy during the COVID-19 pandemic through the lens of mortality densities. *Spatial Statistics*, 49:Article 100541. 21
- Shi, X., Styner, M., Lieberman, J. A., Ibrahim, J. G., Lin, W., and Zhu, H. (2009). Intrinsic regression models for manifold-valued data. *Journal of the American Statistical Association*, 5762:192–199. 16
- Silverman, B. W. (1996). Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics*, 24(1):1–24. 7
- Srivastava, A., Jermyn, I., and Joshi, S. (2007). Riemannian analysis of probability density functions with applications in vision. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE. 21
- Srivastava, A. and Klassen, E. (2016). *Functional and Shape Data Analysis*. Springer Series in Statistics. Springer New York. 11, 12, 17
- Srivastava, A., Klassen, E., Joshi, S. H., and Jermyn, I. H. (2011). Shape analysis of elastic curves in Euclidean spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(7):1415–1428. 19, 20
- Steyer, L. (2022). elasdics: Elastic analysis of sparse, dense and irregular curves. *CRAN*. R package version 1.1.1. 22
- Stöcker, A., Brockhaus, S., Schaffer, S. A., von Bronk, B., Opitz, M., and Greven, S. (2021). Boosting functional response models for location, scale and shape with an application to bacterial competition. *Statistical Modelling*, 21(5):385–404. 10
- Tucker, J. D. (2020). *fdasrvf: Elastic Functional Data Analysis*. R package version 1.9.4. 20
- Turaga, P., Veeraraghavan, A., and Chellappa, R. (2008). Statistical analysis on stiefel and grassmann manifolds with applications in computer vision. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE. 13
- van den Boogaart, K. G., Egozcue, J. J., and Pawlowsky-Glahn, V. (2014). Bayes hilbert spaces. *Australian & New Zealand Journal of Statistics*, 56(2):171–194. 21
- Vitelli, V., Sangalli, L. M., Secchi, P., and Vantini, S. (2010). Functional clustering and alignment methods with applications. *Communications in Applied and Industrial Mathematics*, 1(1):205–224. 11
- Volkman, A., Stöcker, A., Scheipl, F., and Greven, S. (2023). Multivariate functional additive mixed models. *Statistical Modelling*, 23(4):303–326. 9
- Wand, M. (2003). Smoothing and mixed models. *Computational Statistics*, 18:223–249. 4
- Wang, H. and Marron, J. (2007). Object oriented data analysis: Sets of trees. *The Annals of*

- Statistics*, 35(5):1849–1873. 13
- Wang, J., Wong, R. K. W., and Zhang, X. (2022). Low-rank covariance function estimation for multidimensional functional data. *Journal of the American Statistical Association*, 117(538):809–822. 6
- Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application*, 3(1):257–295. 1
- Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 26(4):359–372. 2
- Wood, S. (2017). *Generalized Additive Models: An Introduction with R, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press. 6
- Wrobel, J., Zipunnikov, V., Schrack, J., and Goldsmith, J. (2019). Registration for exponential family functional data. *Biometrics*, 75(1):48–57. 11
- Wu, C. O. and Chiang, C.-T. (2000). Kernel smoothing on varying coefficient models with longitudinal dependent variable. *Statistica Sinica*, 10(2):433–456. 10
- Xiao, L., Li, C., Checkley, W., and Crainiceanu, C. (2018). Fast covariance estimation for sparse functional data. *Statistics and Computing*, 28:Article 245. 6
- Xie, H. and Kong, L. (2023). Gaussian copula function-on-scalar regression in reproducing kernel hilbert space. *Journal of Multivariate Analysis*, 198:Article 105226. 10
- Yao, F., Müller, H., and Wang, J. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590. 5, 7, 8
- You, K. and Park, H.-J. (2021). Re-visiting riemannian geometry of symmetric positive definite matrices for the analysis of functional connectivity. *NeuroImage*, 225:Article 117464. 13
- Ziezold, H. (1977). 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. 14

Contributing manuscripts

2. Paper I: Elastic Analysis of Irregularly or Sparsely Sampled Curves

Paper I introduces spline functions to model Fréchet means (Subsection 1.2.1) of irregularly and sparsely observed curves within the square-root-velocity (SRV) framework (Subsection 1.3.2). It establishes identifiability statements for individual spline representations under reparameterization and addresses their limitations. Additionally, the paper demonstrates the application of the elastic distance for clustering and classification. This is exemplified by clustering undocumented walking paths on the Tempelhof field in Berlin and computing smooth mean paths for them. In a second application, the paper classifies spirals drawn in a test for Parkinson’s disease based on their distance to a smooth mean. All methods are made readily available in the R-package “elasdics”.

Contributing article:

Steyer, L., Stöcker, A., and Greven, S. (2023). Elastic analysis of irregularly or sparsely sampled curves. *Biometrics*, 79:2103–2115. DOI: 10.1111/biom.13706

Supplementary material provided in Appendix A.

Declaration on personal contributions:

The main parts of this project were carried out by the author, including the implementation in the R package “elasdics”. Sonja Greven and Almond Stöcker contributed with important and detailed advice and discussions to this research. Additionally, Almond Stöcker assisted in the development of the central research questions. This work is also a part of Almond Stöcker’s dissertation.

Elastic analysis of irregularly or sparsely sampled curves

Lisa Steyer  | Almond Stöcker  | Sonja Greven 

School of Business and Economics, Chair of Statistics, Humboldt-Universität zu Berlin, Berlin, Germany

Correspondence

Lisa Steyer, Humboldt-Universität zu Berlin, School of Business and Economics, Chair of Statistics, Unter den Linden 6, 10099 Berlin, Germany.
Email: lisa.steyer@hu-berlin.de

Funding information

Deutsche Forschungsgemeinschaft, Grant/Award Number: GR 3793/3-1

Abstract

We provide statistical analysis methods for samples of curves in two or more dimensions, where the image, but not the parameterization of the curves, is of interest and suitable alignment/registration is thus necessary. Examples are handwritten letters, movement paths, or object outlines. We focus in particular on the computation of (smooth) means and distances, allowing, for example, classification or clustering. Existing parameterization invariant analysis methods based on the elastic distance of the curves modulo parameterization, using the square-root-velocity framework, have limitations in common realistic settings where curves are irregularly and potentially sparsely observed. We propose using spline curves to model smooth or polygonal (Fréchet) means of open or closed curves with respect to the elastic distance and show identifiability of the spline model modulo parameterization. We further provide methods and algorithms to approximate the elastic distance for irregularly or sparsely observed curves, via interpreting them as polygons. We illustrate the usefulness of our methods on two datasets. The first application classifies irregularly sampled spirals drawn by Parkinson's patients and healthy controls, based on the elastic distance to a mean spiral curve computed using our approach. The second application clusters sparsely sampled GPS tracks based on the elastic distance and computes smooth cluster means to find new paths on the Tempelhof field in Berlin. All methods are implemented in the R-package "elasdics" and evaluated in simulations.

KEYWORDS

curve alignment, Fisher–Rao Riemannian metric, functional data analysis, multivariate functional data, registration, square-root-velocity transformation, warping

1 | INTRODUCTION

In the biomedical sciences, data are increasingly collected that take the form of open or closed curves $\beta : [0, 1] \rightarrow \mathbb{R}^d$, $d \in \mathbb{N}$. Examples for such curves in two or three dimensions are (human) movement patterns (e.g., Backenroth et al., 2018), handwritten letters or symbols (e.g., Dryden and Mardia, 2016; Isenkul et al., 2014), protein structures (Srivastava et al., 2010), or the outline of an (e.g., anatomic) object, such as the corpus callosum (Joshi et al., 2013). The two applications we consider in this paper concern a spiral

drawing test for the detection of Parkinson's disease, and GPS-recorded movement tracks. In most of the named cases, only the image of the curve represents the object of interest. An "elastic" analysis is then required, that is, a statistical analysis of the curves' image in \mathbb{R}^d that does not take their parameterization over $[0, 1]$ into account and is invariant under different parameterizations. Ideally, it should also yield an optimal alignment of different curves to allow point-to-point comparison, as illustrated in the example in Figure 1. As in this example, curves are often observed at a differing number of discrete points. The aim

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *Biometrics* published by Wiley Periodicals LLC on behalf of International Biometric Society.

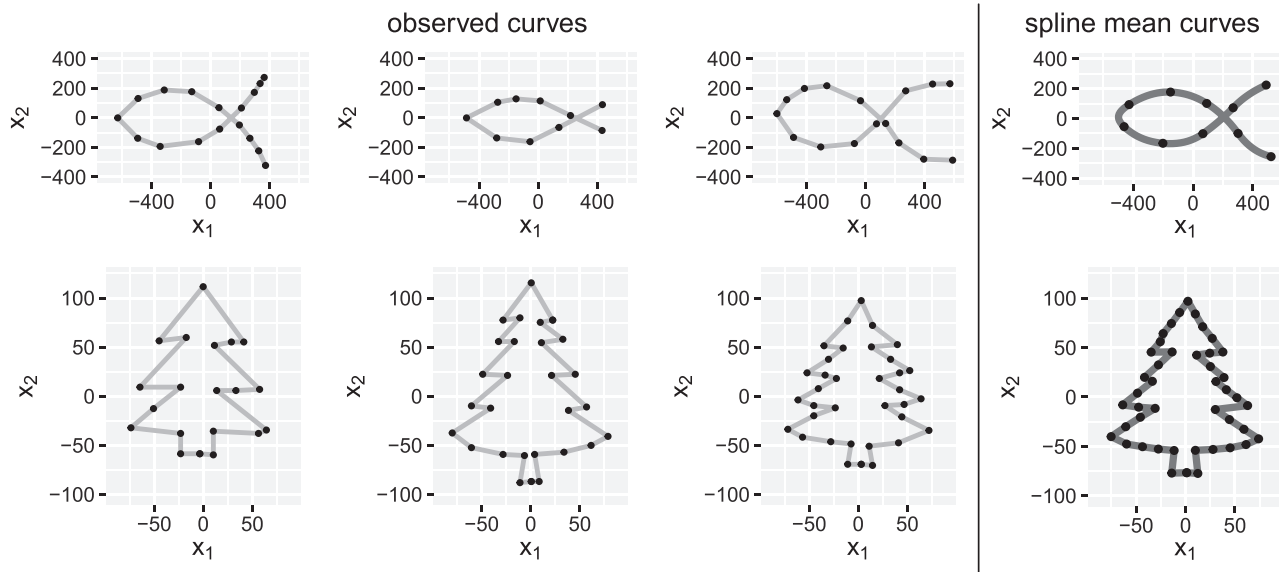


FIGURE 1 Two toy examples of sparsely and irregularly observed curves in \mathbb{R}^2 with observed points indicated as black dots and linear interpolation (first three columns). Ideally, the analysis should yield an optimal alignment of different trees curves to allow comparison of corresponding points such as bumps and other features (the mouth of the fish/the branches of the trees). Smooth or polygonal spline means (last column in dark gray) are computed using our methods, with black dots indicating values at the model-based spline knots

of this paper is to extend elastic statistical methodology to such realistic cases where curves are irregularly and sparsely sampled. In particular, we develop suitable elastic spline models for (Fréchet) mean curves of samples of such curves, and show that certain first- and second-order splines meet the identifiability properties required in a modulo parameterization context. These means can be smooth curves, such as shown for the fish in Figure 1, or polygonal curves, better suited for curves with sharp corners like the trees in Figure 1. To this end, we also propose suitable algorithms for alignment and distance computation of irregularly or sparsely sampled curves—necessary for mean computation, but also useful for distance-based analyses such as clustering or classification. In particular, we derive a useful simplification of the warping (reparameterization, alignment) problem when interpreting the observed curves as polygons.

The alignment problem for curves in \mathbb{R}^d is closely related to the registration problem in functional data analysis (Ramsay and Silverman, 2005), which corresponds to the case $d = 1$. For two functions f_1 and f_2 , warping has commonly been treated as an optimization problem $\inf_{\gamma \in \Gamma} \|f_1 - f_2 \circ \gamma\|_{L_2}$ on a suitable function space Γ of warping functions γ . This choice is problematic as $\inf_{\gamma \in \Gamma} \|f_1 - f_2 \circ \gamma\|_{L_2}$ does not define a proper distance on the space of curves modulo parameterization. The mapping is not symmetric and can be zero even if f_2 is not a warped version of f_1 , which is related to the so-called

“pinching” problem (Marron et al., 2015). Intuitively, this “pushes” the integration mass to parts of the domain where f_1 and f_2 are close. To avoid this “pinching” effect, a regularization term can be added to the loss function (Ramsay and Silverman, 2005). This is done in various dynamic time warping algorithms, where usually large values of the derivative of the warping function are penalized (Sakoe and Chiba, 1978). Alternatively, one can choose a small number of basis functions for the warping or combine both approaches to use penalized basis functions (Ramsay and Li, 1998). Moreover, Bayesian approaches to modeling warping functions have been suggested (e.g., Lu et al., 2017, or Matuk et al., 2021 for sparse one-dimensional functions).

All of these approaches restrict the amount of warping; thus, the analysis is not completely independent of the observed parameterization. This seems more suitable for one-dimensional functions ($d = 1$) where one seeks to separate phase (parameterization) and amplitude (image) but considers both as informative. If we analyze curves in \mathbb{R}^d , $d > 1$, however, we are usually only interested in the image representing the curve, that is, the equivalence class of the curve with respect to (w.r.t.) parameterization, which makes penalized, restricted, or Bayesian approaches for the warping less suitable.

Srivastava et al. (2010) propose a proper metric on the resulting quotient space via minimizing the distance between the square-root-velocity (SRV) transformed

curves. For more details on this framework, see Srivastava and Klassen (2016) and Subsection 2.1. Their perspective is focused on the curves as functions (rather than discrete observations) that, in practice, requires interpolating the curves on a regular grid for the mean computation. This works well in the case of densely observed curves. Often, however, for example, in our applications, curves are only observed at a relatively small number of discrete points, where the number differs between curves (sparse and irregular setting). We show in examples that (elastic) methods designed for densely observed curves have limitations for such sparse settings. This problem is well known in functional data analysis ($d = 1$), where spline representations or other smoothing methods are frequently used to model sparsely and/or irregularly observed functions (e.g., Greven and Scheipl, 2017; Yao et al., 2005).

The main contributions of this paper thus are to carefully introduce spline functions to model elastic (Fréchet) mean curves in \mathbb{R}^d on SRV or curve level, to show that the proposed model is identifiable via its spline coefficients modulo parameterization, and to discuss limitations of this identifiability. This extends approaches for functional data to curves in \mathbb{R}^d , $d \geq 2$ and to the elastic setting.

As part of the mean estimation, but also of interest in its own right, we also develop algorithms to align open and closed curves if at least one of them is piecewise linear, for instance, a sparsely observed curve treated as a polygon, and show local maximization properties of our algorithm for open curves. We show the usefulness of our methods for statistical analysis of irregularly or sparsely observed curves in two applications to a Parkinson spiral drawing test and to GPS movement tracks, involving mean computation, clustering, and classification of curves. Proofs of all formal statements are provided in Web Appendix B.

2 | ELASTIC ANALYSIS OF OBSERVED CURVES

In Section 2.1, we briefly review the SRV framework for analyzing curves modulo parameterization. Then, in Sections 2.2 and 2.3, we introduce our methods for elastic distance computation for irregularly or sparsely sampled curves, a building block for the spline-based Fréchet mean that we propose, and additionally of interest for distance-based analysis methods such as clustering or classification. In Sections 2.4 and 2.5, we introduce spline functions to model smooth or polygonal elastic mean curves and discuss identifiability of these modulo parameterization in Section 2.6. For all proposed methods, we focus on open curves for better readability and present adapted versions for closed curves in Web Appendix A.

2.1 | Square-root-velocity framework

Srivastava et al. (2010) show that for two absolutely continuous curves β_1 and β_2 , the Fisher–Rao metric can be simplified to the L_2 -distance between the corresponding SRV-curves, which can be minimized over the warping to obtain an elastic distance between the two curves.

Definition 1 Elastic distance; Srivastava et al., 2010. Let $\beta_1, \beta_2 : [0, 1] \rightarrow \mathbb{R}^d$ be absolutely continuous and $[\beta_1]$ and $[\beta_2]$ their respective equivalence classes modulo parameterization and translation. Then the elastic distance between $[\beta_1]$ and $[\beta_2]$ is

$$d([\beta_1], [\beta_2]) = \inf_{\gamma_1, \gamma_2 \in \Gamma} \|(\mathbf{q}_1 \circ \gamma_1) \cdot \sqrt{\dot{\gamma}_1} - (\mathbf{q}_2 \circ \gamma_2) \cdot \sqrt{\dot{\gamma}_2}\|_{L_2}, \quad (1)$$

with Γ being the set of boundary-preserving diffeomorphisms $\gamma : [0, 1] \rightarrow [0, 1]$, $\|\mathbf{q}\|_{L_2}^2 = \int_0^1 \|\mathbf{q}(t)\|^2 dt$ and SRV transformations \mathbf{q}_1 and \mathbf{q}_2 of β_1 and β_2 defined via

$$\mathbf{q}_i(t) = \begin{cases} \frac{\dot{\beta}_i(t)}{\sqrt{\|\dot{\beta}_i(t)\|}} & \text{if } \dot{\beta}_i(t) \neq 0 \\ 0 & \text{if } \dot{\beta}_i(t) = 0 \end{cases} \quad \text{for } i = 1, 2.$$

Here, $(\mathbf{q}_i \circ \gamma_i) \cdot \sqrt{\dot{\gamma}_i}$ is the SRV transformation of the reparameterized curve $\beta_i \circ \gamma_i$, $i = 1, 2$.

Srivastava and Klassen (2016) showed that it is sufficient to align one of the curves in (1),

$$d([\beta_1], [\beta_2]) = \inf_{\gamma \in \Gamma} \|\mathbf{q}_1 - (\mathbf{q}_2 \circ \gamma) \cdot \sqrt{\dot{\gamma}}\|_{L_2}. \quad (2)$$

Moreover, they pointed out that to obtain a proper quotient space structure on the space of absolutely continuous curves, we need to consider the closure of SRV-curves w.r.t. parameterization as equivalence classes. That is, for a curve β with SRV transformation \mathbf{q} , $[\beta]$ consists of all curves whose SRV transformation is in the closure of $\{(\mathbf{q}_i \circ \gamma) \cdot \sqrt{\dot{\gamma}} | \gamma \in \Gamma\}$.

Note that any analysis based on this elastic distance will be modulo translation as a result of taking derivatives. If the position of the curve in space is of interest, it has to be analyzed separately. On the other hand, if curves are used to model shape objects, translation invariance is a desired property. In classic shape data analysis (Dryden and Mardia, 2016), the analysis should additionally be invariant under rotation and scaling, and parameterization invariance presents a further key aspect in functional shape analysis (Srivastava and Klassen, 2016). In this paper, we solely discuss parameterization invariance and

give examples of handwritten spirals and GPS tracks where this elastic analysis is suitable.

A solution to the variational problem in the distance (2) is usually approximated using a dynamic programming algorithm or gradient-based optimization (e.g., in Srivastava et al., 2010). Both approaches discretize the warping space Γ . The dynamic programming algorithm, for instance, assumes a discrete grid for the domain of the warping function. An extension by Bernal et al. (2016) allows for an unequal number of points on both curves and improves computation time. Lahiri et al. (2015) provide an algorithm to align two piecewise linear curves and show that an optimal warping exists if at least one curve is piecewise linear. Such an optimal warping also exists if both curves are continuously differentiable (Bruveris, 2016).

2.2 | Elastic distance for discretely observed curves

In practice, we observe curves in \mathbb{R}^d , $d \in \mathbb{N}$, not continuously but only discretely via evaluations of these curves on discrete (and potentially sparse and curve-specific) grids. An elastic analysis needs to explicitly address this point. We propose to treat a discretely observed curve β as a polygon parameterized with constant speed between the observed corners $\beta(s_0), \dots, \beta(s_m)$. This is illustrated in the toy examples (Figure 1) with observed points marked as black dots and the polygon connecting the observations indicated by gray lines. If, as in this example, no parameterization over $[0,1]$ is given for the observed points, we will parameterize the polygon by arc length. Note that we address the case of sparsely observed curves here, whereas the problem of fragmented curves (i.e., curves with unobserved start or end points) generally cannot be handled by the proper distance defined in (1).

If β is such a polygon, the problem of finding an optimal reparameterized curve $\beta \circ \gamma$ to another arbitrary curve can be simplified (similarly as in Lahiri et al., 2015). We show that instead of solving the minimization problem (2) over the space Γ of warping functions, we only need to solve a maximization problem over a subset of \mathbb{R}^{m-1} w.r.t. the new parameterizations $t_1 = \gamma^{-1}(s_1), \dots, t_{m-1} = \gamma^{-1}(s_{m-1})$ at the observed corners.

Lemma 1. *Let β be a polygon in \mathbb{R}^d with constant speed parameterization between its corners $\beta(s_0), \dots, \beta(s_m)$. For its piecewise constant SRV transformation \mathbf{q} , denote $\mathbf{q}|_{[s_j, s_{j+1}]} = \mathbf{q}_j \in \mathbb{R}^d$ for all $j = 0, \dots, m-1$. Let $\tilde{\beta}$ be an absolutely continuous curve with SRV transformation \mathbf{p} , $\|\mathbf{p}\|_\infty < \infty$. Then calculating the optimal γ in (2) to obtain the elastic distance $d([\beta], [\tilde{\beta}])$ is equivalent to the following problem:*

$$\text{Maximize } \Phi(\mathbf{t}) = \sum_{j=0}^{m-1} \sqrt{(s_{j+1} - s_j) \int_{t_j}^{t_{j+1}} \langle \mathbf{p}(t), \mathbf{q}_j \rangle_+^2 dt} \quad (3)$$

$$\text{w.r.t. } \mathbf{t} = (t_1, \dots, t_{m-1}), \quad 0 = t_0 \leq t_1 \leq \dots \leq t_m = 1,$$

where $\langle \cdot, \cdot \rangle_+$ denotes the positive part of the scalar product in \mathbb{R}^d . For a maximizer \mathbf{t} of (3), there is a $\gamma : [0, 1] \rightarrow [0, 1]$ with $\gamma(t_j) = s_j$ for all $j = 1, \dots, m-1$ that minimizes (2).

The proof includes an explicit construction of the minimizing warping function $\gamma \in \bar{\Gamma}$ (or a minimizing sequence of warping functions), where $\bar{\Gamma}$ is the set of absolutely continuous curves $\gamma : [0, 1] \rightarrow [0, 1]$, onto and with $\dot{\gamma} \geq 0$ almost everywhere. The statement for Γ follows as Γ is dense in $\bar{\Gamma}$ and the warping action of $\bar{\Gamma}$ continuous (Bruveris, 2016). Thus, the warping problem can be simplified if one of the SRV-curves is piecewise constant, independent of the form of the second SRV-curve \mathbf{p} . If \mathbf{p} is at least continuous, for example, the SRV-curve of a model-based smooth mean curve like the fish mean in Figure 1 on the top right, the loss function in (3) is differentiable. We propose to tackle the remaining maximization problem with a gradient descent algorithm that can handle linear constraints (for instance, method BFGS in `constrOptim` from R-package “stats;” R Core Team, 2020) and provide a derivation of the gradient in Web Appendix B.

2.3 | Elastic distance for two piecewise linear curves

We present an algorithm that can be used to find an optimal warping function, and therefore, compute the elastic distance, when both curves are piecewise linear. This is relevant either because we model one of the curves as a linear spline (mean) (see Subsection 2.4), as we do for the tree shapes in Figure 1, or because we want to compute the elastic distance between two observed curves, for example, two different discretely observed fish or trees. The latter allows any distance-based analysis of the data such as clustering or classification.

To obtain an optimal warping for a curve with piecewise constant SRV transformation \mathbf{q} to a curve with SRV transformation \mathbf{p} , we first note that the maximization in one t_j direction of the objective function in (3) only depends on the current values of t_{j-1} and t_{j+1} for any \mathbf{p} . If \mathbf{p} is also a piecewise constant SRV-curve, we can even derive a closed-form solution of the maximization problem in (3) w.r.t. each $t_j \in [t_{j-1}, t_{j+1}]$ (cf. Web Appendix B). Hence, we propose a coordinate wise maximization procedure in Algorithm 1, iterating updates of odd and even indices.

Algorithm 1: Elastic distance for two open polygons

Input: piecewise constant SRV-curves \mathbf{p}, \mathbf{q} ; convergence tolerance $\epsilon > 0$;
 starting values $0 \leq t_1^{(0)} \leq \dots \leq t_{m-1}^{(0)} \leq 1$ // e.g. relative arc length
for $k \in \mathbb{N}$ **do**
 for $j = 1, \dots, m - 1$ **do**
 if $j - k$ *even* **then**
 $t_j^{(k)} = \operatorname{argmax}_{t_j \in [t_{j-1}^{(k-1)}, t_{j+1}^{(k-1)}]} \Phi |_{\{t_{j'} = t_j^{(k-1)}, j' \neq j\}}$
 else if $j - k$ *odd* **then**
 $t_j^{(k)} = t_j^{(k-1)}$
 if $\|\mathbf{t}^{(k)} - \mathbf{t}^{(k-2)}\| < \epsilon$ *and* $\|\mathbf{t}^{(k-1)} - \mathbf{t}^{(k-3)}\| < \epsilon$ **then**
 return $\mathbf{t}^{(k)} = (t_1^{(k)}, \dots, t_{m-1}^{(k)})$

The warping problem for two (open) piecewise linear curves has been previously discussed by Lahiri et al. (2015). They propose a precise matching algorithm, which produces a globally optimal reparameterization of \mathbf{q} , but is arguably demanding to implement. Our algorithm can be seen as an alternative, which is much more straightforward to understand and to extend to the closed case (cf. Web Appendix A) not explicitly addressed by Lahiri et al. (2015). We provide an implementation in the R-package “elastics.” Although our algorithm does not guarantee finding a globally optimal solution, we observe convincing results in simulations (Section 3) and can prove local maximization in the following sense:

Theorem 1. *Every accumulation point of the sequence $(\mathbf{t}^{(k)})_{k \in \mathbb{N}} = (t_1^{(k)}, \dots, t_{m-1}^{(k)})_{k \in \mathbb{N}}$ resulting from Algorithm 1 is a local maximizer of Φ in (3).*

To prove this theorem, we first establish that the directional derivatives exist and are nonpositive for all coordinate directions. Then we show that this carries over to all directional derivatives using local concavity of the objective function.

If the sequence $(\mathbf{t}^{(k)})_{k \in \mathbb{N}}$ has more than one accumulation point, they all give the same value $\Phi(\mathbf{t})$. They then correspond to different reparameterizations of the second curve, but give the same distance between the two curves. This can happen as the warping problem does not guarantee unique solutions (see Web Appendix C for an example). In practice, one can pick any maximizing \mathbf{t} to obtain a locally optimal warping function. As we cannot guarantee this \mathbf{t} to also be a global maximizer, we propose using varying starting points to find a global maximum.

Our algorithm computes the elastic distance between two piecewise linear and continuous curves. These curves form a subspace in the space of absolutely continuous curves and are called splines of degree 1. For modeling smooth (differentiable) curves, for example, for a mean function, a spline space of a higher degree may be more suitable.

2.4 | Modeling spline curves or spline SRV-curves

As common in functional data analysis (Ramsay and Silverman, 2005), we like to model curves or means for samples of curves as splines. This is in particular beneficial for sparsely observed curves, which cannot be evaluated at arbitrary points. Moreover, splines impose parsimonious models for smooth curves, which can help to avoid overfitting the observed curves given limited information.

Definition 2 (Spline curves). We call $\xi = (\xi_1, \dots, \xi_d)^T : [0, 1] \rightarrow \mathbb{R}^d$ with $d \in \mathbb{N}$ a d -dimensional spline curve of degree $l \in \mathbb{N}_0$ if all its components $\xi_1, \dots, \xi_d : [0, 1] \rightarrow \mathbb{R}$ are spline curves of degree l with a common knot set $0 = \kappa_0 < \kappa_1 < \dots < \kappa_{K-1} < \kappa_K = 1$ for some $K \geq 2$. That means that ξ_1, \dots, ξ_d are piecewise polynomial of degree l between the knots $\kappa_0, \dots, \kappa_K$, as well as continuous and $(l - 1)$ -times continuously differentiable on the whole domain $[0, 1]$ for $l \geq 1$. Denote by $S_{K; \kappa_0, \dots, \kappa_K}^l$ the set of all such spline curves.

We can either model the curve β as a d -dimensional spline curve, or its SRV transformation \mathbf{p} (see Figure 2). If β is a spline of degree $l \geq 2$, the corresponding SRV-curve \mathbf{p}

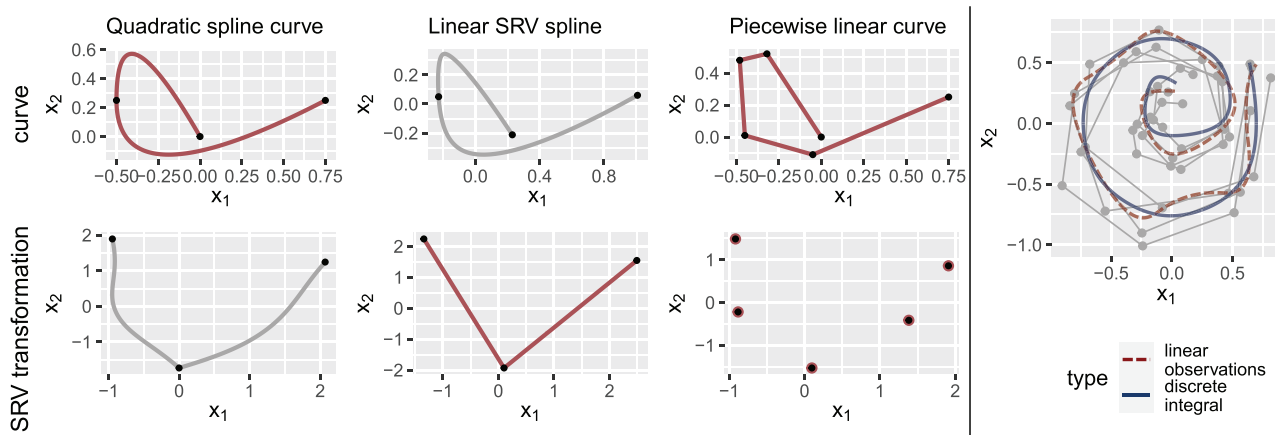


FIGURE 2 Left: Two-dimensional curves and corresponding SRV transformations. Spline curves are plotted as red curves with their values at knots marked as black dots; other curves are gray. Note that the SRV-curve in the sixth panel is piecewise constant in t and t is not visible in the image. Right: Smooth means (with 11 knots each) for four spiral curves based on linear splines on SRV level. The dashed mean curve is based on assuming piecewise linear observations for the integral approximations and the solid mean curve is based on the integral approximation using the mean value theorem

will not be a spline curve. The same holds true for curve β if \mathbf{p} is a spline of degree $l \geq 1$. Only if β is piecewise linear ($l = 1$), then both β and its piecewise constant SRV transformation are splines. However, if we use linear spline curves, we need a large number of knots to obtain similarly smooth curves as using linear splines on SRV level, and thus, expect less parsimonious models.

To use these spline curves or spline SRV-curves as model spaces modulo warping, we need to ensure model identifiability, that is, that each equivalence class contains at most one spline curve. The unique spline representative then allows to identify and interpret the equivalence class of a curve modulo warping via its spline basis coefficients. We will see in Subsection 2.6 that this is true for quadratic or cubic splines on curve level and for linear spline SRV-curves (under mild conditions). Linear spline curves are identifiable under additional assumptions.

Therefore, we can use the space of cubic, quadratic, or linear spline curves as a model space for smooth curves. However, using quadratic or cubic splines on the curve level would not imply a vector space structure on the SRV level, where the distance is computed. We therefore propose to consider linear spline (and thus continuous) SRV-curves to model smooth curves. If \mathbf{p} is a continuous SRV transformation of β , the backtransform $\beta(t) = \beta(0) + \int_0^t \mathbf{p}(s) \|\mathbf{p}(s)\| ds$ is differentiable, as the norm $\|\cdot\|$ is also continuous. Alternatively, constant spline SRV-curves can be used to model less regular, polygonal mean curves. We thus work with a linear or constant spline model on SRV level in the following.

2.5 | Elastic means for samples of curves

As the space of curves modulo parameterization and translation does not form a Euclidean space, standard statistical techniques for describing probability distributions cannot be applied directly. In particular, we cannot define the expected value as an integral or the mean as a weighted average, which would require a linear structure of the space. To generalize the mean as a notion of location to arbitrary metric spaces, Fréchet (1948) proposed to use its property of being the minimizer of the expected squared distances.

Definition 3 Fréchet mean; Fréchet, 1948. Let (Ω, \mathcal{F}, P) be a probability space and \mathcal{X} a metric space with distance function d , equipped with the Borel- σ -Algebra. For a random variable $X : \Omega \rightarrow \mathcal{X}$, we call every element in $\operatorname{arginf}_{A \in \mathcal{X}} E_P(d(X, A)^2)$ an expected element of X . For a set of observations $x_1, \dots, x_n \in \mathcal{X}$, we define the Fréchet mean as an element in $\operatorname{arginf}_{A \in \mathcal{X}} \sum_{i=1}^n d(x_i, A)^2$.

Thus, Fréchet means are empirical versions of expected elements and neither of them need to exist or be unique. For a uniform distribution on the sphere, for example, every point on the sphere is a valid Fréchet mean. This nonuniqueness can occur for the elastic distance as well, see the example given in Web Appendix C. Nevertheless, Ziezold (1977) showed a set version of the law of large numbers for the Fréchet mean, which means that for independently and identically distributed random variables

$X_1, \dots, X_n : \Omega \rightarrow \mathcal{X}$, the set of Fréchet means converges to the set of the expected elements.

As discussed in the previous subsection, we propose to use linear or constant splines on SRV level as model spaces for the Fréchet mean. For a set of curves with SRV transformations $\mathbf{q}_1, \dots, \mathbf{q}_n$ and for a given degree $l \in \{0, 1\}$ and a given set of knots $\kappa_0, \dots, \kappa_K$, we thus define

$$\bar{\mathbf{p}} \in \underset{\mathbf{p} \in S_{K;\kappa_0, \dots, \kappa_K}^l}{\operatorname{arginf}} \sum_{i=1}^n \inf_{\gamma_i} \left\| \mathbf{p} - (\mathbf{q}_i \circ \gamma_i) \sqrt{\gamma_i} \right\|_{L_2}^2 \quad (4)$$

as the SRV transformation of the spline Fréchet mean (i.e., SRV transformation of the Fréchet mean restricted to the spline SRV space) w.r.t. the elastic distance (2). The corresponding restricted Fréchet mean $\bar{\boldsymbol{\beta}}$ is thus either a polygon or a smooth curve. Similarly to the proposal of Srivastava and Klassen (2016) for densely observed curves, we tackle the minimization problem (4) with an iterative approach in Algorithm 2, alternating between

Algorithm 2: Elastic spline Fréchet mean

Input: SRV transformations \mathbf{q}_i of discretely observed curves $\boldsymbol{\beta}_i, i = 1, \dots, n$;

initial mean $\bar{\mathbf{p}}_{new} = \operatorname{arginf}_{\bar{\mathbf{p}}} \sum_{i=1}^n \|\bar{\mathbf{p}} - \mathbf{q}_i\|_{L_2}^2$; convergence tolerance $\epsilon > 0$

while $\|\bar{\mathbf{p}}_{old} - \bar{\mathbf{p}}_{new}\| > \epsilon$ **do**

$\bar{\mathbf{p}}_{old} = \bar{\mathbf{p}}_{new}$;

$\gamma_i = \operatorname{arginf}_{\gamma} \|\bar{\mathbf{p}}_{old} - (\mathbf{q}_i \circ \gamma) \sqrt{\gamma}\|_{L_2}^2, \quad \forall i = 1, \dots, n$; // warping step

$\bar{\mathbf{p}}_{new} = \operatorname{arginf}_{\bar{\mathbf{p}}} \sum_{i=1}^n \|\bar{\mathbf{p}} - (\mathbf{q}_i \circ \gamma_i) \sqrt{\gamma_i}\|_{L_2}^2$ // L_2 spline fit via (weighted)

 least-squares

return $\bar{\mathbf{p}} = \bar{\mathbf{p}}_{new}$

fitting the mean and optimizing the warping for each of the observations, but now using our warping approach for sparse curves and modeling the mean with a constant or linear spline. If we were to model the Fréchet mean in a spline space on curve level instead of SRV level, the mean fitting step would be a minimization problem in a nonlinear space, hence more challenging. That is why we refrain from using splines on curve level, although we show that quadratic and cubic splines are identifiable via their coefficients as well (Theorem 2).

For the warping step, we update the optimal warpings γ_i of the observed curves $\boldsymbol{\beta}_i, i = 1, \dots, n$ via interpreting them as observed polygons with piecewise constant SRV transformations $\mathbf{q}_i, i = 1, \dots, n$, as in Lemma 1. We tackle the remaining maximization problem (3) using a gradient descent algorithm as discussed before if $\bar{\mathbf{p}}$ is piecewise linear and Algorithm 1 if $\bar{\mathbf{p}}$ is piecewise constant. In the L_2 spline fitting step, the integrals

$\|\bar{\mathbf{p}} - (\mathbf{q}_i \circ \gamma_i) \sqrt{\gamma_i}\|_{L_2}^2$ in the sum need to be approximated, because the curves $\boldsymbol{\beta}_i$ are only observed on a finite grid $0 = s_{i,0} \leq s_{i,1} \leq \dots \leq s_{i,m_i} = 1$, and the SRV-curves $\mathbf{q}_1, \dots, \mathbf{q}_n$ are thus unobserved. One option is to assume that the SRVs \mathbf{q}_i of the observed curves are piecewise constant as in the warping step. As $\bar{\mathbf{p}}$ is piecewise linear, $(\mathbf{q}_i \circ \gamma_i) \sqrt{\gamma_i}$ also is (see proof of Lemma 1 in Online Appendix B), which leads to a closed-form solution of the integral. Alternatively, we derive an approximation of the integrals in the L_2 fitting step of Algorithm 2 using the mean value theorem and the monotonicity of the warping in Web Appendix B.5. Both approaches lead to a (weighted) least-squares problem for the spline coefficients of $\bar{\mathbf{p}}$. (An adapted algorithm for closed curves in Web Appendix A uses an additional penalty for openness with increasing weight.) We compare them using an example in Figure 2 on the right, where the second approach here leads to a better fit of the estimated spiral shape (and is used in the following).

2.6 | Identifiability of spline curves

We model curves or means for samples of curves using basis representations. If we study equivalence classes of curves modulo reparameterization, we have to ensure unique spline representatives in each class, meaning that elements of the quotient space are identifiable via their basis coefficients. To see why this is not self-evident, consider as a simple counterexample in \mathbb{R}^1 the space of quadratic polynomials $P : [0, 1] \rightarrow \mathbb{R}$, a subspace of the quadratic spline space. Note that $\gamma_a(x) = ax^2 + (1 - a)x$ defines a feasible warping function for all $a \in]0, 1[$, because γ_a is differentiable with $\gamma'_a(x) \geq 0$ and $\gamma_a(0) = 0, \gamma_a(1) = 1$. Hence, all quadratic polynomials of the form $P(x) = p_1 \gamma_a(x) + p_0$ with $p_0, p_1 \in \mathbb{R}$ are elements of the same equivalence class, although they have varying basis coefficients $ap_1, (1 - a)p_1$ and p_0 for $a \in]0, 1[$ w.r.t. the monomial basis expansion. This counterexample shows in

particular that one-dimensional spline functions do not have unique representatives in the space of functions modulo reparameterization. Moreover, every 1d function is in the orbit of a linear spline with at least as many knots as the function has local extrema. As identifiability is essential in any modeling approach, it is fortunate that in contrast to $d = 1$, we can show that in \mathbb{R}^d with $d \geq 2$, nearly all quadratic or cubic spline curves have unique basis representations.

Theorem 2. *Let $d \geq 2$ and $Q, P : [0, 1] \rightarrow \mathbb{R}^d$ be quadratic or cubic spline curves, where Q has a nonlinear image between each of its knots. Moreover, let $\gamma : [0, 1] \rightarrow [0, 1]$ be monotonically increasing and onto. Then $P = Q \circ \gamma \Rightarrow \gamma = id$.*

Thus, nearly all equivalence classes modulo reparameterization contains at most one spline curve. Hence we can identify these curves modulo warping via their spline basis coefficients. The only exception are splines with linear image, which occur if and only if the splines in each coordinate direction are multiples of each other modulo translation. Note that we do not make any assumptions on the knots here, in particular the knots could be different for Q and P . That means there is almost always a unique representative modulo warping in $\bigcup_{K, \kappa_0, \dots, \kappa_K} S_{K; \kappa_0, \dots, \kappa_K}^l$ for given $l = 2, 3$, that is, in the union of all spline spaces with varying (also varying number of) knots. Considering only quadratic or cubic splines is crucial, as this statement is not true for nonprime spline degrees. We show a counterexample for splines of degree four in Web Appendix C. The result for cubic spline curves also implies uniqueness of representatives for linear spline SRV-curves, another useful result for identifiable modeling of elastic curves.

Corollary 1. *Let $\beta_1, \beta_2 : [0, 1] \rightarrow \mathbb{R}^d$ with SRV functions q_1 and q_2 , respectively. If q_1 and q_2 are nowhere constant linear splines and $q_2(t) = q_1(\gamma(t))\sqrt{\dot{\gamma}(t)}$, then $q_1 = q_2$.*

In summary, the space of linear SRV spline curves seems particularly suitable to model smooth elastic curves as they are identifiable, that is, there is a unique representation in this space, and the corresponding curves are differentiable, which leads to visually smooth curves. In our toy example, we used linear spline SRV-curves to model the smooth fish mean (Figure 1, top right).

Remark 1 (Linear spline curves). Linear spline curves or equivalently piecewise constant SRV-curves are identifiable via their spline basis coefficients modulo warping, if we consider one spline space $S_{K; \kappa_0, \dots, \kappa_K}^1$ but not the union of several such spaces, and assume that the curve is not differentiable at all of its knots (i.e., no knot is superfluous). For an illustration, see Web Appendix C.

Hence, with this weaker identifiability result, piecewise constant SRV-curves are a suitable model space as well, with curves modeled as polygons. This is more appropriate for mean curves that are assumed to have sharp corners, like the trees in Figure 1.

As we use these spline spaces for estimation of smooth or polygonal curves, we need the following result on continuity of the embedding. It allows us to interpret estimated coefficients—for instance, compare the coefficients of two estimated group means to investigate local differences—as it ensures convergence of the spline coefficients if we construct a converging sequence of curves. For instance, we aim to construct such a sequence for the elastic mean in Algorithm 2. We show that this continuity property holds whenever the model space Ξ is a (subset of a) finite-dimensional spline space of the following form. Note that, for simplicity, we do not consider unions of spline spaces here.

Definition 4. Let Ξ be one of the following for given fixed $K \geq 2, 0 = \kappa_0 < \dots < \kappa_K = 1$: (i) a subset of $S_{K; \kappa_0, \dots, \kappa_K}^l, l = 2, 3$, which consists of identifiable splines as described in Theorem 2, additionally centered (i.e., with integral zero) to account for translation; (ii) a set of identifiable curves with linear spline SRV-curves in $S_{K; \kappa_0, \dots, \kappa_K}^1$ from Corollary 1; or (iii) the set of curves with piecewise constant SRV-curves in $S_{K; \kappa_0, \dots, \kappa_K}^0$ from Remark 1.

Lemma 2 (Topological embedding). *Let $f : (\Xi, \|\cdot\|) \rightarrow (\mathcal{A}, d)$ be the embedding of the spline coefficients defining the functions in Ξ , equipped with the usual Euclidean distance $\|\cdot\|$, into the space \mathcal{A} of absolutely continuous curves w.r.t. the elastic distance d . Then f is a topological embedding, that is, f is a homeomorphism on its image.*

Thus, the distance of spline coefficients and the elastic distance of curves modulo translation are topologically equivalent on suitable spline spaces. Consequently, a sequence of curves converges w.r.t. the spline coefficients if, and only if, it converges w.r.t. the elastic distance. Overall, we see that any spline model Ξ in Definition 4 yields an identifiable model for the Fréchet mean of observed curves, with the possibility to interpret spline coefficients. This also holds for converging series of estimators which we aim to construct in our algorithms.

3 | SIMULATION

We test our methods, which we made available for public use in the R-package “elasdics,” on simulated data. A first simulation focuses on the special case of equal numbers of observed points on the curves, where we can

compare our methods to an existing implementation of the SRV framework in the R package “*fdasrvf*” (Tucker, 2020) based on Srivastava et al. (2010). Results presented in Web Appendix D show that Algorithm 1 (and its variant for closed curves) produce clearly better alignment for sparsely and irregularly sampled curves. The corresponding average elastic distance is smaller for our method in all cases, for example, a reduction of 25% and 26% on average for 30 observed points per curve in the open and closed setting, respectively. As expected, this difference decreases if 90 points of the closed butterfly shapes are selected (1% reduction on average), as in this case, the points are nearly observed on a regular, fairly dense grid, which is the setting “*fdasrvf*” is designed for. This simulation also shows that a highly unbalanced distribution of observed points on the curves causes difficulties for the mean computation in “*fdasrvf*” as well, which is not the case for our methods.

Here we mainly discuss the second simulation, focusing on the convergence and the identifiability of the newly proposed spline means and their associated coefficients. As we vary the number of points per curve, there is no competitor to compare our methods with. For a given template curve β with known B-spline coefficients $\vartheta_1, \dots, \vartheta_B$, we generate a sample of observed curves β_1, \dots, β_n by independently sampling the coefficients $\vartheta_{i,b} \sim \mathcal{N}(\vartheta_b, \sigma^2)$ for all $i = 1, \dots, n$, $b = 1, \dots, B$. If the template curve is closed, we additionally close the sampled curves via minimizing a penalty function penalizing openness in gradient direction. The penalty is given in Web Appendix A for estimating a closed mean. The points $t_{i,1}, \dots, t_{i,m_i-1}$ on which β_i is observed are sampled uniformly on $[0, 1]$, where the number of observed points m_i is sampled uniformly either from $\{10, \dots, 15\}$ (very sparse and unbalanced) or $\{30, \dots, 50\}$ (less sparse but unbalanced).

Examples for curves sampled with standard deviation $\sigma = 4$ from a heart-shaped template curve, modeled as linear spline on SRV level with 10 equally spaced inner knots, are displayed in Figure 3. Two further examples for open curves are given in Web Appendix C. The samples in the very sparse setting are hardly recognizable as heart shapes (Figure 3, right). However, the elastic mean curve over $n = 5$ observations, estimated using the true knot set and linear SRV splines to allow a comparison of estimated and true coefficients, represents the original heart surprisingly well even in this challenging setting. We repeated this simulation 100 times each for varying numbers of observations $n \in \{5, 20\}$ and observed points per curve m_i (Figure 3, left). For $m_i \in \{10, \dots, 15\}$ observations per curve, we generally obtain a heart-shaped mean, which seems smaller and shows less pronounced features than

the template. Increasing the number of observed curves from $n = 5$ to $n = 20$ decreases the variance of the mean curve, but a certain bias due to undersampling the curves remains. Likewise, the variance of the spline mean coefficients is smaller for $n = 20$ than for $n = 5$, but their distribution is still not centered at the coefficients of the template (indicated as black dots in Figure 3).

If we increase the number of points on each curve to $m_i \in \{30, \dots, 50\}$, the estimated means w.r.t. the elastic distance adapt closer to the template. Moreover, the variance of the estimated spline coefficients decreases as well as their distance to the template. The reduction of variance indicates convergence of the spline coefficients for $n \rightarrow \infty$, although we do not expect them to precisely converge to the coefficients of the template in this simulation setup, not even if $m_i \rightarrow \infty$ for all $i = 1, \dots, n$. This is because we draw the sample curves β_1, \dots, β_n such that β is the mean w.r.t. the L_2 distance on SRV level, but this does in general not imply that β is the mean w.r.t. the elastic distance. Nevertheless, we expect this difference to be small, as the coefficients in the rightmost boxplot are close to the black dots that indicate the template’s coefficients. In addition, their low variance for $n = 20$ confirms our theoretical results on identifiability of spline coefficients in our model (Corollary 1) and continuity of the embedding (Lemma 2).

As expected, the run time of our elastic mean algorithm grows with the number of observed curves as well as with the number of observed points per curve. On a standard Windows PC, we report run times of 19 s ($n = 5$) and 30 s ($n = 20$) on average for one mean in the very sparse setting. In the less sparse setting, $m_i \in \{30, \dots, 50\}$, the run times increase to 22 and 88 s for $n = 5$ and $n = 20$, respectively.

So far, we have discussed the convergence of correctly specified spline means, as in this case, convergence of elastic means corresponds to convergence of the corresponding spline coefficients (Lemma 2). As correct specification is questionable in practice, we demonstrate the behavior of our methods in the case of model misspecification (varying spline degree and number of knots) in a further simulation given in Web Appendix D. We observe that both smooth and polygonal means reproduce the original template well and that results are not very sensitive to the number of knots, given that it is sufficiently large. Generally, the elastic distance to the template decreases for an increasing number of knots. Distances to the template are smaller for the smooth than for the polygonal model means for a fixed number of knots, and decrease to a lower level, indicating more parsimonious models and less undersampling bias for truly smooth means when using linear SRV-curve models.

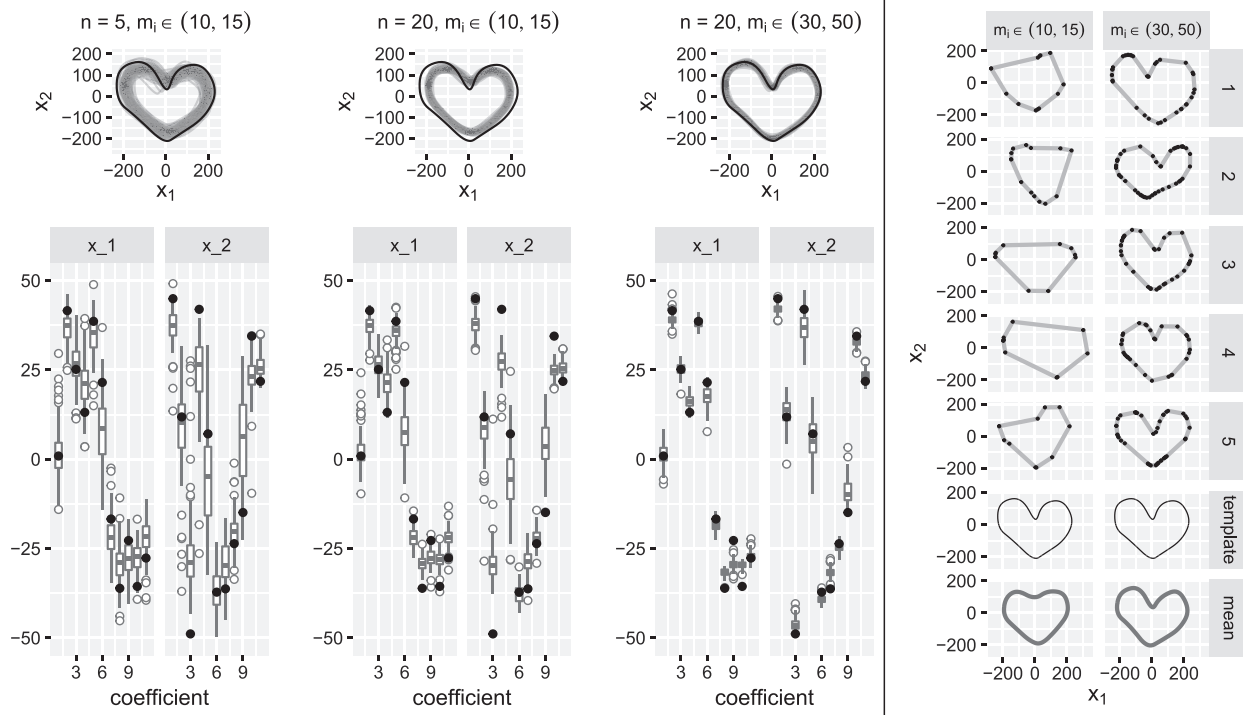


FIGURE 3 Top left: Smooth means (in gray) computed for a set of n simulated curves drawn from the heart-shaped template curve (in black) via sampling its B-spline coefficients from a normal distribution with standard deviation $\sigma = 4$ and $m_i, i = 1, \dots, n$ points observed per curve. The means are computed using linear SRV splines and the same knot set as the template (10 equally spaced inner knots). Bottom left: Corresponding distribution of spline mean coefficients (in gray) and template coefficients (in black). Right: Simulated data $i = 1, \dots, 5$ with observed values marked as black dots and corresponding smooth elastic means over $n = 5$ observations in gray

4 | APPLICATIONS ON REAL DATA

As our main goal is to develop statistical (elastic) analysis methods for discretely observed data curves, we demonstrate their practicality on two datasets.

4.1 | Classifying spiral curve drawings for detecting Parkinson's disease

(Isenkul et al., 2014) provide a dataset of spiral curve drawings by Parkinson patients and healthy controls in a so-called Archimedes spiral-drawing test, which is a common, noninvasive tool for diagnosing patients with Parkinson's disease. The data have been obtained in two different settings: In the “static spiral test,” the participants had to follow a template on a digital tablet; in the “dynamic test,” the template curve appeared and disappeared in certain time intervals. We propose an intuitive classifier mimicking a doctor's decision of the form: Classify as “Parkinson” if the distance of the drawn curve to the template curve exceeds a threshold for one or for both of the settings. As the template curve has not been recorded, we use the elastic mean (see Subsection 2.5) of all curves

from the static spiral test with piecewise constant splines and 201 knots on SRV level, instead. Then we compute the elastic distance of each observed spiral curve to the template using Algorithm 1. We report a leave-one-curve-out cross-validated accuracy of 72.5% for the static, 90.0% for the dynamic setting, and 92.5% for the classifier based on both, which indicates good separation in particular for the dynamic spiral test.

A detailed description of our analysis and a comparison to the methods implemented in the “fdatasrvf” package can be found in Web Appendix E. Our methods lead to better classification accuracy in this application and the mean calculation proves to be faster.

4.2 | Clustering and modeling smooth means of GPS-tracks

The second dataset is an example of increasingly common human movement data and comprises GPS waypoints tracked on Tempelhof Field, a former airfield (up to 2008) in Berlin, which is now used as a recreation area. The dataset consists of 55 paths with 15–45 waypoints each, recorded by members of our working group using their

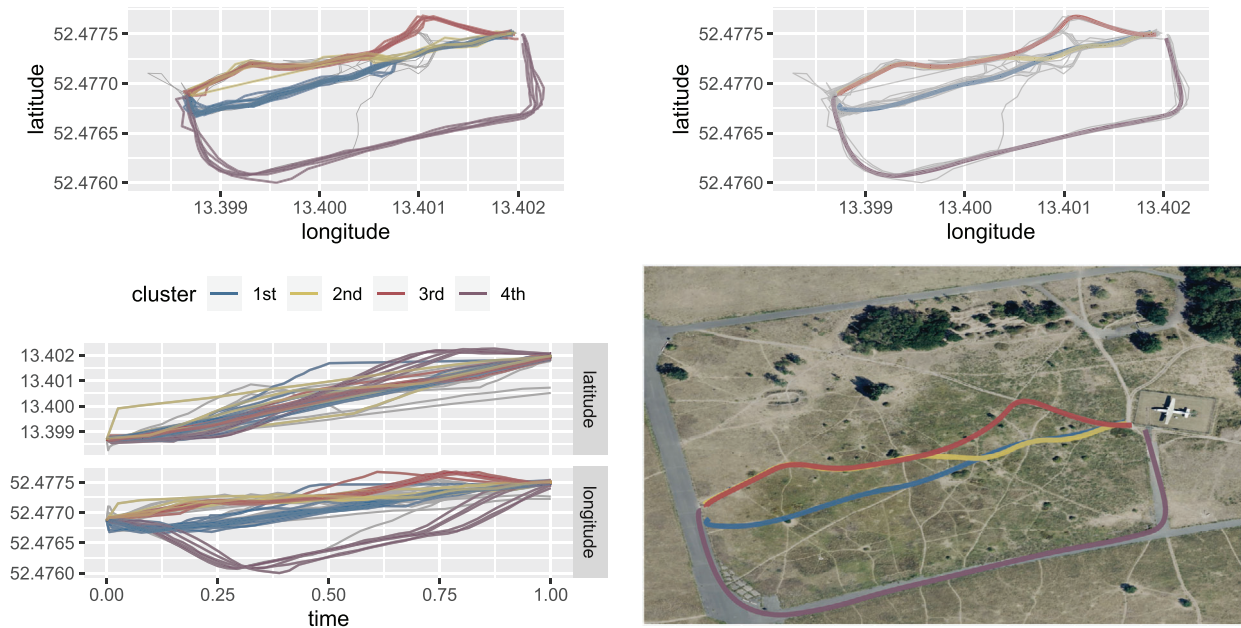


FIGURE 4 Top left: The observed trajectories with elements of the four largest clusters indicated by color. Bottom left: Longitude and latitude for the trajectories (with the four largest clusters indicated by the same colors) over relative time. Top right: Smooth means modeled as linear SRV-curves with 10 inner knots for the four largest clusters and centered at the mean center of the observed paths per cluster to account for translation. Bottom right: Cluster means plotted on Microsoft Bing Map accessed via the R package “OpenStreetMap” (Fellows, 2019)

mobile phones for tracking. Due to the variety of mobile devices used, the number of points per curve differs considerably, resulting in irregularly and quite sparsely observed data. We are solely interested in analyzing the paths (Figure 4, bottom right) the participants walked on, not the trajectories over time. Separately looking at longitude and latitude over time suggests that the individuals had quite different walking patterns and did not move with constant speed. This implies that standard (nonelastic) functional data analysis is not suitable here.

Clustering and smooth mean estimation allow us to recover the paths that the individuals walked on. In a further step, these could be used to identify new paths on Tempelhof field not yet included in existing maps. In a first step, the tracks are clustered using average linkage based on the elastic distance and the elbow criterion for stopping. Here we apply Algorithm 1 to approximate the pairwise distance between the sparsely observed open tracks. In a second step, we compute a smooth elastic Fréchet mean for each of the four largest clusters using Algorithm 2 and linear splines on SRV level with 10 inner knots. The clustering result displayed in Figure 4, top row, is visually satisfying. Looking at longitude and latitude separately clearly indicates that clustering based on the L_2 distance would not work well.

The smooth mean curves for each of the four largest clusters (Figure 4, top right) seem to describe the observed tracks well, despite the dimension reduction (24 spline

coefficients compared to 30–90 observations per curve) and also match the actual paths visible in the satellite image (Figure 4, bottom right) provide by Microsoft Bing and made available for R in the package “OpenStreetMap” (Fellows, 2019).

5 | DISCUSSION

Although our approach addresses the discrete and often sparse nature of observed curves explicitly, the interpretation as polygons with observed values at the corners underestimates the curvature of the real unobserved curves. This leads to a kind of shrinkage bias for the estimated elastic mean for sparsely observed curves. Although this bias toward curves with smaller curvature decreases with increasing observations per curve, it would be of interest to develop correction methods for (very) sparse settings in future work.

We have shown that the SRV splines modulo parameterization used for modeling the elastic mean is in general identifiable via their coefficients and we have confirmed this result in simulations. Although we did not explicitly address the choice of the optimal number of knots for such splines, a further simulation has shown that the estimation of the mean curve is not sensitive to the specific spline degree and choice of knots, given the number of knots is sufficiently large. As the union of any spline

space with fixed degree but varying knots is dense in the space of absolutely continuous curves w.r.t. the elastic distance, using an increasing number of knots would ensure that the mean curve can be arbitrarily well approximated. For a finite dataset, this would lead to overfitting the curves though, which may be addressed via penalized estimation, although the interpretation of coefficients and convergence properties would need to be studied in this setting.

Another appealing direction for further research is to include our methods for sparsely and irregularly sampled curves in existing approaches for functional shape analysis. Here the curves have to be aligned w.r.t. scaling and/or rotation in addition to the alignment w.r.t. parameterization and translation. As this is usually done iteratively, it seems promising to combine this with the iterative warping and mean fitting steps in our methods. Furthermore, elastic mean estimation for irregularly and/or sparsely sampled curves can be seen as a first step toward elastic regression models for such data. That means our methods might be useful building blocks for modeling curves or shapes depending on covariates using splines.

ACKNOWLEDGMENTS

The authors gratefully acknowledge funding by grant GR 3793/3-1 from the German research foundation (DFG). We thank the members of the Chair of Statistics who contributed to data collection on Tempelhof field, and Manuel Pfeuffer for alerting us to the Parkinson's data.

DATA AVAILABILITY STATEMENT

The data that support the findings in this paper are available in the Supporting Information of this article, with the exception of the data analyzed in the Parkinson's spirals application, which are available from <https://www.researchgate.net/publication/291814924>.

ORCID

Lisa Steyer  <https://orcid.org/0000-0002-6987-1520>

Almond Stöcker  <https://orcid.org/0000-0001-9160-2397>

Sonja Greven  <https://orcid.org/0000-0003-0495-850X>

REFERENCES

- Backenroth, D., Goldsmith, J., Harran, M.D., Cortes, J.C., Krakauer, J.W. & Kitago, T. (2018) Modeling motor learning using heteroscedastic functional principal components analysis. *Journal of the American Statistical Association*, 113(523), 1003–1015.
- Bernal, J., Dogan, G. & Hagwood, C.R. (2016) Fast dynamic programming for elastic registration of curves. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 111–118).
- Bruveris, M. (2016) Optimal reparametrizations in the square root velocity framework. *SIAM Journal on Mathematical Analysis*, 48(6), 4335–4354.
- Dryden, I. & Mardia, K. (2016) *Statistical shape analysis: with applications in R*. Wiley Series in Probability and Statistics. Chichester: Wiley.
- Fellows, I. (2019) *OpenStreetMap: Access to Open Street Map Raster Images*. R package version 0.3.4.
- Fréchet, M. (1948) Les éléments aléatoires de nature quelconque dans un espace distancié. In: *Annales de l'institut Henri Poincaré*, volume 10 (pp. 215–310).
- Greven, S. & Scheipl, F. (2017) A general framework for functional regression modelling. *Statistical Modelling*, 17(1–2), 1–35.
- Isenkul, M., Sakar, B., Kursun, O. et al. (2014) Improved spiral test using digitized graphics tablet for monitoring Parkinson's disease. In: *The 2nd International Conference on e-Health and Telemedicine (ICEHTM-2014)*, volume 5 (pp. 171–175).
- Joshi, S.H., Narr, K., Phillips, O., Nuechterlein, K., Asarnow, R., Toga, A. & Woods, R. (2013) Statistical shape analysis of the corpus callosum in schizophrenia. *NeuroImage*, 64, 547–559.
- Lahiri, S., Robinson, D. & Klassen, E. (2015) Precise matching of PL curves in R^N in the square root velocity framework. *Geometry, Imaging and Computing*, 2, 133–186.
- Lu, Y., Herbei, R. & Kurtek, S. (2017) Bayesian registration of functions with a Gaussian process prior. *Journal of Computational and Graphical Statistics*, 26(4), 894–904.
- Marron, J.S., Ramsay, J.O., Sangalli, L.M. & Srivastava, A. (2015) Functional data analysis of amplitude and phase variation. *Statistical Science*, 30(4), 468–484.
- Matuk, J., Bharath, K., Chkrebti, O. & Kurtek, S. (2021) Bayesian framework for simultaneous registration and estimation of noisy, sparse, and fragmented functional data. *Journal of the American Statistical Association*, 1–17, in press.
- R Core Team. (2020) *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Ramsay, J. & Silverman, B. (2005) *Functional data analysis*. Springer Series in Statistics. New York, NY: Springer.
- Ramsay, J.O. & Li, X. (1998) Curve registration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2), 351–363.
- Sakoe, H. & Chiba, S. (1978) Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26, 159–165.
- Srivastava, A. & Klassen, E. (2016) *Functional and shape data analysis*. Springer Series in Statistics. New York: Springer.
- Srivastava, A., Klassen, E., Joshi, S. & Jermyn, I. (2010) Shape analysis of elastic curves in Euclidean spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(7), 1415–1428.
- Steyer, L. (2021) *elasdics: elastic analysis of sparse, dense and irregular curves*. R package version 0.2.0.
- Tucker, J.D. (2020) *fdasrvf: elastic functional data analysis*. R package version 1.9.7.
- Yao, F., Müller, H.G. & Wang, J.L. (2005) Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470), 577–590.
- Ziezold, H. (1977) 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* (pp. 591–602). Springer.

SUPPORTING INFORMATION

Web Appendices A, B and C referenced in Section 2 and Web Appendix D referenced in Sections 3 and 4 are available with this paper at the Biometrics website on Wiley Online Library. All developed methods are implemented in the R-package *elasdics* (Steyer, 2021) available on CRAN and the code to reproduce the findings of this paper is available in the [Supporting Information](#) of this article.

Figure 1: First three iterations of the algorithm for closed mean curves on a toy dataset

Figure 2: Left: Two piecewise linear curves in gray with Frechet mean curves in red and blue

Figure 3: Three constant SRV splines (right) with corresponding linear spline curves (middle)

Figure 4: Comparison of the optimal alignment produced by our method CWO and the one computed with DP

Figure 5: Elastic means for irregularly sampled curves

Figure 6: Example simulated data in gray with observed values marked as black dots and corresponding smooth elastic means over $n = 5$ observations in blue

Figure 7: Top: Smooth means (in blue) computed for a set of n curves drawn from the open template curve (in red) via sampling its B-spline coefficients from a normal distribution with standard deviation $\sigma = 0.3$ and $m_i, i=1, \dots, n$ points observed per curve

Figure 8: Top: Smooth means (in blue) computed for a set of n curves drawn from the open template curve (in red) via sampling its B-spline coefficients from a normal distribution with standard deviation $\sigma=0.4$ and $m_i, i=1, \dots, n$ points observed per curve

Figure 9: Left: Smooth mean based on linear splines on SRV level with varying number of knots and therefore coefficients computed on a sample of 20 curves with $m_i \in \{30, 50\}$ points per curve

Figure 10: Left: Spiral curves drawn by either a healthy control group or by patients with Parkinson's disease in two different settings

Figure 11: Left: Distance of the curves drawn by the participants to the mean spiral curve for both settings

Figure 12: Optimal warping in both settings separated by the actual status and the predicted status using the classifiers based on only the corresponding distance each and leave-one-out cross-validation

Table 1: Classification accuracy in the dynamic setting with a varying fraction of points per curve

Table 2: Comparison of the classification accuracy in the dynamic setting with a varying number of points per curve

Table 3: Run-times for the mean computation of the spiral data in seconds

Figure 13: Left: Comparison of means for the spirals in the static setting with 100 observations per curve

Data S1

How to cite this article: Steyer, L., Stöcker, A., and Greven, S. (2023). Elastic analysis of irregularly or sparsely sampled curves. *Biometrics*, 79, 2103–2115. <https://doi.org/10.1111/biom.13706>

3. Paper II: Elastic Full Procrustes Analysis of Plane Curves via Hermitian Covariance Smoothing

In Paper II, elastic (Fréchet) mean estimation for irregularly/sparsely sampled curves is extended to elastic planar shapes (see Subsection 1.3.2), referring to 2-dimensional curves with respect to translation, rotation, rescaling, and reparameterization. This is achieved by combining the elastic distance with the full Procrustes distance for shapes (Subsection 1.3.1). The focus on planar shapes allows for the identification of shapes with complex-valued functions, which helps to establish a link between full Procrustes mean estimation and covariance estimation in irregular/sparse functional data analysis. For this purpose, Hermitian covariance smoothing is developed as a generalization of symmetric covariance smoothing (Subsection 1.1.2) for complex-valued stochastic processes. This allows the derivation of an estimator of the full Procrustes mean when the shapes are sparsely observed. The performance of the elastic shape mean estimation is then illustrated in the phonetic analysis of tongue shapes during speech production.

Contributing article:

Stöcker, A., Pfeuffer, M., Steyer, L., and Greven, S. (2022). Elastic Full Procrustes Analysis of Planar Curves via Hermitian Covariance Smoothing. *arXiv pre-print*, arXiv:2203.10522

Declaration on personal contributions:

Based on Almond Stöcker's proposal for estimating means of functional two-dimensional shapes, the author of this thesis and Almond Stöcker jointly developed the idea of an elastic extension and co-supervised Manuel Pfeuffer's master's thesis on this topic. The author provided advice in all phases of the project and made significant contributions to the proofs. This work is also part of Almond Stöcker's dissertation.

Elastic Full Procrustes Analysis of Plane Curves via Hermitian Covariance Smoothing

Almond Stöcker^{1,2,*}, Manuel Pfeuffer¹, Lisa Steyer¹, and Sonja Greven¹

¹Chair of Statistics, School of Business and Economics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany

²Department of Mathematics, École polytechnique fédérale de Lausanne (EPFL), Station 8, CH-1015 Lausanne, Switzerland

December 15, 2022

Abstract

Determining the mean shape of a collection of curves is not a trivial task, in particular when curves are only irregularly/sparsely sampled at discrete points. We newly propose an elastic full Procrustes mean of shapes of (oriented) plane curves, which are considered equivalence classes of parameterized curves with respect to translation, rotation, scale, and re-parameterization (warping), based on the square-root-velocity (SRV) framework. Identifying the real plane with the complex numbers, we establish a connection to covariance estimation in irregular/sparse functional data analysis. We introduce Hermitian covariance smoothing and show how to employ this extension of existing covariance estimation methods for obtaining an estimator of the (in)elastic full Procrustes mean, also in the sparse case not yet covered by existing (intrinsic) elastic shape means. For this, we provide different groundwork results which are also of independent interest: we characterize (the decomposition of) the covariance structure of rotation-invariant bivariate stochastic processes using complex representations, and we identify sampling schemes that allow for exact observation of derivatives/SRV transforms of sparsely sampled curves. We demonstrate the performance of the approach in a phonetic study on tongue shapes and in different realistic simulation settings, *inter alia* based on handwriting data.

Keywords: Complex Gaussian process; Functional data; Phonetic tongue shape; Principal component analysis; Shape analysis; Square-root-velocity.

1 Introduction

When comparing the shape of, say, a specific outline marked on medical images across different patients, the concrete coordinate system used for recording is often arbitrary and not of interest: the shape neither depends on positioning in space, nor on orientation or size. Analogously, the outline can be mathematically represented via a parameterized curve $\beta : [0, 1] \rightarrow \mathbb{R}^2$, but the particular parameterization of the outline curve is often not of interest, only its image. We study datasets where an observational unit is the shape of a plane curve, defined as equivalence class a) over the shape invariances translation, rotation and scale and b) over re-parameterization. More specifically, we generalize the notion of a full Procrustes mean from discrete landmark shape analysis (Dryden and Mardia, 2016) to elastic shape analysis of curves, in particular to achieve improved estimation properties in irregular/sparsely measured scenarios compared to existing “intrinsic” elastic mean shape estimation methods relying on geodesic distances. To allow this generalization of landmark shape means to curves, we also present two results characterizing the covariance structure of rotation-invariant bivariate stochastic processes via their complex representations. To enable derivative-based elastic analysis of sparsely/irregularly sampled curves, which are common in practice but for which existing methods have problems, we provide a result on the feasibility to exactly observe the necessary derivatives under such sampling. While these results are important building blocks in preparing the proposed elastic Full Procrustes mean estimation, they are also of independent interest in their own right.

For landmark shapes, different notions of mean shape are well-established including, in addition to the full Procrustes mean, in particular also the intrinsic shape mean, i.e. the Riemannian center of mass in the shape space. Dryden et al. (2014) discuss properties of different shape mean concepts, pointing out that the

full Procrustes mean is more robust with respect to outliers than the intrinsic mean or the *partial* Procrustes mean fixing scale to unit size. Further discussion of these three mean concepts, which all present Fréchet means based on different distances, can be found in Huckemann (2012). The full Procrustes mean also arises as the mode of a complex Bingham distribution (Kent, 1994) on (unit-norm) landmark configurations $\mathbf{X} \in \mathbb{C}^k$ of k landmarks, which is commonly used to model planar landmark shapes, identifying the real plane $\mathbb{R}^2 \cong \mathbb{C}$ with the complex numbers. Moreover, it corresponds to the leading eigenvector of the complex covariance matrix of \mathbf{X} , an important point we generalize for the estimation strategy proposed for curve mean shapes in this paper.

Compared to landmark shapes, different additional challenges arise for shapes of curves: invariance with respect to re-parameterization (warping) is one that is highly related to the registration problem in function data analysis (FDA, Ramsay and Silverman, 2005). In the context of shape analysis of curves, Srivastava et al. (2011) propose an *elastic* re-parameterization invariant metric, allowing to define a proper distance between two curves via optimal warping alignment. Greatly simplifying the formulation of the metric by working with square-root-velocity (SRV) transformations of the curves, this lead to a rapidly growing literature on *functional* shape analysis of curves in the SRV-framework (see e.g., Srivastava and Klassen, 2016). However, so far the focus lay on elastic generalization of the intrinsic shape mean instead of the (potentially more robust) full Procrustes mean which we generalize here. In simulations, we illustrate how the novel elastic full Procrustes mean estimation yields improved mean estimates in irregularly/sparsely sampled data (sometimes even as an estimator of the intrinsic shape mean, compared to existing estimators designed for this alternative mean).

Sparsely/irregularly observed curves have been considered in the SRV-framework by Steyer et al. (2022), however, only restricting to re-parameterization invariance and not investigating shape means. Such data with a comparatively low number of samples per curve often results in practice when the sampling rate of a measurement device is limited, or the resolution of images used for curve segmentation is coarse. In FDA, sparse/irregular functional data is commonly distinguished from dense/regular data, as it requires explicit treatment. Models for sparse/irregular data are often based on smooth (spline) function bases and commonly involve assumption of (small) measurement errors on the discrete curve evaluations (Greven and Scheipl, 2017).

Focusing on shape analysis of sparsely/irregularly measured curves, we consider the full Procrustes mean concept particularly attractive due to its robustness known from landmark shape analysis, and due to its direct connection to the covariance structure of the data, which allows relying on a core estimation strategy in sparse/irregular FDA: following Yao et al. (2005), covariance smoothing has become a major tool for sparse/irregular FDA, allowing to reconstruct the functional covariance structure based on sparse evaluations. Cederbaum et al. (2018); Reiss and Xu (2020) discuss (symmetric) tensor-product spline smoothing for this purpose, considering univariate functional data. Happ and Greven (2018) generalize univariate approaches to conduct functional principal component analysis also for multivariate sparse/irregular data.

In this paper, our contributions are to 1. characterize the complex covariance (decomposition) of rotation-invariant bivariate stochastic processes. This gives us the basis to 2. develop Hermitian covariance smoothing, which we 3. use to propose a covariance-based estimation method for the 4. novel (elastic) full Procrustes means we propose as a more robust notion of elastic shape mean, with a particular focus also on sparsely/irregularly sampled curves. For such realistic curve measurements we 5. characterize scenarios where exact sampling of the necessary SRVs/derivatives is feasible.

In the following, we first discuss in Section 2 complex stochastic processes as random elements of Hilbert spaces, illustrating their convenience for rotation-invariant bivariate FDA and propose Hermitian tensor-product smoothing for complex functional principle component analysis. This lays the groundwork for the second part of the paper in Section 3, where we introduce the notion of elastic (and inelastic) full Procrustes mean shapes of plane curves based on the SRV-framework. We show conditions under which exactly observing SRVs (i.e., curve derivatives) of sparsely/irregularly measured curves is feasible and propose estimation of their full Procrustes means via Hermitian covariance smoothing. Finally, we present an elastic full Procrustes analysis of tongue outlines observed from participants of a phonetic study and validate the proposed approach in three simulation scenarios in Sections 5 and 4. Proofs for all propositions are given in an online supplement. A ready to use implementation is offered in the R-package `elastes` (github.com/mpff/elastes).

2 Hermitian covariance smoothing

2.1 Complex processes and rotation invariance

Although functional data analysis traditionally focuses on Hilbert spaces over \mathbb{R} (compare, e.g., Hsing and Eubank, 2015), underlying functional analytic statements cover Hilbert spaces over \mathbb{C} as well (e.g., Rynne and Youngson, 2007). This lets us formulate principal component analysis for complex-valued functional data and underlying concepts in analogy to the real case in the following. Subsequently, we present two results on the relation of complex to bivariate (real) functional data and on the convenience of a complex viewpoint under rotation invariance that will be key in our estimation approach. Although complex stochastic processes have been discussed in the literature (Neeser and Massey, 1993), we are not aware of any previous discussion of the results we present in this section. In the complex viewpoint, the real plane \mathbb{R}^2 is identified with the complex numbers \mathbb{C} via the canonical vector space isomorphism $\kappa : \mathbb{C} \rightarrow \mathbb{R}^2$, $z \mapsto \mathbf{z} = (\Re(z), \Im(z))^\top$ mapping $z \in \mathbb{C}$ to its real part $\Re(z)$ and imaginary part $\Im(z)$. By z^\dagger we denote the complex conjugate $\Re(z) - \mathbf{i}\Im(z)$ of $z \in \mathbb{C}$, with $\mathbf{i}^2 = -1$, or more generally the Hermitian adjoint (conjugate transpose) for complex matrices or operators. Rotation of $\mathbf{z} \in \mathbb{R}^2$ by $\omega \in \mathbb{R}$ radians simplifies to scalar multiplication $\exp(\mathbf{i}\omega)z \in \mathbb{C}$ in complex representation.

Let Y be a complex-valued stochastic process with realizations $y : \mathcal{T} \rightarrow \mathbb{C}$ in $\mathbb{L}^2(\mathcal{T}, \mathbb{C})$, where \mathcal{T} is a compact metric space with finite measure ν . Here, $\mathcal{T} = [0, 1]$ is typically the unit interval with ν the Lebesgue measure, and $t \in \mathcal{T}$ is referred to as ‘‘time’’. The complex, separable Hilbert space $\mathbb{L}^2(\mathcal{T}, \mathbb{C})$ of square-integrable complex-valued functions is equipped with the inner product $\langle x, y \rangle = \int x^\dagger(t)y(t) d\nu(t)$ for $x, y \in \mathbb{L}^2(\mathcal{T}, \mathbb{C})$ and the corresponding norm $\|\cdot\|$.

Definition 1. *i) Y is called random element in a real or complex Hilbert space \mathbb{H} if $\langle x, Y \rangle$ is measurable for all $x \in \mathbb{H}$ and the distribution of Y is uniquely determined by the (marginal) distributions of $\langle x, Y \rangle$ over $x \in \mathbb{H}$.*

ii) The mean $\mu \in \mathbb{H}$ and covariance operator $\Sigma : \mathbb{H} \rightarrow \mathbb{H}$ of a random element Y are defined via $\langle \mu, x \rangle = \mathbb{E}(\langle Y, x \rangle)$ and $\langle \Sigma(x), y \rangle = \mathbb{E}(\langle x, Y - \mu \rangle \langle Y - \mu, y \rangle)$ for all $x, y \in \mathbb{H}$.

In the following, we assume Y is a random element of $\mathbb{L}^2(\mathcal{T}, \mathbb{C})$. Being self-adjoint and compact, its covariance operator Σ admits a representation $\Sigma(f) = \sum_{k \geq 1} \lambda_k \langle e_k, f \rangle e_k$ via countably many eigenfunctions $e_1, e_2, \dots \in \mathbb{L}^2(\mathcal{T}, \mathbb{C})$, $\Sigma(e_k) = \lambda_k e_k$, with real eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq 0$ of Σ (see Supplement). The $\{e_k\}_k$ form an orthonormal basis of the Hilbert subspace formed by the closure of the image of Σ . The random element can be represented as $Y = \mu + \sum_{k \geq 1} \langle e_k, Y - \mu \rangle e_k$ with probability one. The scores $Z_k = \langle e_k, Y - \mu \rangle$, $k \geq 1$, are complex random variables with mean zero and covariance $\text{Cov}(Z_k, Z_{k'}) = \mathbb{E}(\langle Y - \mu, e_k \rangle \langle e_{k'}, Y - \mu \rangle) = \lambda_k 1_{\{k=k'\}}(k)$, where $1_{\mathcal{S}}(t) = 1$ if $t \in \mathcal{S}$ and 0 else for a set \mathcal{S} .

Y is canonically identified with the bivariate real process $\mathbf{Y} = \kappa(Y) = (\Re(Y), \Im(Y))^\top$, random element in the Hilbert space $\mathbb{L}^2(\mathcal{T}, \mathbb{R}^2)$ with the inner product of $\mathbf{x} = \kappa(x), \mathbf{y} = \kappa(y)$, $x, y \in \mathbb{L}^2(\mathcal{T}, \mathbb{C})$, defined by $\langle \mathbf{x}, \mathbf{y} \rangle = \int \Re(x(t)) \Re(y(t)) d\nu(t) + \int \Im(x(t)) \Im(y(t)) d\nu(t) = \Re(\langle x, y \rangle)$.

Theorem 1. *Define the pseudo-covariance operator Ω of Y with mean μ by $\langle \Omega(x), y \rangle = \mathbb{E}(\langle Y - \mu, x \rangle \langle Y - \mu, y \rangle)$ for all $x, y \in \mathbb{L}^2(\mathcal{T}, \mathbb{C})$, and let Σ denote the covariance operator of $\mathbf{Y} = \kappa(Y)$. Then the covariance and pseudo-covariance operators Σ and Ω of Y together determine Σ via*

$$\kappa^{-1} \circ \Sigma \circ \kappa = (\Sigma + \Omega)/2.$$

Aiming at shape analysis, we are particularly interested in rotation-invariant distributions $\mathfrak{L}(\mathbf{Y})$ of $\mathbf{Y} = \kappa(Y)$, corresponding to $\mathfrak{L}(Y) = \mathfrak{L}(\exp(\mathbf{i}\omega)Y)$ for all $\omega \in \mathbb{R}$. In this case, $\mathfrak{L}(Y)$ is typically referred to as ‘proper’, ‘circular’ or ‘complex symmetric’ (Neeser and Massey, 1993; Picinbono, 1996; Kent, 1994) and the simplification by taking a complex approach becomes evident:

Theorem 2. *A random element Y in $\mathbb{L}^2(\mathcal{T}, \mathbb{C})$ with covariance operator Σ with eigenbasis $\{e_k\}_k$ and corresponding eigenvalues $\{\lambda_k\}_k$ follows a complex symmetric distribution if and only if all scores $Z_k = \langle e_k, Y - \mu \rangle$ with $\lambda_k > 0$ do, and additionally the mean of Y is $\mu = 0$. In this case,*

- i) the pseudo-covariance Ω of Y vanishes, i.e. $\Omega(y) = 0$ for all $y \in \mathbb{L}^2(\mathcal{T}, \mathbb{C})$, and the covariance operator Σ of the bivariate process $\mathbf{Y} = \kappa(Y)$ is completely determined by Σ ;*
- ii) the pairs $\mathbf{e}_k = \kappa(2^{-1/2}e_k), \mathbf{e}_{-k} = \kappa(\mathbf{i}2^{-1/2}e_k) \in \mathbb{L}^2(\mathcal{T}, \mathbb{R}^2)$ yield an eigendecomposition $\Sigma(\mathbf{f}) = \sum_{k \neq 0} \lambda_k \langle \mathbf{e}_k, \mathbf{f} \rangle \mathbf{e}_k$ of Σ . With probability one, $\mathbf{Y} = \sum_{k \neq 0} \mathbf{e}_k \mathbf{Z}_k$ with uncorrelated real scores \mathbf{Z}_k with mean zero, variance $\text{var}(\mathbf{Z}_k) = \lambda_k$ and $\kappa(Z_k) = (\mathbf{Z}_k, \mathbf{Z}_{-k})^\top$.*

While rotation invariance of $\mathcal{L}(\mathbf{Y})$ leads to even multiplicities in the eigenvalues of the bivariate covariance operator Σ , it does not pose a constraint on the complex eigenvalues and eigenfunctions of Σ , which would complicate the eigendecomposition. Here, rotation invariance of $\mathcal{L}(\mathbf{Y})$ instead translates to complex symmetry of the distribution of the scores Z_k .

Mean and covariance structure of Y can also be approached from the point-wise mean $\mu^*(t) = \mathbb{E}(Y(t))$ and Hermitian covariance surface $C(s, t) = \mathbb{E}((Y(s) - \mu^*(s))^\dagger(Y(t) - \mu^*(t))) = C(t, s)^\dagger$. Under complex symmetry, we obtain again $\mu^*(t) = 0$, while the auto-covariances $\mathbb{E}(\Re(Y(s))\Re(Y(t))) = \mathbb{E}(\Im(Y(s))\Im(Y(t))) = \Re(C(s, t))$ and cross-covariances $\mathbb{E}(\Re(Y(s))\Im(Y(t))) = -\mathbb{E}(\Im(Y(s))\Re(Y(t))) = \Im(C(s, t))$ of the bivariate \mathbf{Y} are completely determined by $C(s, t)$, as shown in the Supplement. The integral operator $\Sigma^*(f)(t) = \int C(s, t)f(s) d\nu(s)$ on $\mathbb{L}^2(\mathcal{T}, \mathbb{C})$ induced by the covariance surface again constitutes a compact and self-adjoint operator and admits, as such, an eigendecomposition. In fact, under standard assumptions such as continuity of $\mu^*(t)$ and $C(s, t)$, Fubini allows switching integrals and the point-wise mean $\mu^* = \mu$ coincides with the mean element and the operator $\Sigma^* = \Sigma$ with the covariance operator. In this case, the eigendecomposition of Σ also yields a decomposition

$$C(s, t) = \sum_{k \geq 1} \lambda_k e_k^\dagger(s) e_k(t)$$

of the covariance surface.

2.2 Hermitian covariance estimation via tensor-product smoothing

Based on a densely/regularly sampled collection of realizations $y_1, \dots, y_n : \mathcal{T} \rightarrow \mathbb{C}$ (with equal grids) of a complex symmetric process Y , the covariance surface $C(s, t)$ of Y can be estimated by the empirical covariance surface $\hat{C}_{emp.}(s, t) = \frac{1}{n} \sum_{i=1}^n y_i^\dagger(s) y_i(t)$ for each pair of grid-points s, t . This is, however, not possible in a sparse/irregular setting where only a limited number of evaluations $y_i(t_{i1}) = y_{i1}, \dots, y_i(t_{in_i}) = y_{in_i}$ are available for $i = 1, \dots, n$ such that, for a given (s, t) -tuple, $\hat{C}_{emp.}(s, t)$ would only be based on few observations if computable at all. Consequently, some kind of smoothing over samples becomes necessary and, following the seminal work of Yao et al. (2005), covariance estimation in the sparse/irregular functional case has widely been approached as a non-/semi-parametric regression problem. We proceed accordingly in the complex case and model $\mathbb{E}(Y^\dagger(s)Y(t)) = C(s, t)$ with a (smooth) regression estimator $\hat{C}(s, t)$ fitted to response products $y_{ij}^\dagger y_{ij}$ at respective tuples $(t_{ij}, t_{ij}) \in \mathcal{T}^2$, for $j, \bar{j} = 1, \dots, n_i$ and $i = 1, \dots, n$. Here, it is often reasonable to assume that, in fact, only measurements $\tilde{y}_{ij} = y_{ij} + \varepsilon_{ij}$ are observed with $\varepsilon_{ij} = \varepsilon_i(t_{ij})$ uncorrelated measurement errors originating from a white noise error process $\varepsilon(t)$, $t \in \mathcal{T}$. This leads to a combined covariance $\tilde{C}(s, t) = C(s, t) + \tau^2(t) 1_{\{s\}}(t)$ with $\tau^2(t) = \text{var}(\varepsilon(t))$ the variance function of $\varepsilon(t)$. Assuming $C(s, t)$ continuous, $\tau^2(t)$ can be distinguished as a discontinuous “nugget effect” at $s = t$.

Generalizing the approach of Cederbaum et al. (2018) for real covariance surfaces to the complex case, we propose to model $C(s, t)$ using a Hermitian tensor-product smooth

$$C(s, t) \approx \sum_{g=1}^m \sum_{k=1}^m \xi_{gk} f_g(s) f_k(t) = \mathbf{f}^\top(s) \Xi \mathbf{f}(t) = \text{vec}(\Xi)^\top (\mathbf{f}(t) \otimes \mathbf{f}(s))$$

with real-valued basis functions $f_k : \mathcal{T} \rightarrow \mathbb{R}$, $k = 1, \dots, m$, stacked to a vector $\mathbf{f}(t) = (f_1(t), \dots, f_m(t))^\top$, and a Hermitian coefficient matrix $\Xi = \{\xi_{kk'}\}_{kk'} = \Xi^\dagger \in \mathbb{C}^{m \times m}$ ensuring $C(s, t)$ is Hermitian as required, with vec stacking the columns of a matrix to a vector. Both the symmetry of the real part $\Re(\Xi) = \Re(\Xi)^\top$ and the anti-symmetry of the imaginary part $\Im(\Xi) = -\Im(\Xi)^\top$ present linear constraints. As such they can be implemented via suitable basis transforms $\mathbf{D}_{\Re}(\mathbf{f} \otimes \mathbf{f})(s, t)$ and $\mathbf{D}_{\Im}(\mathbf{f} \otimes \mathbf{f})(s, t)$ of the tensor-product basis $(\mathbf{f} \otimes \mathbf{f})(s, t) = (f_1(s)f_1^\top(t), \dots, f_m(s)f_m^\top(t))^\top$ with transformation matrices $\mathbf{D}_{\Re} \in \mathbb{R}^{(m^2+m)/2 \times m^2}$ and $\mathbf{D}_{\Im} \in \mathbb{R}^{(m^2-m)/2 \times m^2}$ for the symmetric and anti-symmetric part, respectively. Since $\mathbb{R}^{m \times m}$ is a direct sum of the vector spaces of symmetric and antisymmetric $m \times m$ matrices, \mathbf{D}_{\Im} can be obtained, e.g., as basis matrix of the null space of \mathbf{D}_{\Re} . A possible construction of \mathbf{D}_{\Re} is described by Cederbaum et al. (2018). In addition to the covariance, we also model the error variance $\tau^2(t) \approx \xi_\tau^\top \mathbf{f}_\tau(t)$ expanded in a real function basis $\mathbf{f}_\tau(t)$. Here, it might be convenient to employ the same basis $\mathbf{f}_\tau(t) = \mathbf{f}(t)$, or to assume constant error variance by setting $\mathbf{f}_\tau(t) = 1$ for all t . At any t with $\tau^2(t) = 0$, the measurement error is excluded from the model. The coefficients $\text{vec}(\hat{\Xi}) = \mathbf{D}_{\Re} \hat{\xi}_{\Re} + \mathbf{i} \mathbf{D}_{\Im} \hat{\xi}_{\Im}$ of the covariance estimator $\hat{C}(s, t)$ minimize the penalized least-squares criterion

$$\text{PLS}(\Xi, \xi_\tau) = \sum_{i, j, \bar{j}} \left| \mathbf{f}^\top(t_{ij}) \Xi \mathbf{f}(t_{ij}) + \xi_\tau^\top \mathbf{f}_\tau(t_{ij}) 1_{\{\bar{j}\}}(j) - y_{ij}^\dagger y_{i\bar{j}} \right|^2 + \text{PEN}(\Xi, \xi_\tau)$$

with quadratic penalty term PEN. They are separately obtained for the real and imaginary part of the covariance using $\text{PLS} = \text{PLS}_{\mathbb{R}} + \text{PLS}_{\mathbb{S}}$ via the well-known linear estimators $\hat{\xi}_{\mathbb{R}} \in \mathbb{R}^{(m^2+m)/2}$, $\hat{\xi}_{\mathbb{I}} \in \mathbb{R}^{m^2}$ minimizing $\text{PLS}_{\mathbb{R}} = \sum_{i,j,j'} (\xi_{\mathbb{R}}^{\top} \mathbf{D}_{\mathbb{R}} (\mathbf{f} \otimes \mathbf{f})(t_{ij}, t_{ij}) + \xi_{\mathbb{R}}^{\top} \mathbf{f}(t_{ij}) \mathbf{1}_{\{j\}}(j) - \Re(y_{ij}^{\dagger} y_{i'j}))^2 + \eta_{\mathbb{R}} \xi_{\mathbb{R}}^{\top} \mathbf{D}_{\mathbb{R}} \mathbf{P}_{\otimes} \mathbf{D}_{\mathbb{R}}^{\top} \xi_{\mathbb{R}} + \eta_{\mathbb{I}} \xi_{\mathbb{I}}^{\top} \mathbf{P}_{\tau} \xi_{\mathbb{I}}$, and $\hat{\xi}_{\mathbb{S}} \in \mathbb{R}^{(m^2-m)/2}$ minimizing $\text{PLS}_{\mathbb{S}} = \sum_{i,j,j'} (\xi_{\mathbb{S}}^{\top} \mathbf{D}_{\mathbb{S}} (\mathbf{f} \otimes \mathbf{f})(t_{ij}, t_{ij}) - \Im(y_{ij}^{\dagger} y_{i'j}))^2 + \eta_{\mathbb{S}} \xi_{\mathbb{S}}^{\top} \mathbf{D}_{\mathbb{S}} \mathbf{P}_{\otimes} \mathbf{D}_{\mathbb{S}}^{\top} \xi_{\mathbb{S}}$. Smoothing parameters $\eta_{\mathbb{R}}, \eta_{\mathbb{I}}, \eta_{\mathbb{S}} > 0$ control the penalty induced by the matrices \mathbf{P}_{τ} and $\mathbf{P}_{\otimes} = \mathbf{P} \otimes \mathbf{I}_m + \mathbf{I}_m \otimes \mathbf{P}$ constructed from a suitable penalty matrix $\mathbf{P} \in \mathbb{R}^{m \times m}$ for the basis coefficients of $\mathbf{f}(t)$ and the $m \times m$ identity matrix \mathbf{I}_m . Assuming the error variance not too heterogeneous over t , the matrix \mathbf{P}_{τ} should typically penalize deviations from the constant. Based on a working normality assumption, $\eta_{\mathbb{R}}, \eta_{\mathbb{I}}$ and $\eta_{\mathbb{S}}$ are obtained via restricted maximum likelihood (REML) estimation (Wood, 2017), avoiding computationally intense hyper-parameter tuning. For practical use, we extended the R package `sparseFLMM` (Cederbaum, 2018) to also offer anti-symmetric tensor-product smooths for the package `mgcv` (Wood, 2017) used for estimation. For asymptotic theory on the used penalized spline estimators, please see Wood et al. (2016).

After estimation, eigenfunctions e_k and eigenvalues λ_k of the covariance operator Σ of Y are estimated by the corresponding eigendecomposition $\hat{C}(s, t) = \sum_{k \geq 1} \hat{\lambda}_k \hat{e}_k^{\dagger}(s) \hat{e}_k(t)$ of the respective covariance operator $\hat{\Sigma}$. Based on $\hat{\Xi}$ and the Gram matrix $\mathbf{G} = \{\langle f_k, f_{k'} \rangle\}_{k,k'=1}^m$, the right eigenvalues of the matrix $\mathbf{G}^{-1} \hat{\Xi}$ yield the eigenvalues $\hat{\lambda}_k$ of $\hat{\Sigma}$. The corresponding eigenvectors $\hat{\theta}_k$ yield the eigenfunctions $\hat{e}_k(t) = \hat{\theta}_k^{\top} \mathbf{f}(t)$ of $\hat{\Sigma}$ for $k = 1, \dots, m$. To ensure positive-definiteness, eigenfunctions with $\lambda_k \leq 0$ are omitted from the basis. Nonnegativity of τ^2 is enforced post-hoc by setting negative values to zero.

3 Elastic full Procrustes analysis

3.1 Full Procrustes analysis in the square-root-velocity framework

To now propose (elastic) full Procrustes means for plane curves, we first introduce some underlying concepts and notation. We understand a *parameterized* curve as a function $\beta : [0, 1] \rightarrow \mathbb{C}$, which is assumed absolutely continuous such that the component-wise derivative $\dot{\beta}(t) = \frac{d}{dt} \Re \circ \beta(t) + \mathbf{i} \frac{d}{dt} \Im \circ \beta(t)$ exists almost everywhere and also the integral $\varphi_{\beta}(t) = \int_0^t |\dot{\beta}(s)| ds < \infty$ exists for $t \in [0, 1]$. Denoting the set of absolutely continuous functions $[0, 1] \rightarrow \mathbb{C}$ by $\mathcal{AC}([0, 1], \mathbb{C})$, we further assume $\beta \in \mathcal{AC}^*([0, 1], \mathbb{C}) = \mathcal{AC}([0, 1], \mathbb{C}) \setminus \{t \mapsto z : z \in \mathbb{C}\}$ excluding constant functions as degenerate curves. Then β has positive length $L(\beta) = \varphi_{\beta}(1) > 0$, and a constant-speed parameterization $\alpha = \beta \circ \varphi_{\beta}^{-1}$ always exists, when taking the generalized inverse $\varphi_{\beta}^{-1}(s) = \inf\{t \in [0, 1] : s L(\beta) \leq \varphi_{\beta}(t)\}$, $s \in [0, 1]$. Two parameterized curves $\beta_1, \beta_2 \in \mathcal{AC}^*([0, 1], \mathbb{C})$ are said to describe the same curve if they have the same constant-speed parameterization $\alpha_1 = \alpha_2$, which yields an equivalence relation $\beta_1 \approx \beta_2$. An *oriented* curve is then defined as equivalence class with respect to ‘ \approx ’. If the context allows it, we commonly refer to both oriented plane curves and their parameterized curve representatives β simply as ‘‘curve’’. A diffeomorphism $\gamma : [0, 1] \rightarrow [0, 1]$ which is orientation-preserving, i.e., with derivative $\dot{\gamma}(t) > 0$ for $t \in [0, 1]$, is called warping function and the set of such warping functions is denoted by Γ . With obviously $\beta \circ \gamma \approx \beta$, warping can equivalently be used to define equivalence of parameterized curves (see, e.g. Bruveris, 2016, which we also recommend for further details). Abstracting also from the particular coordinate system for \mathbb{C} , the shape of an (oriented) curve with parameterization β is then defined by $[\beta] = \{\tilde{\beta} \in \mathcal{AC}^*([0, 1], \mathbb{C}) : u \tilde{\beta} + v \approx \beta \text{ for some } u, v \in \mathbb{C}\}$, its equivalence class under translation by any v , rotation by $u/|u| = \exp(\mathbf{i}\omega)$, $\omega \in \mathbb{R}$, re-scaling by $|u|$, and warping. This presents our ultimate object of interest. In establishing a metric on the quotient space $\mathfrak{B} = \{[\beta] : \beta \in \mathcal{AC}^*([0, 1], \mathbb{C})\}$, we follow and extend the idea of the full Procrustes distance in landmark shape analysis and define

$$d_{\Psi}([\beta_1], [\beta_2]) = \inf_{\substack{a \geq 0, v_i \in \mathbb{C}, \\ \omega_i \in \mathbb{R}, \gamma_i \in \Gamma}} \|\Psi(\exp(\mathbf{i}\omega_1) \beta_1 \circ \gamma_1 + v_1) - a \Psi(\exp(\mathbf{i}\omega_2) \beta_2 \circ \gamma_2 + v_2)\| \quad (1)$$

for $\beta_1, \beta_2 \in \mathcal{AC}^*([0, 1], \mathbb{C})$, with a pre-shape map $\Psi : \mathcal{AC}^*([0, 1], \mathbb{C}) \rightarrow \mathbb{L}^2([0, 1], \mathbb{C})$, $\beta \mapsto q$ discussed below allowing to base computation on the \mathbb{L}^2 -metric while optimizing over all involved invariances. Acting differently than the other curve-shape preserving transformations (see, e.g., Srivastava and Klassen, 2016, Chap. 3.7), scale invariance is generally accounted for by a normalization constraint $\|\Psi(\beta)\| = \|q\| = 1$ for all β . Fixing $a = 1$ in (1) would yield a partial-Procrustes-type distance instead. Replacing also the norm by the arc length on the \mathbb{L}^2 -sphere would correspond to an intrinsic shape distance. To obtain a proper and sound metric, Ψ has to be carefully chosen. It is well-known that directly applying the \mathbb{L}^2 -metric on

the level of parameterized curves β is problematic, since in this case the warping action of $\gamma \in \Gamma$ is not by isometries (Srivastava and Klassen, 2016).

We set $\tilde{\Psi}(\beta)$ to the SRV-transformation (Srivastava et al., 2011), representing a curve β by its square-root-velocity (SRV) transform $q : [0, 1] \rightarrow \mathbb{C}$ given by $q(t) = \dot{\beta}(t)/|\dot{\beta}(t)|^{1/2}$ wherever this is defined and $q(t) = 0$ elsewhere. Indeed, q is square-integrable with $\|q\|^2 = \int_0^1 |q(t)|^2 dt = L(\beta)$. Since $\tilde{\Psi}(u\beta \circ \gamma + v)(t) = (u/|u|^{1/2}) q \circ \gamma(t) \dot{\gamma}(t)^{1/2}$, warping and rotation act by isometries with $\|\tilde{\Psi}(a \exp(i\omega) \beta_1 \circ \gamma + v) - \tilde{\Psi}(a \exp(i\omega) \beta_2 \circ \gamma + v)\| = a^{1/2} \|\tilde{\Psi}(\beta_1) - \tilde{\Psi}(\beta_2)\|$ for any two curves β_1, β_2 and $\gamma \in \Gamma, a \geq 0, \omega \in \mathbb{R}, u, v \in \mathbb{C}$. The \mathbb{L}^2 -metric on the SRV-transforms induces a metric on the space of parameterized curves modulo translation (Bruveris, 2016). It is commonly referred to as “elastic” metric due to the isometric action of γ allowing to construct a metric on oriented curves via optimal warping alignment. $\tilde{\Psi}$ is surjective but not injective, with $\tilde{\Psi}^{-1}(\{\tilde{\Psi}(\beta)\}) = \{\beta + v : v \in \mathbb{C}\} \subset [\beta]$. Without loss of generality, we can, thus, set $\tilde{\Psi}^{-1}(q)(t) = \int_0^t \dot{\beta}(s) ds = \int_0^t q(s) |q(s)| ds$ when discussing shapes $[\beta]$.

Proposition 1. *With $\Psi(\beta) = \tilde{\Psi}(\beta/L(\beta)) = \tilde{\Psi}(\beta)/\|\tilde{\Psi}(\beta)\|$ the normalized SRV-transform, d_Ψ defines a metric on \mathfrak{B} , referred to as elastic full Procrustes distance $d_\mathcal{E}$. It takes the form*

$$d_\mathcal{E}^2([\beta_1], [\beta_2]) = \inf_{u \in \mathbb{C}, \gamma \in \Gamma} \|q_1 - u q_2 \circ \gamma \dot{\gamma}^{1/2}\|^2 = 1 - \sup_{\gamma \in \Gamma} |\langle q_1, q_2 \circ \gamma \dot{\gamma}^{1/2} \rangle|^2$$

for $q_i = \Psi(\beta_i)$ unit-norm SRV-transforms of curve shape representatives $\beta_1, \beta_2 \in \mathcal{AC}^*([0, 1], \mathbb{C})$.

With a metric at hand, we may proceed by considering random shapes and define the concept of a Fréchet mean induced by the metric (compare, e.g., Huckemann, 2012; Ziezold, 1977). A random element A in a metric space (\mathfrak{A}, d) is a Borel-measurable random variable taking values in \mathfrak{A} . A (population) Fréchet mean or expected element $\mathfrak{m} \in \mathfrak{A}$ is defined as a minimizer of the expected square distance

$$\mathbb{E}(d^2(\mathfrak{m}, A)) = \sigma^2 = \inf_{\mathfrak{a} \in \mathfrak{A}} \mathbb{E}(d^2(\mathfrak{a}, A)).$$

assuming a finite variance $\sigma^2 < \infty$.

Definition 2. *A random (plane curve) shape $[B]$ is a random element in the shape space \mathfrak{B} equipped with the elastic full Procrustes distance $d_\mathcal{E}$. We call a Fréchet mean $[\mu_\mathcal{E}] \in \mathfrak{B}$ of $[B]$, represented by $\mu_\mathcal{E} \in \mathcal{AC}^*([0, 1], \mathbb{C})$, an elastic full Procrustes mean of the random shape $[B]$.*

As distance computation is carried out on SRV-transforms, it is, however, typically more convenient to consider the mean shape on SRV-level, i.e. via a distribution $\mathfrak{L}(Q)$ of a random element $Q = \Psi(B)$ in the Hilbert space $\mathbb{L}^2([0, 1], \mathbb{C})$ inducing the shape distribution $\mathfrak{L}([B])$.

Proposition 2. *Consider a random element Q in $\mathbb{L}^2([0, 1], \mathbb{C})$ with $\|Q\| = 1$ almost surely. The elastic full Procrustes means $[\mu_\mathcal{E}]$ of the induced random shape $[B] = [\Psi^{-1}(Q)]$ are determined by their SRV-transform $\psi_\mathcal{E} = \Psi(\mu_\mathcal{E})$ fulfilling*

$$\psi_\mathcal{E} \in \operatorname{argmax}_{y: \|y\|=1} \mathbb{E}(\sup_{\gamma \in \Gamma} |\langle y, Q \circ \gamma \dot{\gamma}^{1/2} \rangle|^2) = \operatorname{argmax}_{y: \|y\|=1} \mathbb{E}(\sup_{\gamma \in \Gamma} \langle y, Q \circ \gamma \dot{\gamma}^{1/2} \rangle \langle Q \circ \gamma \dot{\gamma}^{1/2}, y \rangle). \quad (2)$$

When fixing γ in Equation (2), the maximum of the quadratic form is obtained at the leading eigenvector of the covariance operator of $Q \circ \gamma \dot{\gamma}^{1/2}$, which is carried out in detail in Proposition 3 considering *inelastic* full Procrustes means of shapes of parameterized plane curves. This allows use of Hermitian covariance smoothing, introduced in Section 2.2, for shape mean estimation. Inelastic mean estimation will present a building block in elastic mean estimation but is also interesting in its own right, especially in data scenarios involving natural curve parameterizations.

Proposition 3. *For $\beta \in \mathcal{AC}^*([0, 1], \mathbb{C})$ define the shape of a parameterized plane curve as $(\beta) = \{u\beta + v : u, v \in \mathbb{C}\}$. Then*

- i) *the inelastic full Procrustes distance $d_\mathcal{E}((\beta_1), (\beta_2)) = \inf_{u \in \mathbb{C}} \|q_1 - u q_2\|$ with $\|q_i\| = 1$ for $\Psi(\beta_i) = q_i, i = 1, 2$, defines a metric on the shape space $\mathfrak{B} = \{(\beta) : \mathcal{AC}^*([0, 1], \mathbb{C})\}$ of parameterized plane curves and can be expressed as $d_\mathcal{E}^2((\beta_1), (\beta_2)) = 1 - |\langle q_1, q_2 \rangle|^2$;*
- ii) *multiplication by $\langle q_1, q_2 \rangle^\dagger / |\langle q_1, q_2 \rangle| = \operatorname{argmin}_{u: |u|=1} \|q_1 - u q_2\|$ yields rotation alignment of β_2 to β_1 ;*

iii) for a complex symmetric random element Q in $\mathbb{L}^2([0, 1], \mathbb{C})$ with covariance operator Σ , let $\mathcal{Y}_1 = \{y : \Sigma(y) = \lambda_1 y\}$ denote the spectrum of the leading eigenvalue λ_1 of Σ . Then, $(\mathcal{Y}_1) = \{(y) : y \in \mathcal{Y}_1\}$ is the set of Fréchet means of the random shape $(B) = (\Psi^{-1}(Q))$ in \mathfrak{B} with respect to $d_{\mathcal{G}}$, which we refer to as inelastic full Procrustes means. In particular, the leading eigenfunction $\psi_{\mathcal{G}} = e_1$ of an eigendecomposition of Σ yields an inelastic full Procrustes mean $(\mu_{\mathcal{G}})$ of (B) with SRV-transform $\psi_{\mathcal{G}} = \Psi(\mu_{\mathcal{G}})$. It is unique if λ_1 has multiplicity 1. The variance of (B) is $\sigma_{\mathcal{G}}^2 = \mathbb{E}(d_{\mathcal{G}}^2((\mu_{\mathcal{G}}), (B))) = 1 - \lambda_1$.

Motivated by Proposition 3 iii), we propose to estimate $\psi_{\mathcal{G}}$ as leading eigenfunction \hat{e}_1 of $\hat{C}(s, t)$ obtained by Hermitian covariance smoothing in Section 3.3, as part of the estimation procedure of $\psi_{\mathcal{E}}$. However, before that we first address the question of how it is still possible to work with derivative-based SRV-curves even in the sparsely observed setting so common in practice.

3.2 The square-root-velocity representation in a sparse/irregular setting

In practice, the shape of an (oriented) plane curve is observed via a vector $\mathbf{b} = (b_0, \dots, b_{n_0})^\top \in \mathbb{C}^{n_0+1}$ of points, which can be considered evaluations $\beta^*(t_j^*) = b_j$ of some continuous parameterization $\beta^* : [0, 1] \rightarrow \mathbb{C}$ of the curve at arbitrary time points $t_0^* < \dots < t_{n_0}^*$. However, fixing the time grid, the derivatives $\dot{\beta}^*(t_i^*)$ are not observable. Instead, evaluations of an SRV-transform describing the curve can be directly obtained from the finite differences $\Delta_j = b_j - b_{j-1}$, if the curve segments $\beta^*((t_{j-1}^*, t_j^*)) \subset \mathbb{C}$ between the observed points in \mathbf{b} have no edges or loops:

Theorem 3 (Feasible sampling). *If β^* is continuous and $\beta^* : (t_{j-1}^*, t_j^*) \rightarrow \mathbb{C}$ is injective and continuously differentiable with $\dot{\beta}^*(t) \neq 0$ for all $t \in (t_{j-1}^*, t_j^*)$, for $j = 1, \dots, n_0$, then for any time points $0 < t_1 < \dots < t_{n_0} < 1$ and speeds $w_1, \dots, w_{n_0} > 0$, there exists a $\gamma \in \Gamma$ such that*

$$q(t_j) = w_j^{1/2} (\beta^*(t_j^*) - \beta^*(t_{j-1}^*)) = w_j^{1/2} \Delta_j \quad (j = 1, \dots, n_0)$$

for the SRV-transform q of $\beta = \beta^* \circ \gamma$.

We call a vector of sampling points \mathbf{b} of a curve *feasible* if the conditions of Lemma 3 hold. This is always fulfilled if there is a $\beta^* \in (\beta)$ such that β^* is continuously differentiable with non-vanishing derivative on all $(0, 1)$ and, in particular, if it describes an embedded one-dimensional differentiable submanifold of \mathbb{R}^2 . If, instead, the curve has edges, they must be contained in \mathbf{b} , as well as a point inside of each loop (i.e. within each closed curve segment).

Note that while discrete observations often result in approximate derivative computations, Theorem 3 ensures that the derivative-based SRV-transform can be *exactly* recovered on a desired grid - up to a re-parameterization not essential in an analysis invariant to re-parameterization. Selected time points $t_1 < \dots < t_{n_0}$ and speeds $w_1, \dots, w_{n_0} > 0$ implicitly determine the parameterization. In principle, they could be arbitrarily selected due to parameterization invariance of the analysis, but with regard to mean estimation it is desirable to initialize them in a coherent way. Without any prior knowledge, constant speed parameterization of underlying curves β presents a canonical choice. To approximate this, we borrow from constant speed parameterization $\hat{\beta}$ of the sample polygon with vertices \mathbf{b} , implying a piece-wise constant SRV-transform $\hat{q}(t) = \sum_{j=1}^{n_0} q_j 1_{[s_{j-1}, s_j)}(t)$ of $\hat{\beta}$ with SRVs $q_j = \Delta_j |\Delta_j|^{-1} L^{1/2}(\hat{\beta})$, with $L(\hat{\beta}) = \sum_{j=1}^{n_0} |\Delta_j|$ the length of the polygon. The nodes $s_j = \sum_{l=1}^j |\Delta_l| / L(\hat{\beta})$ indicate the vertices $\hat{\beta}(s_j) = b_j$, $j = 0, \dots, n_0$. In accordance with that, we set $q(t_j) = q_j$ and select time points $t_j = (s_j + s_{j-1})/2$ in the center of the edges, for $j = 1, \dots, n_0$. Depending on the context other choices might be preferable, but we generally expect this choice to imply reasonable starting parameterizations.

3.3 Estimating elastic full Procrustes means via Hermitian covariance smoothing

Consider a collection of sample vectors $\mathbf{b}_i \in \mathbb{C}^{n_i+1}$ of n curves $\beta_i \in \mathcal{AC}^*([0, 1], \mathbb{C})$, $i = 1, \dots, n$, realizations of a random plane curve shape $[B]$. For scale-invariance, sample polygons are normalized to unit-length. Moreover, the \mathbf{b}_i are assumed feasibly sampled to represent them by evaluations $q_i(t_{ij}) = q_{ij}$ at time points t_{ij} , $j = 1, \dots, n_i$, of the SRV-transform q_i of β_i as described in the previous Section 3.2. We model an elastic full Procrustes mean $[\mu]$ of $[B]$ via the SRV-transform ψ of $\mu \in \mathcal{AC}^*([0, 1], \mathbb{C})$ expanded as $\psi(t) = \sum_{k=1}^m \theta_k f_k(t) = \boldsymbol{\theta}^\top \mathbf{f}(t) = (f_1(t), \dots, f_m(t))^\top$ of functions $f_k \in \mathbb{L}^2([0, 1], \mathbb{R})$, $k = 1, \dots, m$, with complex coefficient vector $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)^\top \in \mathbb{C}^m$. For the basis, piece-wise linear B-splines of order 1 present an attractive choice, since they have been proven identifiable under warping-invariance (Steyer et al., 2022) while still implying continuity of ψ and a differentiable mean curve μ .

The idea of alternating between a) mean estimation on aligned data and b) alignment of the data to the current mean is used for estimation of landmark full Procrustes means (Dryden and Mardia, 2016, p. 139) and intrinsic elastic mean curve shapes (Srivastava and Klassen, 2016, p. 319). We follow a similar strategy to find an estimator $\hat{\psi}(t) = \hat{\theta}^\top \mathbf{f}(t)$ for ψ but estimate an inelastic full Procrustes mean in a) and base the estimate on Hermitian covariance smoothing for irregularly/sparsely sampled curves. The covariance estimate is also used for estimating normalization and rotation alignment multipliers, which are not directly computable for sparse curve data. For warping alignment in b), we utilize the approach of Steyer et al. (2022), which has proven suitable also for irregularly/sparsely sampled curves. The single steps of the algorithm are detailed in the following and a discussion of its empirical performance is given in the next section.

Initialize in iteration $h = 0$ SRV-representations $q_i^{[h]}(t_{ij}^{[h]}) = q_{ij}^{[h]}$ with $q_{ij}^{[0]} = q_{ij}$ and $t_{ij}^{[0]} = t_{ij}$ as in Section 3.2 for all i, j , and repeat the following steps for $h = 1, 2, \dots$:

- I. **Covariance estimation:** We estimate the covariance surface $C^{[h]}(s, t)$ of a complex symmetric process Q underlying $q_1^{[h]}, \dots, q_n^{[h]}$ with a tensor-product estimator $\hat{C}^{[h]}(s, t) = \mathbf{f}(s)^\top \hat{\Xi}^{[h]} \mathbf{f}(t)$ with coefficient matrix $\hat{\Xi}^{[h]} \in \mathbb{C}^{m \times m}$. While for dense sampling, an estimate can be directly obtained from the covariance of the $\langle q_i^{[h]}, f_k \rangle$ (see Supplement), we propose Hermitian covariance smoothing as described in Section 2 for sparse/irregular data. This yields eigenfunctions $\hat{e}_k^{[h]}$ and eigenvalues $\hat{\lambda}_k^{[h]}$, $k = 1, \dots, m$, of the corresponding covariance operator $\hat{\Sigma}^{[h]}$, as well as an estimate $\hat{\tau}^{2[h]}(t) \geq 0$ of the variance of a white noise zero mean residual process $\varepsilon(t)$ at $t \in [0, 1]$, if measurement uncertainty on observations $Q(t_{ij}) + \varepsilon(t_{ij})$ is assumed.
- II. **Mean estimation:** Set $\hat{\psi}^{[h]}(t) = \hat{e}_1^{[h]}(t) = \hat{\theta}_1^{[h]\top} \mathbf{f}(t)$ to the leading eigenfunction of $\hat{\Sigma}^{[h]}$ obtained from the leading right eigenvector $\hat{\theta}_1^{[h]}$ of $\mathbf{G}^{-1} \hat{\Xi}^{[h]}$ with Gramian \mathbf{G} of \mathbf{f} . This yields an inelastic full Procrustes mean estimate $[\hat{\mu}^{[h]}] = [\Psi^{-1}(\hat{\psi}^{[h]})]$ of the curves with the current parameterization (Proposition 3), presenting the current estimate of the elastic full Procrustes mean.
- III. **Rotation alignment and re-normalization:** For $u_i^{[h]} = (z_{i1}^{[h]} / |z_{i1}^{[h]}|)^\dagger (L^{[h]}(\beta_i))^{-1/2}$ with $z_{i1}^{[h]} = \langle \hat{e}_1^{[h]}, q_i \rangle$, the multiplied $u_i^{[h]} q_i^{[h]}$ has norm 1 and is rotation aligned to $\hat{\psi}^{[h]}$. We estimate $u_i^{[h]}$ by $\hat{u}_i^{[h]}$ for $i = 1, \dots, n$ based on the covariance estimation by plugging in conditional expectations $\hat{z}_{i1}^{[h]} = \mathbb{E}(\langle \hat{e}_1^{[h]}, Q \rangle \mid Q(t_{ij}) + \varepsilon(t_{ij}) = q_{ij}^{[h]}, j = 1, \dots, n_i)$ and $\hat{L}^{[h]}(\beta_i) = \mathbb{E}(\|Q\|^2 \mid Q(t_{ij}) + \varepsilon(t_{ij}) = q_{ij}^{[h]}, j = 1, \dots, n_i)$ under a working normality assumption, an estimation approach in the spirit of Yao et al. (2005). Expressions can be found in the Supplement.
- IV. **Warping alignment:** Based on its rotation aligned SRV evaluations, the i th curve is (approximately) warping aligned to $\hat{\mu}^{[h]}$ using the approach of Steyer et al. (2022), where SRV-transforms are approximated as piece-wise constant functions $\hat{q}_i^{[h]}(t) \approx q_i^{[h]}(t)$ to find the infima of $\|\hat{\mu}^{[h]} - \hat{q}_i^{[h]} \circ \gamma_i \hat{\gamma}_i^{1/2}\|$ over $\gamma_1, \dots, \gamma_n \in \Gamma$. This yields new parameterization time-points $t_{ij}^{[h+1]}$, $j = 1, \dots, n_i$, and corresponding SRVs $q_{ij}^{[h+1]} = w_{ij}^{[h]} \hat{u}_i^{[h]} q_{ij}^{[h]}$, with $w_{ij}^{[h]} > 0$ depending on the $t_{ij}^{[h]}$ and $t_{ij}^{[h+1]}$, passed forward to proceed with the next iteration at Step I. Details can be found in the Supplement.

Stop the algorithm when $\|\hat{\psi}^{[h]} - \hat{\psi}^{[h-1]}\|$ is below a specified threshold in Step II. An additional execution of Steps III and IV then yields rotation aligned representations of approximately unit-length curves and current time points.

4 Adequacy and robustness of elastic full Procrustes mean estimation in realistic curve shape data

Familiar everyday shapes offer an ideal platform for evaluation of shape mean estimation, allowing for intuitive visual assessment of results. We consider three different such datasets for investigating the performance of elastic full Procrustes mean shape estimation and comparing it to other mean concepts: 1. `digit3.dat` from Dryden and Mardia (2016), in R package `shapes`, comprising a total of 30 handwritten digits “3” sampled at 13 landmarks each; 2. irregularly sampled spirals $\beta(t) = t \exp(13 \mathbf{i} t)$, $t \in [0, 1]$, with random $n_i \in \{17, \dots, 22\}$ sampling points per spiral or with $n_i \in \{4, \dots, 7\}$ in a very sparse setting, additionally provided with small measurement errors and random rotation, translation and scaling; and 3. handwritten letters “f” extracted from the `handwrit` data in Ramsay and Silverman (2005), in R package

`fd`, comprising 20 repetitions of the letter with a total of 501 samples per curve. While we focus on one letter here for simplicity, example fits on the entire “`fd`” writings can be found in Figure S1 in the Online Supplement.

Based on `digit3.dat`, we compare our elastic full Procrustes mean estimator $\hat{\mu}_E$ with its inelastic analog $\hat{\mu}_G$ and with an elastic curve mean estimator $\hat{\mu}_C$ taking shape invariances not into account (fitted with R package `elasdics`). Moreover, we investigate fitting performance of $\hat{\mu}_E$ for $n = 4, 10, 30$ observed digits in a simulation. All estimators are fitted using piece-wise constant and piece-wise linear B-splines with 13 equally spaced knots on SRV-level applying 2nd order difference penalties in the covariance estimation for $\hat{\mu}_E$ and $\hat{\mu}_G$. No penalty is available for $\hat{\mu}_C$. Figure 1 shows the estimates fitted on the first $n = 4$ digits in the dataset. Without warping alignment, $\hat{\mu}_G$ does not capture the pronounced central nose in the digit “3” as distinctly as $\hat{\mu}_E$. The difference is somewhat smaller when fitting on all $n = 30$ digits (not shown), yet only marginally. Since the data is roughly rotation and scaling aligned, $\hat{\mu}_C$ is very close to $\hat{\mu}_E$ when fitting on all digits. When fitting only on the first $n = 4$ digits in the data, however, $\hat{\mu}_C$ substantially deviates, in particular for the smooth estimator using linear splines, as shown in Figure 1 (top left). This can presumably be attributed to a) $\hat{\mu}_C$ being more affected by the one outlying “3” (top-left) than $\hat{\mu}_E$, and b) the nose pointing into different directions depending on the handwriting. Overall, deficiencies in warping and rotation alignment tend to mask features in the curve shapes by averaging over different orientations and parameterizations, similarly to the effect of measurement error in covariates in a regression model. With missing scale alignment, the shape of the estimated mean is mainly driven by the shape of the largest curve(s) in the data. Good estimation quality is also confirmed in simulations that compare elastic full Procrustes mean estimates $\hat{\mu}_l, l = 1, \dots, 101$, estimated on independently drawn bootstrap samples of the digits (with $n = 4, 10, 30$), with the mean μ estimated on the original dataset and taken as true mean. While single mean estimates for as few curves as $n = 4$ might considerably deviate, the majority visually resembles μ well, including $\hat{\mu}_{(0.75)}$ where $\hat{\mu}_{(a)}$ denotes the bootstrap estimator with $d_{(a)}$ the a -quantile of the distances $d_l = d_E([\hat{\mu}_l], [\mu]), l = 1, \dots, 101$. Except for two outliers, all estimates with $n = 10$ and $n = 30$ are better than $\hat{\mu}_{(0.75)}$ for $n = 4$ (Figure 1, top middle).

We illustrate the role of sparsity in shape mean estimation in the spiral data with its varying level of detail over the curve (i.e. varying curvature) and random irregular grids sampled roughly at constant angle distances (Figure 1, bottom). Elastic full Procrustes mean estimates are based on piece-wise linear splines on SRV-level with 20 knots and 2nd order penalties in covariance smoothing. With a moderate number of sample points $n_i \in \{17, \dots, 22\}$, the estimate based on $n = 9$ curves regains the original spiral shape close to perfectly. Only the inner end of the spiral with the most curvature shows some deviation. With $n_i \in \{4, \dots, 7\}$ and $n = 20$, the estimator does not capture the higher curvature in the inner part of the spiral but otherwise fits its shape well despite extreme sparsity. In sparse functional data analysis, borrowing of strength across curves allows for consistent estimation of principle components based on a minimum number of sampling points n_i for each curve under mild conditions (Yao et al., 2005). However, this cannot equally be expected under shape invariances, as indicated by the fact that no shape information remains when curves are observed at $n_i < 3$ points, and in particular when warping-alignment can only be approximated on sparse samples. Still, we observe that bias becomes vanishingly small when the sampling points cover the curve sufficiently well. As this is often the case in real data, elastic full Procrustes mean estimation performs reliably well in practice already for comparably sparse data in our experience.

Based on $n = 20$ handwritten letters “`f`”, we compare to R package `fdasrvf` (Tucker, 2017), which offers state-of-the-art elastic (intrinsic, not full Procrustes) shape mean estimation for regularly and densely observed curves. To test different degrees of sparsity, we consider three scenarios with $n_{points} = 10, 20, 30$ sampling points per curve. For each, we draw $l = 1, \dots, 101$ bootstrap samples with $n_1 = \dots = n_{20} = n_{points}$ points subsampled from the total recorded points of each “`f`” giving a higher acceptance probability to points important for curve reconstruction. This leads to datasets of sparse but still recognizable letters. For all three settings our elastic full Procrustes mean estimator is fitted using piece-wise constant B-splines with 30 equally spaced knots on SRV-level and applying a 2nd order difference penalty in the covariance estimation. This leads to polygonal means on curve level as in `fdasrvf` where the number of knots is, unlike in our approach, always equal to n_{points} . As they estimate a different, intrinsic shape mean based on the elastic geodesic shape distance ρ , a fair comparison is not possible. We thus tailor the comparison to favor `fdasrvf` by comparing (also our full Procrustes) to their intrinsic shape mean on the full data, and using their distance ρ . Figure 1 (top right) illustrates performance based on their “true mean” $[\mu]$, estimated on the complete original data. In the very sparse $n_{points} = 10$ setting, differences in the mean concept are clearly dominated by the gain of using our mean estimator, which shows stable estimates gradually improving with n_{points} . With more densely observed curves the differences in fitting performance become smaller and the `fdasrvf` implementation gains a distinct computational advantage due to quadratic increase of the design

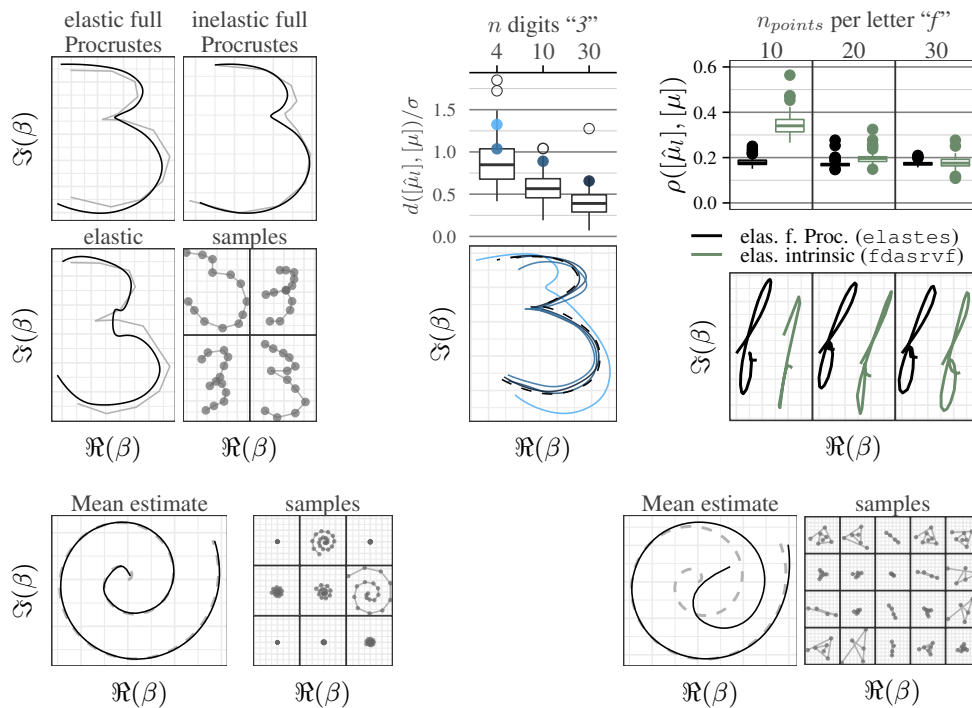


Figure 1: *Top left:* Different digit “3” mean curves (*black*: order 1, *grey*: order 0 B-splines on SRV-level) estimated on the first $n = 4$ sample polygons in `digit3.dat` shown in the bottom-right. *Top center:* Simulation results from 101-fold bootstrap samples of different sample sizes on `digit3.dat`. Four bootstrap estimates as examples of cases with relatively high deviations from μ (95% and for $n = 4$ also 75% distance quantiles) are depicted in the bottom and marked in the top panel (*filled dots*). Here, distances to $[\mu]$ are provided relative to the standard deviation σ estimated on the original dataset (as described below in Section 5). However, in some sense, σ is an underestimate as it does not include variation induced by irregular/sparse sampling. *Top right:* Performance comparison of our elastic full Procrustes mean and the `fdasrvf` elastic intrinsic mean estimator based on 101-fold bootstrap with $n_{points} = 10, 20, 30$ points sampled per letter “f”. Top shows the distribution of geodesic distances of estimated means to the overall intrinsic mean “f” $[\mu]$ (computed with `fdasrvf`). For `fdasrvf`, three outliers for $n_{points} = 10$ and one for $n_{points} = 20$ above 0.6 are omitted for the sake of visibility. Bottom shows example means of median geodesic distance in each setting. *Bottom:* Elastic full Procrustes means estimated on the spiral samples displayed to their right, in front of the original spiral (*grey, dashed line*).

matrix dimension in Hermitian covariance smoothing. While also in the $n_{points} = 30$ scenario fitting time remains below 1.5 minutes on a standard computer, it can dramatically increase with the numbers of knots and sampling points. In dense scenarios, we, thus, recommend utilizing an alternative covariance estimator for elastic full Procrustes mean estimation as described in the Online Supplement. Still, also in this denser setting, our approach estimating the elastic full Procrustes mean is at least as good in recovering the elastic intrinsic mean as `fdasrvf` which is, unlike our estimator, designed to estimate this mean.

5 Phonetic analysis of tongue shapes

The modulation of tongue shape presents an integral part of articulation (Hoole, 1999). Several authors investigate the shape variation in different phonetic tasks by analyzing tongue surface contours during speech production (Stone et al., 2001; Iskarous, 2005; Davidson, 2006) to obtain insights into speech mechanics. They model tongue contour shapes with (penalized) B-splines fitted through points marked on the tongue surface in ultrasound or MRT images of the speaker profile. While different measures to register/superimpose the tongue contour curves are undertaken, shape and warping invariances are not explicitly incorporated into their statistical analysis so far. In particular, reducing tongue shapes to one dimensional curves over an angle as in Davidson (2006) brings the problem that the different functions (due to different tongue shapes for different sounds) extend over different angle domains, which is ignored in the analysis. We suggest elastic full Procrustes analysis to appropriately handle the inherently two-dimensional curves. This approach accounts for the lack of a coordinate system in the ultrasound image, different positioning of ultrasound devices and size differences of speakers (Procrustes analysis) as well as flexibility of the tongue muscle to adjust its shape (elastic analysis). We illustrate the approach in experimental data kindly provided by Marianne Pouplier: tongue contour shapes are recorded in an experimental setting from six native German speakers ($\mathcal{S} = \{1, \dots, 6\}$) repeating the same set of fictitious words, such as “pada”, “pidi”, “pala” or “pili”. The words implement different combinations of two flanking vowels in $\mathcal{V} = \{a * a, i * i\}$ around a consonant in $\mathcal{C} = \{d, l, n, s\}$. Each combination is repeated multiple times by each of the speakers (1-8 times), observing tongue contour shapes formed at the central time point of consonant articulation (estimated from the acoustic signal). In total, this yields $n = 299$ sample polygons with nodes $\mathbf{b}_i \in \mathbb{C}^{n_i}$, $i = 1, \dots, n$, each sampled at $n_i = 29$ points from the tongue root to the tongue tip. A feature vector $X_i = (v_i, c_i, s_i)^\top \in \mathcal{X} = \mathcal{V} \times \mathcal{C} \times \mathcal{S}$ identifies the word-speaker combination of the i th curve. We investigate the different sources of shape variability (consonants, vowel context, speakers, repetitions) by elastic Full Procrustes analysis on different levels of hierarchy. Let $[\hat{\mu}_{\mathcal{A}}] \in \mathfrak{B}$ denote the elastic full Procrustes mean estimated for all i with $X_i \in \mathcal{A} \subset \mathcal{X}$. Figure 2 depicts the overall shape mean $[\hat{\mu}_{\mathcal{X}}]$, separate means $[\hat{\mu}_{\{(c,v)\} \times \mathcal{S}}]$ for the consonants $c \in \{d, s\}$ in both vowel contexts $v \in \mathcal{V}$, and speaker-word means $[\hat{\mu}_{\{(c,v,s)\}}]$ reflecting individual articulation by speaker $s \in \mathcal{S}$. Not displayed consonants “l” and “n” yield very similar shapes as “d”. Shape means are estimated using linear B-splines on SRV level with 13 equidistant knots and a 2nd order difference penalty for the basis coefficients. Homogeneous measurement error variance is assumed. Fitting the overall mean in this setting takes about 3 minutes on a standard computer.

For quantitative assessment of the hierarchical variation structure, we consider the conditional variances $\sigma_{\mathcal{A}}^2 = \mathbb{E}(d_{\mathcal{E}}^2([B], [\mu_{\mathcal{A}}]) \mid X \in \mathcal{A})$ with X constrained on a subset $\mathcal{A} \subset \mathcal{X}$. Motivated by $\sigma_{\mathcal{A}}^2 = 1 - \lambda_{\mathcal{A},1}$ (Proposition 3 iii) with $\lambda_{\mathcal{A},1}$ the largest eigenvalue of the respective conditional covariance operator, we estimate $\hat{\sigma}_{\mathcal{A}}^2 = 1 - \hat{\lambda}_{\mathcal{A},1}(\sum_{k=1}^m \hat{\lambda}_{\mathcal{A},k})^{-1}$ with $\hat{\lambda}_{\mathcal{A},1}, \dots, \hat{\lambda}_{\mathcal{A},m}$ the positive eigenvalues of the covariance operator obtained in the final iteration of estimating $[\mu_{\mathcal{A}}]$. In a dense setting, where observations can be exactly normalized, the estimator $\check{\sigma}_{\mathcal{A}}^2 = 1 - \hat{\lambda}_{\mathcal{A},1}$ can be used directly, since when $\|Q\| = 1$ almost surely also $\mathbb{E}(\|Q\|^2) = \sum_{k \geq 1} \lambda_k = 1$. In a sparse setting, however, dividing by $\sum_{k=1}^m \hat{\lambda}_{\mathcal{A},k}$ in $\hat{\sigma}_{\mathcal{A}}^2$ ensures non-negative variance estimates.

In analogy to standard analysis of variance, we define the coefficient of determination for \mathcal{A}_1 in some decomposition $\mathcal{A}_1 \times \mathcal{A}_2 = \mathcal{X}$ as $R_{\mathcal{A}_1}^2 = 1 - (|\mathcal{X}| \hat{\sigma}_{\mathcal{X}}^2)^{-1} |\mathcal{A}_2| \sum_{a \in \mathcal{A}_1} \hat{\sigma}_{\{a\} \times \mathcal{A}_2}^2$ reflecting the variance reduction achieved by conditioning on the features in \mathcal{A}_1 . Inspecting these measures underpins the visual impression from Figure 2: although the tongue movement is induced by consonant pronunciation, the vowel context appears more dispositive for the tongue shape during articulation explaining more than half of the total variation ($R_{\mathcal{V}}^2 = 0.68$, $R_{\mathcal{C}}^2 = 0.11$), which increases only to $R_{\mathcal{V} \times \mathcal{C}}^2 = 0.73$ when also distinguishing consonants. Comparing the different vowel contexts, we observe nearly double variation for $a * a$ than for $i * i$ with $\hat{\sigma}_{\{a*a\} \times \mathcal{C} \times \mathcal{S}}^2 / \hat{\sigma}_{\{i*i\} \times \mathcal{C} \times \mathcal{S}}^2 = 1.95$, which might potentially relate to different pronunciations of “a” in German dialects. When considering single word articulation of a speaker ($R_{\mathcal{V} \times \mathcal{C} \times \mathcal{S}}^2 = 0.93$) about 7 percent of the variation remain as residual variance, indicating that, while there is still non-negligible intra speaker variation, the inter speaker variance is considerably higher.

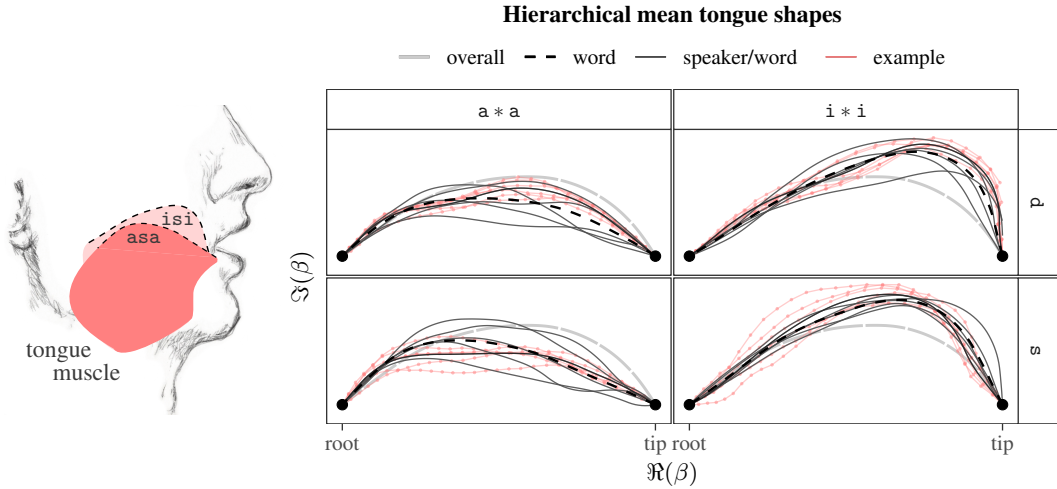


Figure 2: *Left*: schematic illustrating the tongue muscle modulation when pronouncing “isi” and “asa”. Dashed lines correspond to the respective mean shapes in the right plot. With its multiple and multi-directional fibers, the tongue muscle almost fills the entire oral cavity and can flexibly adjust its shape. In particular, not only tongue tip but also tongue root can move relatively freely. *Right*: elastic full Procrustes mean tongue shape estimates for different levels of aggregation. Tongue shapes are depicted in Bookstein coordinates, i.e. with the tongue roots at $\beta(0) = 0$ and the tongue tips at $\beta(1) = 1$. Each panel shows the overall mean shape in the dataset (*light gray, thick long-dashed line*), the vowel-consonant mean shape (*black, dashed line*), and speaker-wise mean shapes (*dark gray, solid lines*) for each combination. In each panel, original sample polygons (*light red, thin lines, dots at sample points*) are added for the speaker with most intra-speaker variation (which is the same speaker except for “idi”).

Recorded via ultrasound images, the shape of tongue surface contours modulo the respective invariances presents a natural object of analysis. Yet, if suitable reference landmarks allowed, the information on positioning, size, orientation and warping of the curve could also be separately investigated.

6 Discussion

While we find good performance of the proposed elastic full Procrustes mean estimator in realistic irregular/sparse curve data, future work should focus on theoretical assessment of estimation quality as well as inference. In particular, evaluation of the bias introduced by sub-optimal alignment of curves based on single discrete measurements is a topic of its own that would be of interest, as well as characterization of suitable sampling schemes where the bias is empirically negligible, which often appears to be the case in practice.

In this paper, we focus on open rather than closed curves, since the presented covariance-based estimation approach is particularly natural in this case. Constraining curves β to be closed, i.e. $\beta(0) = \beta(1)$, induces the non-linear constraint $\int_0^1 q(t)|q(t)| dt = 0$ on SRV-level, which prevents direct application of Proposition 3 in the estimation. To still obtain a closed mean estimator for closed observations, the presented estimator could be closed along the lines of Srivastava and Klassen (2016, Chapter 10.6.2) either post-hoc or in each step of the fitting algorithm.

As it can be analytically computed, inelastic full Procrustes analysis can also serve as a good starting point for estimating other types of shape means of plane curves. In addition, the estimated covariance structure supports estimation of inner products in sparse/irregular data scenarios, which are involved also in estimation of, e.g., other types of shape means. The presented results thus have relevance beyond the estimation of the (elastic) full Procrusted mean for plane shapes.

Acknowledgements

We sincerely thank Marianne Pouplier and Philip Hoole for providing their carefully recorded phonetic tongue shape data and Paula Giesler and Sophia Schaffer for their help in understanding and visualizing its anatomical background. We gratefully acknowledge funding by grant GR 3793/3-1 from the German research foundation (DFG).

References

- Bruveris, M. (2016). Optimal reparametrizations in the square root velocity framework. *SIAM Journal on Mathematical Analysis* **48**, 4335–4354.
- Cederbaum, J. (2018). *sparseFLMM: Functional Linear Mixed Models for Irregularly or Sparsely Sampled Data*. R package version 0.2-2.
- Cederbaum, J., Scheipl, F., and Greven, S. (2018). Fast symmetric additive covariance smoothing. *Computational Statistics & Data Analysis* **120**, 25–41.
- Davidson, L. (2006). Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance. *The Journal of the Acoustical Society of America* **120**, 407–415.
- Dryden, I. L., Le, H., Preston, S. P., and Wood, A. T. (2014). Mean shapes, projections and intrinsic limiting distributions. *Journal of Statistical Planning and Inference* **145**, 25–32.
- Dryden, I. L. and Mardia, K. V. (2016). *Statistical Shape Analysis: With Applications in R*. John Wiley & Sons.
- Greven, S. and Scheipl, F. (2017). A general framework for functional regression modelling (with discussion and rejoinder). *Statistical Modelling* **17**, 1–35 and 100–115.
- Happ, C. and Greven, S. (2018). Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association* **113**, 649–659.
- Hoole, P. (1999). On the lingual organization of the german vowel system. *The Journal of the Acoustical Society of America* **106**, 1020–1032.
- Hsing, T. and Eubank, R. (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*. John Wiley & Sons.
- Huckemann, S. F. (2012). On the meaning of mean shape: manifold stability, locus and the two sample test. *Annals of the Institute of Statistical Mathematics* **64**, 1227–1259.
- Iskarous, K. (2005). Patterns of tongue movement. *Journal of Phonetics* **33**, 363–381.
- Kent, J. T. (1994). The complex bingham distribution and shape analysis. *Journal of the Royal Statistical Society: Series B (Methodological)* **56**, 285–299.
- Neeser, F. D. and Massey, J. L. (1993). Proper complex random processes with applications to information theory. *IEEE transactions on information theory* **39**, 1293–1302.
- Picinbono, B. (1996). Second-order complex random vectors and normal distributions. *IEEE Transactions on Signal Processing* **44**, 2637–2640.
- Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer New York.
- Reiss, P. T. and Xu, M. (2020). Tensor product splines and functional principal components. *Journal of Statistical Planning and Inference* **208**, 1–12.
- Rynne, B. and Youngson, M. A. (2007). *Linear functional analysis*. Springer Science & Business Media.
- Srivastava, A., Klassen, E., Joshi, S. H., and Jermyn, I. H. (2011). Shape analysis of elastic curves in Euclidean spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33**, 1415–1428.
- Srivastava, A. and Klassen, E. P. (2016). *Functional and Shape Data Analysis*. Springer-Verlag.

- Steyer, L. (2021). *elasdics: Elastic Analysis of Sparse, Dense and Irregular Curves*. R package version 0.1.3.
- Steyer, L., Stöcker, A., and Greven, S. (2022). Elastic analysis of irregularly or sparsely sampled curves. *Biometrics* .
- Stone, M., Davis, E. P., Douglas, A. S., Aiver, M. N., Gullapalli, R., Levine, W. S., and Lundberg, A. J. (2001). Modeling tongue surface contours from cine-mri images.
- Tucker, J. D. (2017). *fdasrvf: Elastic Functional Data Analysis*. R package version 1.8.3.
- Wood, S. (2017). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2nd edition.
- Wood, S. N., Pya, N., and Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association* **111**, 1548–1563.
- Yao, F., Müller, H., and Wang, J. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* **100**, 577–590.
- Ziezold, H. (1977). 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.

A Hermitian covariance smoothing

A.1 Complex processes and rotation invariance

In the following, we detail prerequisites on linear operators and proof Theorem 1 and 2. Subsequently, Proposition 5 substantiates the relation of complex and real covariance surfaces indicated in the main manuscript.

We widely follow Hsing and Eubank (2015) in their introduction of functional data fundamentals, but re-state required statements underlying Section 2.1 for the complex case, since they nominally focus on real Hilbert spaces. Moreover, we give a Bochner integral free definition of mean elements and covariance operators to avoid introduction of additional notions.

Let \mathbb{H} denote a Hilbert space over \mathbb{C} or \mathbb{R} .

Theorem 4. *Let Ω be a compact self-adjoint operator on \mathbb{H} . Then there exists a sequence of countably many real eigenvalues $\lambda_1, \lambda_2, \dots \in \mathbb{R}$ of Ω with corresponding orthogonal eigenvectors $e_1, e_2, \dots \in \mathbb{H}$ and $\lambda_1 \geq \lambda_2 \geq \dots$ such that $\{e_k\}_k$ (called eigenbasis of Ω) is an orthonormal basis of the closure $\overline{\Omega(\mathbb{H})}$ of the image of Ω and for every $x \in \mathbb{H}$*

$$\Omega(x) = \sum_{k \geq 1} \lambda_k \langle e_k, x \rangle e_k.$$

Proof. Compare Rynne and Youngson (2007), Chapter 7.3. □

Definition 3. *Let Y be a random element in \mathbb{H} with $\mathbb{E}(\|Y\|^2) < \infty$. Then*

- i) *the mean element $\mu \in \mathbb{H}$ of Y is defined by $\langle f, \mu \rangle = \mathbb{E}(\langle f, Y \rangle)$ for all $f \in \mathbb{H}$.*
- ii) *the covariance operator $\Sigma : \mathbb{H} \rightarrow \mathbb{H}$ of Y is defined by $\langle \Sigma(e), f \rangle = \mathbb{E}(\langle Y - \mu, f \rangle \langle e, Y - \mu \rangle)$ for all $e, f \in \mathbb{H}$.*

Proposition 4. *Consider μ and Σ as above.*

- i) *μ and Σ are well-defined.*
- ii) *Σ is a nonnegative-definite (thus self-adjoint), trace-class and, hence, also compact linear operator.*

Proof. i) Since $\mathbb{E}(\|Y\|^2) < \infty$, Jensen's inequality yields $\mathbb{E}(\|Y\|) < \infty$, and therefore $\mathbb{E}(\langle f, Y \rangle) < \infty$ and also $\mathbb{E}(\langle Y - \mu, f \rangle \langle e, Y - \mu \rangle) < \infty$ for all $e, f \in \mathbb{H}$. Uniqueness of μ and Σ follows from the Riesz Representation Theorem.

- ii) Set $\mu = 0$ without loss of generality. Self-adjointness $\langle \Sigma(e), f \rangle = \mathbb{E}(\langle Y, f \rangle \langle e, Y \rangle) = \langle e, \Sigma(f) \rangle$ and nonnegative-definiteness $\langle \Sigma(e), e \rangle = \mathbb{E}(\langle Y, e \rangle \langle e, Y \rangle) = \mathbb{E}(|\langle e, Y \rangle|^2)$ immediately follow from the definition. Σ is trace-class, since for an orthonormal basis $\{e_k\}_k$ of \mathbb{H} it holds that

$$\sum_k \langle \Sigma(e_k), e_k \rangle = \sum_k \mathbb{E}(|\langle e_k, Y \rangle|^2) = \mathbb{E}(\|Y\|^2) < \infty$$

as assumed in the definition. Trace-class operators are compact. □

Corollary 1. *The covariance operator Σ of Y with $\mathbb{E}(\|Y\|^2) < \infty$ has an eigenbasis as described in Theorem 4.*

Proof. Immediately follows from Theorem 4 and the self-adjointness and compactness of Σ shown in Proposition 4. □

We proceed by proving Theorem 1 and 2 in the main manuscript characterizing the relation of the covariance of a complex process Y and the covariance of the corresponding bivariate real process \mathbf{Y} :

Theorem 1. For $x, y \in \mathbb{L}(\mathcal{T}, \mathbb{C})$ and assuming $\mu = 0$ without loss of generality, $\Re(\langle \Sigma(x) + \Omega(x), y \rangle) = \Re(\mathbb{E}(\langle x, Y \rangle \langle Y, y \rangle + \langle Y, x \rangle \langle Y, y \rangle)) = \Re(\mathbb{E}(2 \Re(\langle Y, x \rangle \langle Y, y \rangle))) = 2 \mathbb{E}(\Re(\langle Y, x \rangle) \Re(\langle Y, y \rangle)) = 2 \langle \Sigma(\kappa(x)), \kappa(y) \rangle.$ □

Theorem 2. From complex symmetry of $\mathfrak{L}(Y)$ it follows that $\mathfrak{L}(\exp(\mathbf{i}\omega)Z_k) = \mathfrak{L}(\langle e_k, \exp(\mathbf{i}\omega)Y \rangle) = \mathfrak{L}(Z_k)$, $\langle \mu, f \rangle = \mathbb{E}(\langle Y, f \rangle) = \mathbb{E}[\langle -Y, f \rangle] = 0$, and $\langle \Omega(e), f \rangle = \mathbb{E}(\langle Y, e \rangle \langle Y, f \rangle) = \mathbb{E}(-\langle Y, e \rangle \langle Y, f \rangle) = 0$ for all ω, k, e, f , which yields the first direction of the characterization via scores and, together with Theorem 1, statement i). ii) follows from Theorem 1, statement i) and the fact that if Z_k is complex symmetric, $\kappa(Z_k)$ has uncorrelated components with equal variance. Since $\exp(\mathbf{i}\omega)Y = \sum_{k \geq 1} \exp(\mathbf{i}\omega)Z_k e_k$ almost surely if $\mu = 0$, the second direction of the characterization via scores follows. \square

Proposition 5. Analogous to Σ , the bivariate covariance surface $\mathbf{C}(s, t)$ of $\mathbf{Y} = \kappa(Y)$ in $\mathbb{L}^2([0, 1], \mathbb{R}^2)$ is characterized by the matrix of covariance and cross-covariance surfaces

$$\begin{aligned} \mathbf{C}(s, t) &= \begin{pmatrix} \mathbb{E}(\Re(Y(s))\Re(Y(t))) & \mathbb{E}(\Im(Y(s))\Re(Y(t))) \\ \mathbb{E}(\Re(Y(s))\Im(Y(t))) & \mathbb{E}(\Im(Y(s))\Im(Y(t))) \end{pmatrix} \\ &= \frac{1}{2} \begin{pmatrix} \Re(C(s, t) + R(s, t)) & \Im(R(s, t) - C(s, t)) \\ \Im(C(s, t) + R(s, t)) & \Re(C(s, t) - R(s, t)) \end{pmatrix} \end{aligned}$$

determined by the pseudo-covariance surface $R(s, t) = \mathbb{E}(Y(s)Y(t))$ in addition to the complex covariance surface $C(s, t)$.

Proof.

$$\begin{aligned} C(s, t) + R(s, t) &= \mathbb{E}(Y^\dagger(s)Y(t) + Y(s)Y(t)) = \mathbb{E}((2\Re(Y(s)) + 0)Y(t)) \\ &= 2\mathbb{E}(\underbrace{\Re(Y(s))\Re(Y(t))}_{\frac{1}{2}\Re(C(s, t) + R(s, t))}) + 2\mathbf{i}\mathbb{E}(\underbrace{\Re(Y(s))\Im(Y(t))}_{\frac{1}{2}\Im(C(s, t) + R(s, t))}) \\ C(s, t) - R(s, t) &= \mathbb{E}(Y^\dagger(s)Y(t) - Y(s)Y(t)) = \mathbb{E}((0 - 2\mathbf{i}\Im(Y(s)))Y(t)) \\ &= -2\mathbf{i}\mathbb{E}(\underbrace{\Im(Y(s))\Re(Y(t))}_{\frac{1}{2}\Im(R(s, t) - C(s, t))}) + 2\mathbb{E}(\underbrace{\Im(Y(s))\Im(Y(t))}_{\frac{1}{2}\Re(C(s, t) - R(s, t))}) \end{aligned}$$

which shows the desired form. \square

B Elastic full Procrustes analysis

B.1 Full Procrustes analysis in the square-root-velocity framework

In the following, we start by proving Proposition 3 and use Proposition 3 i) to show Proposition 1 before proving Proposition 2 subsequently.

Proposition 3 i) and ii). $d_{\mathcal{G}}$ defines a metric on $\tilde{\mathfrak{B}}$:

$$\begin{aligned} d_{\mathcal{G}}^2((\beta_1), (\beta_2)) &= \inf_{u \in \mathbb{C}} \|q_1 - u q_2\|^2 = \inf_{u \in \mathbb{C}} \left[1 - \overbrace{u}^{=r_1 \exp(\mathbf{i}\omega_1)} \underbrace{\langle q_1, q_2 \rangle}_{=r_2 \exp(\mathbf{i}\omega_2)} - u^\dagger \langle q_2, q_1 \rangle + |u|^2 \right] \\ &= \inf_{r_1 > 0, \omega_1 \in \mathbb{R}} \left[1 - r_1 r_2 \exp(\mathbf{i}(\omega_1 + \omega_2)) - r_1 r_2 \exp(-\mathbf{i}(\omega_1 + \omega_2)) + r_1^2 \right] \\ &= \inf_{r_1 > 0, \omega_1 \in \mathbb{R}} \left[1 - 2r_1 r_2 \cos(\omega_1 + \omega_2) + r_1^2 \right] \stackrel{\omega_1 = -\omega_2}{=} \inf_{r_1 > 0} \left[1 - 2r_1 r_2 + r_1^2 \right] \quad (3) \\ &= \inf_{r_1 > 0} \left[1 - r_2^2 + (r_1 - r_2)^2 \right] \stackrel{r_1 = r_2}{=} 1 - |\langle q_1, q_2 \rangle|^2 = \|q_1 - \langle q_2, q_1 \rangle q_2\|^2 \quad (4) \end{aligned}$$

Clearly, $d_{\mathcal{G}}$ is well-defined (i.e., does not depend on the choice of $\beta_i \in (\beta_i)$), symmetric, positive. It is zero if and only if $|\langle q_2, q_1 \rangle| = 1$ and, hence, $(\beta_1) = (\int_0^t q_1(s)|q_1(s)| ds) = (\langle q_2, q_1 \rangle \int_0^t q_2(s)|q_2(s)| ds) = (\beta_2)$.

To show the triangle inequality let $(\beta_3) \in \tilde{\mathfrak{B}}$ with $q_3 = \Psi(\beta_3)$ and $v^* = \langle q_2, q_1 \rangle$. Then $d_{\mathcal{G}}((\beta_1), (\beta_3)) =$

$$\inf_{u \in \mathbb{C}} \|q_1 - u q_3\| \stackrel{\text{tr. ineq.}}{\leq} \underbrace{\|q_1 - v^* q_2\|}_{\stackrel{(4)}{=} \inf_{v \in \mathbb{C}} \|q_1 - v q_2\|} + \underbrace{\inf_{u \in \mathbb{C}} \|v^* q_2 - u q_3\|}_{= |v^*| \inf_{u \in \mathbb{C}} \|q_2 - u q_3\|} \stackrel{|v^*| \leq 1}{\leq} d_{\mathcal{G}}((\beta_1), (\beta_2)) + d_{\mathcal{G}}((\beta_2), (\beta_3)).$$

This shows i). ii) directly follows from (3), since $\exp(-\mathbf{i}\omega_2) = \langle q_1, q_2 \rangle / |\langle q_1, q_2 \rangle|$. \square

Proposition 3 iii). $\min_{(\beta) \in \mathfrak{B}} \mathbb{E} \left(d_{\mathcal{E}}^2((\beta), (B)) \right) = \min_{y: \|y\|=1} \mathbb{E} (1 - |\langle y, Q \rangle|^2) = 1 - \max_{y: \|y\|=1} \mathbb{E} (|\langle y, Q \rangle|^2)$.

Hence, $\psi_{\mathcal{E}} \in \operatorname{argmax}_{y: \|y\|=1} \mathbb{E} (|\langle y, Q \rangle|^2)$, and $\mathbb{E} (|\langle y, Q \rangle|^2) = \langle y, \Sigma(y) \rangle = \langle y, \sum_k \lambda_k \langle e_k, y \rangle e_k \rangle = \sum_k \lambda_k |\langle e_k, y \rangle|^2 \leq \lambda_1 \sum_k |\langle e_k, y \rangle|^2 = \lambda_1 \|y\|^2 = \lambda_1$, due to $\lambda_k \leq \lambda_1$ and $\|y\| = 1$, with equality attained by all $y = \frac{x}{\|x\|}$ with $x \in \mathcal{Y}_1$. This also yields $(\mu_{\mathcal{E}})$ and $\sigma_{\mathcal{E}}^2$. \square

Proposition 1. $d_{\mathcal{E}}$ defines a metric on \mathfrak{B} and allows for the provided expression:

$$\begin{aligned} d_{\mathcal{E}}^2([\beta_1], [\beta_2]) &= \inf_{a \geq 0, v_i \in \mathbb{C}, \omega_i \in \mathbb{R}, \gamma_i \in \Gamma, i=1,2} \left\| \exp(\mathbf{i}\omega_1) q_1 \circ \gamma_1 \dot{\gamma}_1^{1/2} - a \exp(\mathbf{i}\omega_2) q_2 \circ \gamma_2 \dot{\gamma}_2^{1/2} \right\|^2 \\ &\stackrel{(*)}{=} \inf_{u \in \mathbb{C}, \gamma \in \Gamma} \|q_1 - u q_2 \circ \gamma \dot{\gamma}^{1/2}\|^2 \stackrel{(**)}{=} 1 - \sup_{\gamma \in \Gamma} |\langle q_1, q_2 \circ \gamma \dot{\gamma}^{1/2} \rangle|^2 \end{aligned}$$

where $(*)$ follows from isometry of rotation and warping action setting $u = a \exp(\mathbf{i}(\omega_2 - \omega_1))$, $\gamma = \gamma_2 \circ \gamma_1^{-1}$; and $(**)$ is analogous to the proof of Proposition 3.

As Γ acts on \mathfrak{B} by isometries, $\inf_{u \in \mathbb{C}, \gamma \in \Gamma} \|q_1 - u q_2 \circ \gamma \dot{\gamma}^{1/2}\| = \inf_{\gamma \in \Gamma} d_{\mathcal{E}}([\beta_1], [\beta_2])$ is a semi-metric. To see that it is also positive-definite, assume $d_{\mathcal{E}}([\beta_1], [\beta_2]) = 0$. Consider any minimizing sequence $\{u_l\}_l$ with $0 = d_{\mathcal{E}}([\beta_1], [\beta_2]) = \inf_{\gamma \in \Gamma} \lim_{l \rightarrow \infty} \|q_1 - u_l q_2 \circ \gamma \dot{\gamma}^{1/2}\|$. Then, $\{u_l\}_l$ is bounded, since $|u_l| \|q_2\| = \inf_{\gamma \in \Gamma} \|u_l\| \|q_2 \circ \gamma \dot{\gamma}^{1/2}\| = \inf_{\gamma \in \Gamma} \|u_l q_2 \circ \gamma \dot{\gamma}^{1/2}\| \leq \inf_{\gamma \in \Gamma} \|u_l q_2 \circ \gamma \dot{\gamma}^{1/2} - q_1\| + \|q_1\| = \|q_1\|$ and $\|q_2\| > 0$ since β_1 is assumed non-constant. Hence, there is a convergent sub-sequence $\lim_{h \rightarrow \infty} u_{l_h} = u$, and $0 = \inf_{\gamma \in \Gamma} \lim_{h \rightarrow \infty} \|q_1 - u_{l_h} q_2 \circ \gamma \dot{\gamma}^{1/2}\| \stackrel{\text{continuity}}{=} \inf_{\gamma \in \Gamma} \|q_1 - u q_2 \circ \gamma \dot{\gamma}^{1/2}\|$ which is known to be a metric on $\mathbf{q}_1 = \kappa(q_1)$, $\mathbf{q}_2 = \kappa(q_2) \in \mathbb{L}^2([0, 1], \mathbb{R}^2)$ (Bruveris, 2016). Hence, also $[\beta_1] = [\beta_2]$ which completes the proof. \square

Proposition 2. In analogy to Proposition 3, $\min_{[\beta] \in \mathfrak{B}} \mathbb{E} (d_{\mathcal{E}}^2([\beta], [B])) = \min_{y: \|y\|=1} \mathbb{E} (1 - \sup_{\gamma \in \Gamma} |\langle y, Q \circ \gamma \dot{\gamma}^{1/2} \rangle|^2) = 1 - \max_{y: \|y\|=1} \mathbb{E} (\sup_{\gamma \in \Gamma} |\langle y, Q \circ \gamma \dot{\gamma}^{1/2} \rangle|^2)$. \square

B.2 The square-root-velocity representation in a sparse/irregular setting

Theorem 5. *Let $\beta : [0, 1] \rightarrow \mathbb{C}$ be continuous, injective, and, for all $t \in (0, 1)$, continuously differentiable with $\dot{\beta}(t) = \frac{d}{dt} \Re \circ \beta(t) + \mathbf{i} \frac{d}{dt} \Im \circ \beta(t) \neq 0$. Then, there exists a $c \in (0, 1)$ such that $\dot{\beta}(c) = \delta (\beta(1) - \beta(0))$ for some $\delta > 0$.*

Proof. Let $\rho = \Re \circ \beta$ and $\zeta = \Im \circ \beta$ denote the real and imaginary part of β . Without loss of generality assume $\beta(0) = 0$ and $\beta(1) = \mathbf{i}$. Choose $0 \leq t_0 < t_1 \leq 1$ with $\rho(t_0) = \rho(t_1) = 0$ such that $\zeta(t) \geq \zeta(t_0)$ for all $t \in [0, 1]$ with $\rho(t) = 0$ and $\zeta(t) \leq \zeta(t_1)$ for all $t \in [t_0, 1]$ with $\rho(t) = 0$. If $\rho(t) = 0$ for all $t \in [t_0, t_1]$ and, hence, $\dot{\beta}(t) = \mathbf{i} \dot{\zeta}(t)$ within (t_0, t_1) , the Mean Value Theorem directly yields existence of the desired $c \in (t_0, t_1)$. We may, thus, assume $\rho(t) \neq 0$ for some $t \in [t_0, t_1]$, say, with $\rho(t) > 0$. Accordingly, a maximizer $c \in [t_0, t_1]$ with $\rho(c) = \max_{t \in [t_0, t_1]} \rho(t) > 0$ lies in (t_0, t_1) and $\dot{\rho}(c) = 0$, since ρ is continuously differentiable. Hence $\dot{\beta}(c) = \mathbf{i} \dot{\zeta}(c) \neq 0$ as β is regular. $t_0 \neq t_1$ and c all exist due to compactness/continuity arguments.

We will now assume $\delta = \dot{\zeta}(c) < 0$ and show that this leads to a contradiction. With some upper/lower bounds $\rho_{\sup} > \rho(c) (> 0)$ and $\zeta_{\inf} < \min_{t \in [0, 1]} \zeta(t)$, we construct the open polygonal curve $\alpha : [c, 1]$ connecting the points $a_1 = \beta(c)$, $a_2 = \rho_{\sup} + \mathbf{i} \zeta_{\inf}$, $a_3 = \mathbf{i} \zeta_{\inf}$ and $a_4 = \beta(t_0) \leq 0$. Then $\beta 1_{[t_0, c]} + \alpha 1_{[c, 1]}$ is a simple closed continuous curve on $[t_0, 1]$, hence splits \mathbb{C} into two connected open components, the interior component $\mathcal{A} \subset \mathbb{C}$ which is bounded and the exterior component $\mathcal{U} = \mathbb{C} \setminus \bar{\mathcal{A}}$ (Jordan curve theorem) where $\bar{\mathcal{A}}$ denotes the closure of \mathcal{A} . The path $\phi : [0, \infty) \rightarrow \mathbb{C}$, $r \mapsto \beta(t_1) + r \mathbf{i}$ does not intersect the boundary $\beta([t_0, c]) \cup \alpha([c, 1]) = \bar{\mathcal{A}} \cap \bar{\mathcal{U}}$ for all $r \geq 0$, since, by construction, $\zeta(t_1) > \zeta(a_k)$ for $k = 2, \dots, 4$ and, for all $t \in [t_0, c]$ with $\rho(t) = 0$, $\zeta(t_1) > \zeta(t)$ as $\zeta(t_1) \geq \zeta(t)$, $c < t_1$ and β injective. Thus, ϕ lies entirely in \mathcal{A} or in \mathcal{U} . Since \mathcal{A} is bounded, the path and, in particular, $\phi(0) = \beta(t_1) \in \mathcal{U}$. Due to the construction of α and injectivity of β that do not permit intersection of the boundary (Jordan curve), $\beta(t)$ lies in \mathcal{A} for all $t > c$ if it lies within \mathcal{A} for some $t > c$. This makes the local behavior at c crucial. Thus, the assumption of $\dot{\zeta}(c) < 0$ entailing $\beta(t) \in \mathcal{A}$ for some $t > 0$ yields, in particular, $\beta(t_1) \in \mathcal{A}$ and, hence, the desired contradiction. \square

Corollary 2 (Feasible sampling). *If $\beta^* : [0, 1] \rightarrow \mathbb{C}$ is continuous and $\beta^* : (t_{j-1}^*, t_j^*) \rightarrow \mathbb{C}$ continuously differentiable for $j = 1, \dots, n_0$, $t_0^* < \dots < t_{n_0}^*$ with non-vanishing derivative, then for any time points $0 < t_1 < \dots < t_{n_0} < 1$ and speeds $w_1, \dots, w_{n_0} > 0$, there exists a $\gamma \in \Gamma$ such that for the SRV-transform q of $\beta = \beta^* \circ \gamma$, $q(t_j) = w_j^{1/2} (\beta^*(t_j^*) - \beta^*(t_{j-1}^*)) = w_j^{1/2} \Delta_j$ for all $j = 1, \dots, n_0$.*

Proof. Since this is a local property, it suffices to consider the case of $n_0 = 1$ and $t_0^* = 0, t_1^* = 1$. By Theorem 5, there exists $c \in (0, 1)$ with $\hat{\beta}(c)^* = a \Delta_1$ for some $a > 0$. Choose $\gamma \in \Gamma$ such that $\gamma(t_1) = c$ and $\dot{\gamma}(t_1) = w_1 a^{-2}$. Then, $q(t_j) = \beta^* \circ \gamma(t_j) \gamma'(t_j)^{1/2} = a \Delta_1 w_1^{1/2} a^{-1} = w_1^{1/2} \Delta_1$ for all $j = 1, \dots, n_0$. \square

B.3 Estimating elastic full Procrustes means via Hermitian covariance smoothing

In the following, we provide additional details for three steps in our proposed elastic full Procrustes mean estimation algorithm. We commence with proposing a more efficient covariance estimation procedure for data with densely observed curves and continue with a discussion of conditional complex Gaussian processes in Proposition 6 underlying our estimation of length and optimal rotation of curves. Finally, we detail the warping alignment strategy proposed for the re-parameterization step.

Covariance estimation for densely observed curves: If curves y_1, \dots, y_n , are sampled densely enough, covariance estimation can be achieved computationally more efficient than by Hermitian covariance smoothing. In fact, for say $n_i > 1000$ samples per curve and m basis functions $\mathbf{f} = (f_1, \dots, f_m)^\top$ for each margin, setting up the joint $(\sum_{i=1}^n n_i^2) \times (m^2 \pm m)/2$ design matrices for tensor-product covariance smoothing may also cause working memory shortage. Using the notation of Section 2.2, we obtain a tensor-product covariance estimator $\hat{C}(s, t) = \mathbf{f}^\top(s) \hat{\mathbf{\Xi}} \mathbf{f}(t)$ of the same form by setting $\hat{\mathbf{\Xi}} = \frac{1}{n} \sum_{i=1}^n \hat{\boldsymbol{\vartheta}}_i \hat{\boldsymbol{\vartheta}}_i^\dagger$ to the empirical covariance matrix of complex coefficient vectors $\hat{\boldsymbol{\vartheta}}_i = (\hat{\vartheta}_{i1}, \dots, \hat{\vartheta}_{im})^\top \in \mathbb{C}^m$ of basis representations $y_i(t) \approx \sum_{k=1}^m \hat{\vartheta}_{ik} f_k(t)$ for $i = 1, \dots, n$. Partitioning the data into $\mathcal{N}_1 \cup \dots \cup \mathcal{N}_N = \{1, \dots, n\}$ subsets for computational efficiency (which might simply be given by $\mathcal{N}_i = \{i\}$), the estimators $\hat{\boldsymbol{\vartheta}}_i$ are fit by minimizing the penalized least-squares criterion

$$\text{PLS}(\boldsymbol{\vartheta}_{i\Re}, \boldsymbol{\vartheta}_{i\Im}) = \sum_{i \in \mathcal{N}_i} \sum_{j=1}^{n_i} |y_{ij} - \boldsymbol{\vartheta}_{i\Re}^\top \mathbf{f}(t_{ij}) - \mathbf{i} \boldsymbol{\vartheta}_{i\Im}^\top \mathbf{f}(t_{ij})|^2 + \eta \boldsymbol{\vartheta}_{i\Re}^\top \mathbf{P} \boldsymbol{\vartheta}_{i\Re} + \eta \boldsymbol{\vartheta}_{i\Im}^\top \mathbf{P} \boldsymbol{\vartheta}_{i\Im}$$

with $\boldsymbol{\vartheta}_{i\Re} = \Re(\boldsymbol{\vartheta}_i)$ and $\boldsymbol{\vartheta}_{i\Im} = \Im(\boldsymbol{\vartheta}_i)$, for $l = 1, \dots, N$. In principle, real and imaginary parts can be separately fit with the same smoothing parameter $\eta \geq 0$ in both parts to achieve rotation invariant penalization. As in Section 2.2, we use the `mgcv` framework for fitting (Wood, 2017) using restricted maximum likelihood (REML) estimation for η . To speed up computation, η can be estimated only on \mathcal{N}_1 and fixed for $l = 2, \dots, N$, or set to $\eta = 0$ if no measurement error is assumed or no penalization is desired. The residual variance yields a constant estimate for τ^2 . Using for instance `mgcv`'s “`gauss`” family, a smooth estimator $\hat{\tau}^2(t)$ could be obtained as well but is not detailed here.

Rotation and length estimation: As proposed by Yao et al. (2005) for predicting scores in functional principal component analysis, we propose to use conditional expectations under a working normality assumption to incorporate the covariance structure of the data into estimation of inner products and quadratic terms. These are used for predicting basis coefficients of a curve (Proposition 6 iii) Equation (5)), its optimal rotation to the mean (6), its length (7), and its distance from the mean (8) or another given curve. We provide required conditional expectations covering both the case of a positive white noise error variance $\tau^2(t) > 0$ and of no white noise error ($\tau^2(t) = 0$) for each time point t . The distinction runs through all formulations and reading might be more convenient when assuming either of the cases is always fulfilled.

Proposition 6 (Conditional Gaussian process). *Consider a random element Y in a complex Hilbert space \mathbb{H} of functions $\mathcal{T} \rightarrow \mathbb{C}$ defined on some set \mathcal{T} . Assume $Y = \sum_{k=1}^m Z_k e_k$ finitely generated with probability one from a finite set $\mathbf{e}(t) = (e_1(t), \dots, e_m(t))^\top$ of functions $e_k \in \mathbb{H}$ with regular Gramian $\mathbf{G} = \{\langle e_k, e_{k'} \rangle\}_{k, k'} \in \mathbb{C}^{m \times m}$ and with $\mathbf{Z} = (Z_1, \dots, Z_m)^\top$ following a complex symmetric multivariate normal distribution with positive-definite covariance matrix $\mathbf{\Lambda}$. Let further denote ε an uncorrelated complex symmetric error process on \mathcal{T} with variance function $\tau^2 : \mathcal{T} \rightarrow \mathbb{R}$. We consider a sequence of $n_* = n_0 + n_+$ points $t_1, \dots, t_{n_*} \in \mathcal{T}$ and values $y_1, \dots, y_{n_*} \in \mathbb{C}$ with $\tau^2(t_1), \dots, \tau^2(t_{n_0}) = 0$ and $\tau^2(t_{n_0+1}), \dots, \tau^2(t_{n_0+n_*}) > 0$. Write $\mathbf{E} = \{e_k(t_j)\}_{jk} = (\mathbf{E}_0^\top, \mathbf{E}_+^\top)^\top$ for the $n_* \times m$ design matrix of function evaluations subdivided into $\mathbf{E}_0 \in \mathbb{C}^{n_0 \times m}$ and $\mathbf{E}_+ \in \mathbb{C}^{n_+ \times m}$ containing the evaluations with zero and positive error variance, respectively, and analogously $\mathbf{y} = (y_1, \dots, y_{n_*})^\top = (\mathbf{y}_0^\top, \mathbf{y}_+^\top)^\top$ for the values and $\mathbf{T}_+ = \text{Diag}(\tau^2(t_1), \dots, \tau^2(t_{n_+}))$ for the diagonal $n_+ \times n_+$ noise covariance matrix. Let $r_0 = \text{rank}(\mathbf{E}_0)$ denote the rank of \mathbf{E}_0 and $\mathbf{Q} = (\mathbf{M}, \mathbf{N})$ be an $m \times m$ Hermitian matrix such that \mathbf{M} is $m \times r_0$ and \mathbf{N} spans the null space of \mathbf{E}_0 . \mathbf{Q} is obtained, e.g., by the QR-decomposition $\mathbf{E}_0^\top = \mathbf{Q}\mathbf{R}$. By convention, matrices are set to 0 if their rank is zero (i.e., if $m - r_0, n_0$, or $n_+ = 0$, respectively). Conditioning on $Y(t_j) + \varepsilon(t_j) = \mathbf{Z}^\top \mathbf{e}(t_j) + \varepsilon(t_j) = y_j$ for $j = 1, \dots, n_*$ we obtain:*

i) $\mathbf{Z} = \mathbf{Z}_+ + \mathbf{z}_0$ is split into a random part $\mathbf{Z}_+ = \mathbf{N}\tilde{\mathbf{Z}}_+$ constrained to the linear sup-space $\text{span}(\mathbf{N})$ spanned by \mathbf{N} , with $\tilde{\mathbf{Z}}_+$ a complex random vector of length $m - r_0$, and a deterministic part $\mathbf{z}_0 = \mathbf{M} \left(\mathbf{M}^\dagger \mathbf{E}_0^\dagger \mathbf{E}_0 \mathbf{M} \right)^{-1} \mathbf{M}^\dagger \mathbf{E}_0^\dagger \mathbf{y}_0$. In fact, under the given assumptions $\mathbf{z}_0 = \mathbf{M}(\mathbf{M}\mathbf{E}_0)^{-\dagger} \mathbf{y}_0$ with probability one, but the generalized inverse is robust with respect to the case where $\mathbf{y}_0 \notin \text{span}(\mathbf{E}_0)$, i.e. where no measurement error is assumed but the curve cannot be exactly fit by the chosen basis.

ii) $\tilde{\mathbf{Z}}_+$ follows a complex normal with covariance $\mathbf{S} = \left(\mathbf{N}^\dagger \left(\mathbf{E}_+^\dagger \mathbf{T}_+^{-1} \mathbf{E}_+ + \mathbf{\Lambda}^{-1} \right) \mathbf{N} \right)^{-1}$, mean $\hat{\mathbf{z}}_+ = \mathbf{S} \mathbf{N}^\dagger \left(\mathbf{E}_+^\dagger \mathbf{T}_+^{-1} (\mathbf{y}_+ - \mathbf{E}_+ \mathbf{z}_0) - \mathbf{\Lambda}^{-1} \mathbf{z}_0 \right)$ and zero pseudo-covariance.

iii) For $x \in \mathbb{H}$ and $\mathbf{g}_x = (\langle e_1, x \rangle, \dots, \langle e_m, x \rangle)$, this provides conditional means

$$\hat{\mathbf{z}} = \mathbb{E}(\mathbf{Z} | Y(t_j) + \varepsilon(t_j) = y_j, j = 1, \dots, n_*) = \mathbf{N}\hat{\mathbf{z}}_+ + \mathbf{z}_0 \quad (5)$$

$$\mathbb{E}(\langle Y, x \rangle | Y(t_j) + \varepsilon(t_j) = y_j, j = 1, \dots, n_*) = \hat{\mathbf{z}}^\dagger \mathbf{g}_x \quad (6)$$

$$\mathbb{E}(\|Y\|^2 | Y(t_j) + \varepsilon(t_j) = y_j, j = 1, \dots, n_*) = \text{tr}(\mathbf{S} \mathbf{G}) + \hat{\mathbf{z}}^\dagger \mathbf{G} \hat{\mathbf{z}}. \quad (7)$$

$$\mathbb{E}(|\langle Y, x \rangle|^2 | Y(t_j) + \varepsilon(t_j) = y_j, j = 1, \dots, n_*) = \mathbf{g}_x^\dagger \mathbf{S} \mathbf{g}_x + \mathbf{g}_x \hat{\mathbf{z}}^\dagger \hat{\mathbf{z}} \mathbf{g}_x^\dagger. \quad (8)$$

Proof. The computation is analogous to the real case. Defining $\mathbf{Y} = (Y(t_1), \dots, Y(t_{n_*}))^\top$, i.e. $\mathbf{Y} = \mathbf{E}\mathbf{Z}$, and $\boldsymbol{\epsilon} = (\varepsilon(t_1), \dots, \varepsilon(t_{n_*}))^\top$, the distribution of $\tilde{\mathbf{Z}} = \mathbf{Q}^\dagger \mathbf{Z} = (\mathbf{M}^\dagger \mathbf{Z}, \mathbf{N}^\dagger \mathbf{Z})^\dagger = (\tilde{\mathbf{Z}}_0^\dagger, \tilde{\mathbf{Z}}_+^\dagger)^\dagger$ conditional on $\mathbf{Y} + \boldsymbol{\epsilon} = \mathbf{y}$ has a density proportional to

$$\begin{aligned} p_{\tilde{\mathbf{Z}}}(\tilde{\mathbf{z}} | \mathbf{Y} + \boldsymbol{\epsilon} = \mathbf{y}) &\propto p_{\tilde{\mathbf{Z}}, \mathbf{Y} + \boldsymbol{\epsilon}}(\tilde{\mathbf{z}}, \mathbf{Y} + \boldsymbol{\epsilon}) \propto p_{\tilde{\mathbf{Z}}, \boldsymbol{\epsilon}}(\underbrace{\tilde{\mathbf{Q}} \tilde{\mathbf{z}}}_{=\mathbf{M}\tilde{\mathbf{z}}_0 + \mathbf{N}\tilde{\mathbf{z}}_+}, \mathbf{y} - \mathbf{E}\mathbf{Q}\tilde{\mathbf{z}}) \\ &\propto \exp\left(-\frac{1}{2} \tilde{\mathbf{z}}^\dagger \mathbf{Q}^\dagger \mathbf{\Lambda}^{-1} \mathbf{Q} \tilde{\mathbf{z}}\right) \\ &\quad \cdot \exp\left(-\frac{1}{2} (\mathbf{y}_+ - \mathbf{E}_+ \mathbf{Q} \tilde{\mathbf{z}})^\dagger \mathbf{T}_+^{-1} (\mathbf{y}_+ - \mathbf{E}_+ \mathbf{Q} \tilde{\mathbf{z}})\right) 1_{\{\mathbf{y}_0\}}(\mathbf{E}_0 \mathbf{Q} \tilde{\mathbf{z}}) \\ &\stackrel{(*)}{\propto} \exp\left(-\frac{1}{2} \left(\tilde{\mathbf{z}}_+^\dagger \mathbf{N}^\dagger \mathbf{\Lambda}^{-1} \mathbf{N} \tilde{\mathbf{z}}_+ \right) - \Re\left(\tilde{\mathbf{z}}_+^\dagger \mathbf{N}^\dagger \mathbf{\Lambda}^{-1} \mathbf{z}_0 \right)\right) \\ &\quad \cdot \exp\left(-\frac{1}{2} \tilde{\mathbf{z}}_+^\dagger \mathbf{N}^\dagger \mathbf{E}_+^\dagger \mathbf{T}_+^{-1} \mathbf{E}_+ \mathbf{N} \tilde{\mathbf{z}}_+ + \Re\left(\tilde{\mathbf{z}}_+^\dagger \mathbf{N}^\dagger \mathbf{E}_+^\dagger \mathbf{T}_+^{-1} (\mathbf{y}_+ - \mathbf{E}_+ \mathbf{z}_0) \right)\right) 1_{\{\mathbf{M}^\dagger \mathbf{z}_0\}}(\tilde{\mathbf{z}}_0) \\ &\propto \exp\left(-\frac{1}{2} \tilde{\mathbf{z}}_+^\dagger \underbrace{\mathbf{N}^\dagger \left(\mathbf{\Lambda}^{-1} + \mathbf{E}_+^\dagger \mathbf{T}_+^{-1} \mathbf{E}_+ \right) \mathbf{N}}_{=\mathbf{S}^{-1}} \tilde{\mathbf{z}}_+ \right. \\ &\quad \left. + \Re\left(\tilde{\mathbf{z}}_+^\dagger \mathbf{N}^\dagger \underbrace{\left(\mathbf{E}_+^\dagger \mathbf{T}_+^{-1} (\mathbf{y}_+ - \mathbf{E}_+ \mathbf{z}_0) - \mathbf{\Lambda}^{-1} \mathbf{z}_0 \right)}_{=\mathbf{S}^{-1} \hat{\mathbf{z}}_+} \right)\right) 1_{\{\mathbf{M}^\dagger \mathbf{z}_0\}}(\tilde{\mathbf{z}}_0) \\ &\propto \exp\left(-\frac{1}{2} (\tilde{\mathbf{z}}_+ - \hat{\mathbf{z}}_+)^\dagger \mathbf{S}^{-1} (\tilde{\mathbf{z}}_+ - \hat{\mathbf{z}}_+)\right) 1_{\{\mathbf{M}^\dagger \mathbf{z}_0\}}(\tilde{\mathbf{z}}_0). \end{aligned}$$

Solving $\mathbf{y}_0 = \mathbf{E}_0 \mathbf{Q} \tilde{\mathbf{z}} = \mathbf{E}_0 \mathbf{M} \tilde{\mathbf{z}}_0$ for $\tilde{\mathbf{z}}_0$ yields $(*)$ and shows i). Deriving the kernel of a Gaussian, the remainder of the computation shows ii). In iii), (5) and (6) follow directly by linearity and (7) from variance decomposition (omitting conditions for brevity):

$$\begin{aligned} \mathbb{E}(\|Y\|^2) &= \mathbb{E}\left(\left\langle \sum_{k=1}^m Z_k e_k, \sum_{k=1}^m Z_k e_k \right\rangle\right) = \mathbb{E}(\mathbf{Z}^\dagger \mathbf{G} \mathbf{Z}) = \mathbb{E}(\text{tr}(\mathbf{Z} \mathbf{Z}^\dagger \mathbf{G})) \\ &\stackrel{\text{linearity}}{=} \text{tr}(\mathbb{E}(\mathbf{Z} \mathbf{Z}^\dagger) \mathbf{G}) = \text{tr}\left(\left(\text{Var}(\mathbf{Z}) + \mathbb{E}(\mathbf{Z}) \mathbb{E}(\mathbf{Z})^\dagger\right) \mathbf{G}\right) \stackrel{\text{ii)}}{=} \text{tr}(\mathbf{S} \mathbf{G}) + \hat{\mathbf{z}}^\dagger \mathbf{G} \hat{\mathbf{z}}. \end{aligned}$$

The computation for (8) is analogous. □

Warping alignment: Generally, we consider it advisable to base warping alignment of the i th curve directly on its original SRV-evaluations $q_{i1}^{[h]}, \dots, q_{in_i}^{[h]}$ but, when considerable measurement error presents an issue, it might also be useful to employ a smoothed reconstruction $\tilde{q}_i : [0, 1] \rightarrow \mathbb{C}$ of the SRV-transform

in the assumed basis. Based on the working normality assumption used also for length and rotation estimation, such a reconstruction is obtained as $\hat{q}_i^{[h]}(t) = (\hat{\mathbf{z}}_i^{[h]} / \|\hat{\mathbf{z}}_i^{[h]}\|)^\top \hat{\mathbf{e}}^{[h]}(t)$ with $\hat{\mathbf{z}}_i^{[h]} = (\hat{z}_{i1}^{[h]}, \dots, \hat{z}_{im}^{[h]})^\top$ the predicted score vector for the eigenbasis $\hat{\mathbf{e}}^{[h]} = (\hat{e}_1^{[h]}, \dots, \hat{e}_m^{[h]})^\top$.

Following Steyer et al. (2022), warping alignment to $\hat{\mu}^{[h]}$ is conducted using another, polygonal approximation of the curve given by a piece-wise constant approximation $\hat{q}_i^{[h]} \in \mathbb{L}^2([0, 1], \mathbb{C})$ of $q_i^{[h]}$. With a hyperparameter $\rho \in [0, 1]$, we control the balance between original $q_{ij}^{[h]}$ (for $\rho = 0$) and smoothed reconstruction \tilde{q}_i (for $\rho = 1$) and set $\hat{q}_{ij}^{[h]} = \hat{u}_i^{[h]} \left(\rho \tilde{q}_i^{[h]}(t_{ij}^{[h]}) + (1 - \rho) q_{ij}^{[h]} \right)$ at nodes $s_{i0}^{[h]} = 0$, $s_{ij}^{[h]} = 2t_{ij}^{[h]} - s_{ij-1}^{[h]}$, $j = 1, \dots, n_i$. This defines $\hat{q}_i^{[h]}(t) = \sum_{j=1}^{n_i} \hat{q}_{ij}^{[h]} 1_{[s_{ij-1}^{[h]}, s_{ij}^{[h]}]}(t)$ already rotated by $\hat{u}_i^{[h]}$.

Warping alignment to $\hat{\mu}^{[h]}$ is achieved for $i = 1, \dots, n$ by finding an optimal $\hat{q}_i^* \in \mathbb{L}^2([0, 1], \mathbb{C})$ with

$$\|\hat{q}_i^* - \hat{\psi}^{[h]}\| \leq \|\hat{q}_i^{[h]} \circ \gamma^{\hat{\gamma}^{1/2}} - \hat{\psi}^{[h]}\| \quad \text{for all } \gamma \in \Gamma \quad (9)$$

where the polygon approximation yields a practically feasible optimization problem and has proven suitable for sparse/irregular curves (Steyer et al., 2022). As shown by Steyer et al. (2022), the optimizers of (9) have the form $\hat{q}_i^*(t) = \sum_{j=1}^{n_i} w_i(t) \hat{q}_{ij}^{[h]} 1_{[s_{ij-1}^{[h+1]}, s_{ij}^{[h+1]}]}(t)$ almost-everywhere, where, denoting $a_+ = \max\{a, 0\}$

for $a \in \mathbb{R}$, the functions $w_i : [0, 1] \rightarrow \mathbb{R}$ are given by $w_i^2(t) = (s_{ij}^{[h]} - s_{ij-1}^{[h]}) \Re \left(\psi^{[h]}(t)^\dagger \hat{q}_{ij}^{[h]} \right)_+^2 / \int_{s_{ij-1}^{[h+1]}^{[h+1]}}^{s_{ij}^{[h+1]}^{[h+1]}} \Re \left(\psi^{[h]}(t)^\dagger \hat{q}_{ij}^{[h]} \right)_+^2 dt$ for $t \in [s_{ij-1}^{[h+1]}^{[h+1]}, s_{ij}^{[h+1]}^{[h+1]})$, and fully determined by the warped time points

$$(s_{i1}^{[h+1]}, \dots, s_{in_i-1}^{[h+1]}) = \arg \max_{0=s_{i0} \leq \dots \leq s_{in_i}=1} \sum_{j=1}^{n_i} \left((s_{ij}^{[h]} - s_{ij-1}^{[h]}) \int_{s_{ij-1}^{[h+1]}^{[h+1]}}^{s_{ij}^{[h+1]}^{[h+1]}} \Re \left(\psi^{[h]}(t)^\dagger \hat{q}_{ij}^{[h]} \right)_+^2 dt \right)^{1/2}.$$

If $s_{ij}^{[h+1]} = s_{ij-1}^{[h+1]}$ for some j , there is a minimizing sequence of functions of the form given for \hat{q}_i^* . After optimization over the $s_{ij}^{[h]}$ with R package `elasdics` (Steyer, 2021), we set new $t_{ij}^{[h+1]} = (s_{ij-1}^{[h+1]} + s_{ij}^{[h+1]})/2$ and $q_{ij}^{[h+1]} = w_j^* q_{ij}^{[h]}$ with $w_j^* = (s_{ij}^{[h]} - s_{ij-1}^{[h]})^{1/2} (s_{ij}^{[h+1]} - s_{ij-1}^{[h+1]})^{-1/2}$ for $s_{ij}^{[h+1]} > s_{ij-1}^{[h+1]}$ and omit double time points for $j = 1, \dots, n_i$. The chosen time-points hereby approximate $t_{ij}^{[h+1]} \approx t_{ij}^* \in (s_{ij-1}^{[h+1]}, s_{ij}^{[h+1]})$ with $w_i(t_{ij}^*) = w_{ij}^*$ existing by the Mean Value Theorem.

C Adequacy and robustness of elastic full Procrustes mean estimation in realistic curve shape data

While we focus on the first letter “*f*” in our simulation studies, Figure 3 exemplifies elastic full Procrustes mean estimation on the entire “*fda*” handwritings contained in the dataset `handwrit.dat` in the R package `fda` (Ramsay and Silverman, 2005). To visualize different degrees of sparsity, means are fitted after subsampling recorded points to $n_i = n_{points}$, $i = 1, \dots, n$, $n = 20$, random sampling points for each curve placing higher acceptance probability on points more important for curve reconstruction, as illustrated in the bottom of the figure. Means are fitted using piece-wise constant 0 order B-splines with 70 knots applying a 2nd order difference penalty in the Hermitian covariance estimation. This results in a nice gradual evolution from a rough “*fda*” approximation for $n_{points} = 21$ to a detailed handwritten “*fda*” for $n_{points} = 71$.

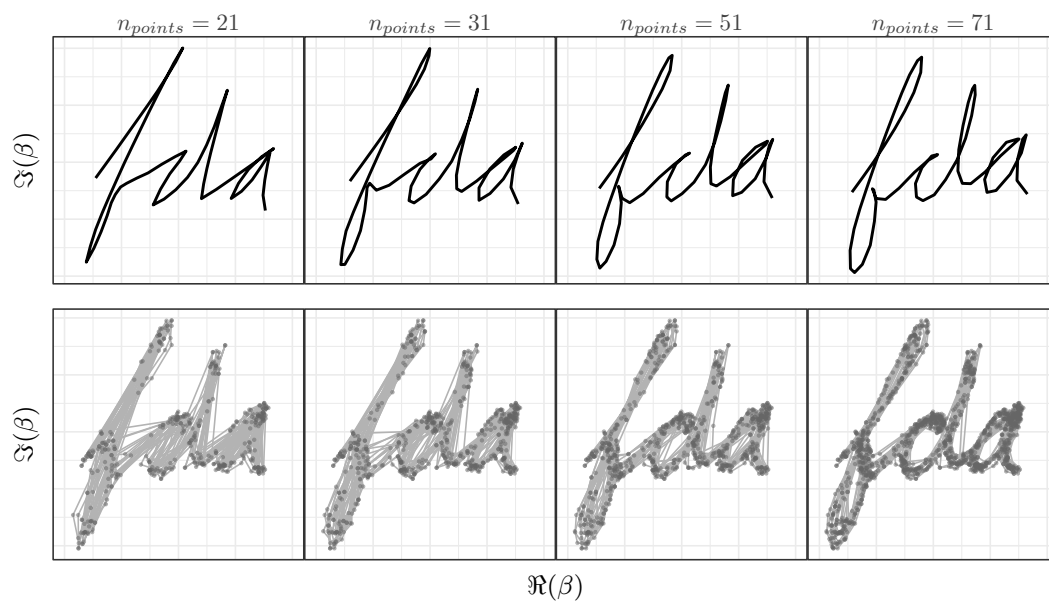


Figure 3: *Top:* Elastic full Procrustes means estimated over 20 handwritten “fda”s sampled with different degrees of sparsity. *Bottom:* Underlying datasets with 20 curves from the `handwrit.dat` dataset subsampled with higher acceptance probability on points important for curve reconstruction. Points sampled for each curve are connected by light-grey lines.

4. Paper III: Regression in Quotient Metric Spaces with a Focus On Elastic Curves

Paper III extends the unconditional (Fréchet) mean estimation for curves modulo re-parametrization (Subsection 1.3.2) considered in Paper I to regression in this metric space. It also proposes “quotient regression” as a generalization of “linear” regression to quotient metric spaces arising from actions by isometries (e.g. rotation) in a broader context. In this way, this paper also contributes to the few available options for regression in metric spaces (Subsection 1.2.2). The proposed regression method is applied to outlines of the human hippocampus from magnetic resonance imaging (MRI), where the shape of the irregularly sampled hippocampus is modeled using age, Alzheimer’s disease, and sex as covariates. All methods are made readily available in the R-package “elasdics”.

Contributing article:

Steyer, L., Stöcker, A. and Greven, S. (2023). Regression in quotient metric spaces with a focus on elastic curves. *arXiv pre-print*, arXiv:2305.02075

Declaration on personal contributions:

The main parts of this research were carried out by the author, including the implementation in the R package “elasdics”. Sonja Greven and Almond Stöcker were involved in this project with important advice and discussions. Almond Stöcker also assisted in the conceptualization of the research objectives.

REGRESSION IN QUOTIENT METRIC SPACES WITH A FOCUS ON ELASTIC CURVES

A PREPRINT

Lisa Steyer, Almond Stöcker & Sonja Greven
for the Alzheimer's Disease Neuroimaging Initiative*

16th June 2023

ABSTRACT

We propose regression models for curve-valued responses in two or more dimensions, where only the image but not the parametrization of the curves is of interest. Examples of such data are handwritten letters, movement paths or outlines of objects. In the square-root-velocity framework, a parametrization invariant distance for curves is obtained as the quotient space metric with respect to the action of re-parametrization, which is by isometries. With this special case in mind, we discuss the generalization of 'linear' regression to quotient metric spaces more generally, before illustrating the usefulness of our approach for curves modulo re-parametrization. We address the issue of sparsely or irregularly sampled curves by using splines for modeling smooth conditional mean curves. We test this model in simulations and apply it to human hippocampal outlines, obtained from Magnetic Resonance Imaging scans. Here we model how the shape of the irregularly sampled hippocampus is related to age, Alzheimer's disease and sex.

Keywords alignment, elastic distance, quotient space regression, sparse functional data, square-root-velocity framework, warping

*Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

1 Introduction: Regression for metric spaces

Regression is a widely used statistical technique for exploring the relationship between covariates and response variables. In the simplest case of linear regression, these variables are elements in the Euclidean space and the relationship between the variables is assumed to be affine linear. Since linear operations are also defined in general Hilbert spaces, the linear regression model can be extended to these spaces (Ramsay and Dalzell, 1991) and in particular to functional data in \mathbb{L}_2 (Ramsay and Silverman, 2005). For more general (metric) response spaces, analogues of linear models are less straightforwardly to define.

The focus of this paper is to develop an 'elastic' regression model for curves modulo parametrization. More precisely, we consider the quotient space \mathcal{A}/Γ as response space, where \mathcal{A} is the set of absolutely continuous curves $\mathbf{y} : [0, 1] \rightarrow \mathbb{R}^d$, $d \in \mathbb{N}$, and Γ is the set of boundary-preserving diffeomorphisms $\gamma : [0, 1] \rightarrow [0, 1]$. These curves occur naturally when we look at the outlines of (e.g., anatomical) objects such as the corpus callosum (Joshi et al., 2013) where only the image but not the parametrization of the curves is of interest. Furthermore, handwritten letters or symbols (e.g. Dryden and Mardia, 2016b), protein structures (Srivastava et al., 2010) or centerlines of the internal carotid artery (Sangalli et al., 2009) can be viewed as curves modulo parametrization in 2d or 3d. In this work, we investigate the variability in outlines of (a representative slice of) the hippocampus of patients suffering from Alzheimer's disease and of a control group, with the aim of differentiating changes due to Alzheimer's from normal aging (Fig. 3). These outlines were extracted from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (Petersen et al., 2010).

In this elastic setting, the response space is a quotient space that only has a metric space structure with no notion of linearity, such that linear models cannot be directly defined. It seems natural to use constant speed geodesics instead of affine linear functions in metric spaces, since they coincide in the case of vector spaces. Fletcher (2013) considers such a geodesic regression model for a scalar covariate and the response variable being an element of a smooth Riemannian manifold. Here, the tangent bundle of the manifold serves as a convenient parametrization of the set of possible geodesics. Conversely, in general metric spaces, there is no such parameterization of the geodesics. Hence, although a geodesic regression model could be defined here as well, the estimation of the minimizing geodesics is difficult to accomplish and the result difficult to interpret. For this reason, to our knowledge, the geodesic model has not been considered for general metric spaces.

Petersen and Müller (2019) develop a non-geodesic global regression model for responses that are elements of general metric spaces and Euclidean covariates. The regression function here is implicitly defined for each possible combination of covariate values as a (potentially negatively) weighted Fréchet mean. This means that no global model parameters are estimated, which makes interpretation difficult. Overall, defining a regression model in general metric spaces that is both interpretable and computable appears difficult if not infeasible. To build such a model for response data in a metric space, it thus seems necessary to make use of the specific structure of the response space.

Besides our current structure of interest \mathcal{A}/Γ , there are various situations where observations can be naturally seen as elements of a quotient space, for instance if the objects of interest are either subject to certain invariances or not fully observed. Classic examples arise in statistical shape data analysis (Dryden and Mardia, 2016a), where objects are considered invariant under translation, rotation, and scaling, as well as occasionally also under reflection, or a subset of these invariances. Other examples include analysis of unlabeled networks (Calissano et al., 2023), data on a Grassmannian (Hong et al., 2016), data of 3D rotations (Fletcher, 2013), compositional data (Pawlowsky-Glahn et al., 2015) and density data (van den Boogaart et al., 2014). Srivastava and Klassen (2016) also combines parametrization invariance with statistical shape analysis – to analyze shapes of curves and also surfaces – and, more recently, also with analysis of unlabeled graphs to model brain arterial networks (Guo et al., 2022).

Given the relevance and variety of data in quotient spaces in the literature, we will motivate our elastic regression model for curves in \mathcal{A}/Γ with a more general discussion of a regression approach for responses in certain quotient metric spaces. More precisely, we will consider quotient spaces where the distance is induced by an isometric group action,

since this is the case for \mathcal{A}/Γ if we equip \mathcal{A} with a semi-metric based on the Fisher-Rao metric (Srivastava et al., 2010). This semi-metric can be simplified to the \mathbb{L}_2 distance using the square-root-velocity (SRV) transformation (essentially $\mathbf{y} \mapsto \frac{\dot{\mathbf{y}}(t)}{\sqrt{\|\dot{\mathbf{y}}(t)\|}}$, $\mathbf{y} \in \mathcal{A}$) and minimization over all possible re-parametrizations in Γ yields a suitable “elastic” distance on \mathcal{A}/Γ modulo translation. While not all of the above examples correspond to such isometric group actions, they comprise – besides re-parameterization groups – also rotation, reflection and permutation groups.

The considered approach, which we refer to as “quotient regression”, is straight-forward and natural in two ways: a) the structure of the model predictor is simply obtained by projecting a suitable predictor in the original space to the quotient, and b) the model is fit based on the distance in the quotient metric space obtained by minimizing the distance in the original space over all possible group actions. Due to the more general perspective, beyond the target “quotient linear regression” for elastic curves, our results on consistency and existence of estimators, as well as inclusion of geodesics in the model space, are also applicable to other quotient regression scenarios. It also allows us to point out close connections to approaches for other response quotient metric spaces, such as the recent approach of Calissano et al. (2022) for unlabeled network responses, corresponding to quotient linear regression over the permutation group, and intrinsic Riemannian regression for responses in shape spaces (Cornea et al., 2017), combining rotation invariance with invariance with respect to non-isometric re-scaling. Curves in \mathcal{A}/Γ , in particular, have not been directly considered before as responses in an elastic regression model. One existing approach (Tucker et al., 2019) examines the case of elastic curves as covariates instead. They introduce elastic functional principal component regression (fPCR) for scalar response variables and 1d-functions as covariates. Here they first align the data curves to their Fréchet mean and then perform principal component analysis (PCA) for both the aligned curves as well as the optimal re-parametrizations and use both parts in a functional regression model. (Guo et al., 2020) proceed similarly but use the principal component scores of the pre-aligned SRV curves as covariates and response in a regression model.

Given this related work, we consider regression (on SRV or on curve level) after pre-alignment natural benchmarks to our model. Specifically, we compare our quotient linear model for curves to 1) linear regression after pre-alignment, a simpler approach that can be used for regression in the quotient of any Hilbert space, 2) to linear regression on curve instead of SRV level basing only alignment on the SRV framework, 3) to the combination of the simplifications in 1) and 2), and 4) to Fréchet regression (Petersen and Müller, 2019) for general metric spaces, which we adapt and implement for this purpose for the case of \mathcal{A}/Γ . In simulation studies, we illustrate when a clear performance gain by our model can be expected and when alternatives yield comparably good results.

In applications, such as in our example of hippocampus outlines, it is often necessary to handle sparsely or irregularly observed curves. We achieve this via employing spline bases, as often done in (sparse) functional data analysis. This is motivated by the work of Steyer et al. (2022) on spline-based unconditional elastic mean estimation, where we show identifiability of spline coefficients modulo parameterization and the adequateness of the approach for sparsely or irregularly observed curves. We provide a ready-to-use implementation of our elastic regression model in the R package `elasdics`. In a simulation study, we validate bootstrap confidence regions either based on spline coefficients - specific to our spline-based modeling - or more generically on distances, and discuss when each is recommended in practice. Both approaches enable data based model selection and assessment of estimation uncertainty. The proposed inference methods allow us to reveal and assess systematic patterns in the hippocampus outlines, which are visually hard to distinguish due to considerable subject-to-subject variation (Fig. 3). Specifically, we are able to compare the effect of Alzheimer’s disease to that of normal aging – two mechanisms that have been related to each other in the literature before (Henneman et al., 2009) – in a more detailed and visually intuitive way (Fig. 4).

We proceed as follows. In Section 2, we first construct the model for responses in general quotient metric spaces before developing the elastic model and our estimation strategy in the particular case of curves modulo parameterization. Here we build on the spline modeling and alignment methods for sparsely and irregularly sampled curves developed in Steyer et al. (2022). In Section 3, we present different alternatives to our model, which have not yet been discussed in this form in the literature, either, but deemed natural competitors by us. Section 4 proposes inference methods for our model. Section 5 compares the performance of our model with the alternative methods described in Section 3 and

validates inference based on the spline coefficients. Finally, in Section 6, we use our method to model the outline of the human hippocampus as a function of age, Alzheimer’s disease status and sex, before concluding in Section 7.

2 Quotient space regression and the particular case of elastic curves

Regression models for elastic curves are a particular case of regression models for quotient metric spaces, where the quotient is induced by an isometry, and we will define a regression model for the quotient by using the structure of the original space. In the case of elastic curves, the reparametrization group acts by isometries on \mathbb{L}_2 , the space of SRV-transformed curves (cf. Srivastava and Klassen, 2016). This means the original space here is \mathbb{L}_2 , which is a Hilbert space and therefore has a linear structure and allows us to base our models on linear regression in \mathbb{L}_2 . With this goal in mind, it is worthwhile to begin with a more general discussion of regression models in metric spaces. In particular, we discuss reasonable model spaces \mathcal{F} for regression in quotient spaces \mathcal{Y}/G over a more general original space \mathcal{Y} on which the group G acts by isometries. This is of independent interest and shows direct connections to regression for unlabeled networks (Calissano et al., 2022) and on shape/form spaces (Cornea et al., 2017; Stöcker et al., 2023).

As stated by Petersen and Müller (2019) in very general terms, traditional regression for the mean is naturally generalized to metric spaces by modeling the conditional Fréchet mean given covariates. This generalizes the least squares problem via replacing the Euclidean metric in the risk minimization with the distance in the metric space. More precisely, for \mathcal{X} being the space of covariates, (\mathcal{Y}, d) a metric space, and (X, Y) random variables taking values in $\mathcal{X} \times \mathcal{Y}$, the conditional Fréchet mean of Y given X is given by

$$\mathcal{E}(Y|X = x) = \operatorname{argmin}_{\mu \in \mathcal{Y}} \mathbb{E}(d(Y, \mu)^2 | X = x). \quad (1)$$

Petersen and Müller (2019) point out that without assuming an algebraic structure on \mathcal{Y} , it is not feasible to directly define a parametric regression model, that is to define a suitable function space \mathcal{F} such that $x \mapsto \mathcal{E}(Y|X = x)$ is an element of \mathcal{F} . For this reason, they develop a generalization of multiple linear regression as a set of weighted Fréchet means, where the weights are given by a known function of the covariates. This allows them to define a regression model in general metric spaces without an explicit model equation or global model parameters. In contrast, as soon as there is any additional structure given on \mathcal{Y} , it can potentially be used to motivate a suitable function space \mathcal{F} , which we refer to as model space in the following.

Definition 2.1 (Model-based conditional Fréchet mean). *Given a model space \mathcal{F} , we define the model-based conditional Fréchet mean as*

$$f^* = \operatorname{argmin}_{f \in \mathcal{F}} \mathbb{E}(\mathbb{E}(d(Y, f(X))^2 | X)) = \operatorname{argmin}_{f \in \mathcal{F}} \mathbb{E}(d(Y, f(X))^2), \quad (2)$$

assuming the total variation $\mathbb{E}(d(Y, f(X))^2) < \infty$ is finite for some $f \in \mathcal{F}$.

Note that, in contrast to Equation (1), the minimization is here over \mathcal{F} rather than point-wise over \mathcal{Y} and that, in general, there does not exist a unique minimizer but $f^* \subseteq \mathcal{F}$ is a set of models. $f^*(x) = \mathcal{E}(Y | X = x)$ coincide, if the model is correctly specified. In practice, this is of course hard to verify and it might be more truthful to model f^* and assume that it reasonably approximates $\mathcal{E}(Y | X = x)$. Since the distinction is subtle, we nonetheless simply refer to $f^*(x) = \mathcal{E}(Y | X = x)$ as conditional Fréchet mean in the following, when \mathcal{F} is clear from the context, while considering f^* as given in Definition 2.1.

For a corresponding estimator $\hat{f} \subseteq \mathcal{F}$ of f^* the following properties are desirable: a) good interpretability, presenting one central advantage of using the structure of \mathcal{Y} , b) consistency and c) computational feasibility, which is practically necessary. While interpretability and computation depend on the structure and will be discussed for quotient space regression in Section 2.1, we may discuss consistency already here at a higher level of generality.

For a given model space \mathcal{F} consider the conditional sample Fréchet mean

$$\hat{f} = \hat{f}_n = \operatorname{argmin}_{f \in \mathcal{F}} \sum_{i=1}^n d(y_i, f(x_i))^2 \quad (3)$$

for a given set of observations $\{(x_i, y_i), i = 1, \dots, n\}$ drawn independently from (X, Y) . We first show that the estimator $\hat{f}_n \subset \mathcal{F}$, again in general not a unique function, is a consistent estimator of f^* in very general metric spaces \mathcal{Y} in the weaker sense established by Ziezold (1977) for the (unconditional) Fréchet mean. He showed that for independently and identically distributed random variables the set of empirical Fréchet means converges to the set of expected elements. Since also neither the (conditional) Fréchet mean (1) nor its empirical analogue, the (conditional) sample Fréchet mean (3) need to be unique, we can only expect a set version of consistency to hold here as well.

Lemma 2.2 (Consistency). *Let \mathcal{X} be compact and \mathcal{Y} separable. Let $\mathcal{F} \subseteq C(\mathcal{X}, \mathcal{Y})$ be a subset of the continuous functions from \mathcal{X} to \mathcal{Y} equipped with the metric $d_{\mathcal{F}}(f_1, f_2) = \sup_{x \in \mathcal{X}} d(f_1(x), f_2(x))$, $\forall f_1, f_2 \in \mathcal{F}$ and let $\mathbb{E}(d(Y, f(X))^2) < \infty \forall f \in \mathcal{F}$. Then \hat{f}_n is a strongly consistent estimator of $f^* \subseteq \mathcal{F}$ in the sense of Ziezold (1977), that is $\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} \hat{f}_k \subseteq f^*$.*

This statement is a consequence of a theorem on strong consistency of generalized Fréchet means (Huckemann, 2011), here given as Theorem A.2. See Subsection A.1 for more details. Lemma 2.2 shows that the sample conditional Fréchet mean is a consistent estimator (for $n \rightarrow \infty$) for continuous regression models, which means that consistency does not impose serious constraints on the quotient metric spaces for which we will define our regression model.

Note that this statement on consistency also holds true if $f^* = \emptyset$, i.e. if there is no $f \in \mathcal{F}$ which minimizes the total variation. To ensure $f^* \neq \emptyset$ strong additional assumptions on \mathcal{F} need to be imposed such as in the following statement (proof in Appendix A.2).

Lemma 2.3 (Existence). *Let \mathcal{X} be compact, \mathcal{Y} complete and $\mathcal{F} \subseteq C(\mathcal{X}, \mathcal{Y})$ closed and totally bounded. Then $f \mapsto \mathbb{E}(d(Y, f(X))^2)$ attains its minimum on \mathcal{F} , i.e. $f^* \neq \emptyset$.*

To motivate now a natural and interpretable model space \mathcal{F} , linear regression will serve as a prototype: in the case where \mathcal{Y} is a Hilbert space and $\mathcal{X} \subset \mathbb{R}^k$, \mathcal{F} can be chosen as the space of affine linear functions $\mathcal{X} \rightarrow \mathcal{Y}$. The minimization problem in (3) then yields an analytical solution, the minimizer \hat{f} is unique and corresponds to the usual linear predictor $\hat{f}(x_1, \dots, x_k) = \hat{\beta}_0 + \sum_{j=1}^k \hat{\beta}_j x_j$ with coefficients estimated analogously as for $\mathcal{Y} = \mathbb{R}$. That is for a design matrix $\Xi = (x_{ij})_{i=1, \dots, n; j=0, \dots, k}$ with $x_{i0} \equiv 1$ and $\mathbf{y} = (y_1, \dots, y_n) \in \otimes_{i=1}^n \mathcal{Y}$, a minimizing function $\hat{f} \in \mathcal{F}$ is given by the coefficients

$$\hat{\beta} = (\hat{\beta}_0, \dots, \hat{\beta}_k)^T = (\Xi^T \Xi)^{-1} \Xi^T \mathbf{y}, \quad (4)$$

where the matrix times vector multiplication in $\otimes_{i=1}^n \mathcal{Y}$ is defined as in \mathbb{R}^n and $\Xi^T \Xi$ is assumed to be invertible. A proof for this statement can be found in Ramsay and Dalzell (1991). Since general metric spaces, however, lack the notion of linearity, linear models cannot be directly defined here. Instead, we will use the quotient space structure to motivate a suitable generalization.

2.1 Regression in quotient metric spaces

Since we are considering regression in quotient metric spaces arising from an isometric group action, we briefly review the relevant concepts before defining our regression model for quotient spaces. We first summarize how a quotient space resulting from an isometric group action can be turned into a metric space.

Definition 2.4 (Quotient metric space). *Let (\mathcal{Y}, d) be a metric space and G a group acting on \mathcal{Y} by isometries. The quotient pseudometric d_G is defined as*

$$d_G(y_1, y_2) = \inf_{g \in G} d(y_1, g \circ y_2)$$

for all $y_1, y_2 \in \mathcal{Y}$. Since d_G defines an equivalence relation on \mathcal{Y} via $y_1 \sim y_2 \Leftrightarrow d_G(y_1, y_2) = 0$, there is a natural quotient metric space $(\mathcal{Y}/d_G, d_G)$ of \mathcal{Y} under the action G . Elements of \mathcal{Y}/d_G are denoted by $[y]$ for $y \in \mathcal{Y}$, and d_G naturally defines a metric on the equivalence classes in \mathcal{Y}/d_G .

A proof that d_G is indeed a pseudometric on \mathcal{Y} and therefore a metric on \mathcal{Y}/d_G can be found in Burago et al. (2001). We will denote $(\mathcal{Y}/G, d_G) = (\mathcal{Y}/d_G, d_G)$, although the topological quotient \mathcal{Y}/G defined via the equivalence relation $y_1 \sim y_2 \Leftrightarrow \exists g \in G : y_1 = g \circ y_2$ and \mathcal{Y}/d_G do in general not coincide. In fact, d_G does not in general define a metric on the topological quotient \mathcal{Y}/G , since there can be elements with $d_G(y_1, y_2) = 0$ for which there is no $g \in G$ such that $y_1 = g \circ y_2$. Nevertheless, the notation \mathcal{Y}/G is common, for example in the SRV-framework (Srivastava and Klassen, 2016), where \mathcal{A}/Γ is used instead of \mathcal{A}/d to denote the set of equivalence classes with respect to the elastic distance d , and we thus use it here for consistency. This notation emphasizes the dependence on the group G instead of the metric d_G it induces.

The following lemma shows that separability and completeness carry over from the original space \mathcal{Y} to the quotient. Thus, assumptions on \mathcal{Y}/G (e.g. such as needed in Lemma 2.2 and 2.3) can be reduced to those on \mathcal{Y} .

Lemma 2.5. *i) \mathcal{Y} separable $\Rightarrow (\mathcal{Y}/G, d_G)$ separable.*

ii) \mathcal{Y} complete $\Rightarrow (\mathcal{Y}/G, d_G)$ complete.

A proof for these statements can be found in the appendix. For the special case of elastic curves, 2.5 ii) was also shown in Bruveris (2016, Lemma 13).

2.1.1 Quotient regression models

Given the construction of such a quotient metric space $(\mathcal{Y}/G, d_G)$, there is a natural way to induce a model space \mathcal{F} for regression on \mathcal{Y}/G from a given model space Φ of functions $\varphi : \mathcal{X} \rightarrow \mathcal{Y}$. Given Φ , e.g. affine linear functions for the case of \mathcal{Y} a Hilbert space, we let \mathcal{F} be the space of (point-wise) projections $x \mapsto [\varphi(x)]$ of functions $\varphi \in \Phi$ on \mathcal{Y}/G which we denote by $\Phi//G = \{f : \mathcal{X} \rightarrow \mathcal{Y}/G, x \mapsto [\varphi(x)] \mid \varphi \in \Phi\}$. $\Phi//G$ is the quotient space of Φ with respect to the equivalence relation $\varphi_1 \sim \varphi_2 \Leftrightarrow \forall x \in \mathcal{X} : \varphi_1(x) \sim \varphi_2(x)$. We refer to regression with model space $\mathcal{F} = \Phi//G$ for a conditional Fréchet mean in \mathcal{Y}/G as *quotient regression* (over Φ). Note that we now focus on regression on \mathcal{Y}/G instead of the original space \mathcal{Y} , i.e we replace \mathcal{Y} by \mathcal{Y}/G in Definition 2.1, while keeping the model space denoted as \mathcal{F} for simplicity.

Definition 2.6 (Quotient regression). *Let $x_1, \dots, x_n \in \mathcal{X}$ be realizations of a random variable X and let $[y_1], \dots, [y_n] \in \mathcal{Y}/G$ be realizations of a random variable $[Y]$ taking values in \mathcal{Y}/G , where (\mathcal{Y}, d) is a metric space and G a group acting on \mathcal{Y} by isometries. Then, for a model space $\Phi = \{\varphi : \mathcal{X} \rightarrow \mathcal{Y}\}$ we define the quotient regression model (over Φ) on \mathcal{Y}/G as the conditional Fréchet mean $\mathcal{E}([Y]|X = x) = f^*(x)$ assuming*

$$f^* = \operatorname{argmin}_{f \in \Phi//G} \mathbb{E}(d_G([Y], f(X)))^2,$$

which is estimated as $\hat{f}(x) = \hat{f}_n(x) = [\hat{\varphi}(x)]$ with

$$\hat{\varphi} = \operatorname{argmin}_{\varphi \in \Phi} \sum_{i=1}^n d_G([y_i], [\varphi(x_i)])^2 = \operatorname{argmin}_{\varphi \in \Phi} \sum_{i=1}^n \inf_{g_i \in G} d(g_i \circ \varphi, \varphi(x_i))^2. \quad (5)$$

While it is not immediately clear that quotient regression is a good model for every combination of \mathcal{Y} , G and Φ , we will, in this section, give some evidence that it is in several cases. In particular, we will later illustrate its benefits in our example of elastic curve modeling based on a multiple linear spline predictor. Another example of quotient regression with model space $\Phi = \{\varphi : \mathbb{R}^k \rightarrow \mathcal{Y}, \varphi \text{ affine linear}\}$ has been suggested by Calissano et al. (2022) for the special case of \mathcal{Y} being the set of networks and G being the permutation group on the set of nodes. In particular, our result on consistency for the quotient regression model (Corollary 2.7) also applies to their case. Note that, by contrast,

approaches inducing a probability distribution on \mathcal{Y}/G via some distribution on \mathcal{Y} (such as in, e.g. offset normal shape distributions, Dryden and Mardia, 2016a, Chap. 11) are, in general, fundamentally different from our distribution-free approach that constructs the model space via projection while the mean is defined to minimize the distance d_G .

Consistency of the quotient regression model carries over from Lemma 2.2 using that separability of \mathcal{Y}/G (based on Lemma 2.5 i)) and continuity of $\Phi//G \subset C(\mathcal{X}, \mathcal{Y}/G)$ carry over from \mathcal{Y} and Φ , respectively.

Corollary 2.7 (Consistency for quotient regression). *Let \mathcal{X} be compact and \mathcal{Y} separable. Let $\Phi \subseteq C(\mathcal{X}, \mathcal{Y})$ be a subset of the continuous functions from \mathcal{X} to \mathcal{Y} , $\Phi//G$ equipped with the metric $d_{\Phi//G}$, and $\mathbb{E}(d(Y, [\varphi(X)]))^2 < \infty$ for all $\varphi \in \Phi$. Then \hat{f}_n is a strongly consistent estimator of $f^* \subseteq \mathcal{F}$ in the sense of Lemma 2.2.*

We can also formulate requirements as in Lemma 2.3 to ensure that the quotient regression model is not empty. Note that all requirements are given for the original space \mathcal{Y} and the model space Φ instead of \mathcal{Y}/G and \mathcal{F} .

Corollary 2.8 (Existence for quotient regression). *Let \mathcal{X} be compact, \mathcal{Y} complete and $\Phi \subseteq C(\mathcal{X}, \mathcal{Y})$ closed and totally bounded. Then $\varphi \mapsto \mathbb{E}(d_G([Y], [\varphi(X)]))^2$ attains its minimum on Φ .*

In the remainder of this subsection, we discuss computational aspects of quotient regression estimators. Here, quotient regression offers a straight-forward estimation scheme if, for realizations $\tilde{y}_1, \dots, \tilde{y}_n$ of a random variable \tilde{Y} in the original space \mathcal{Y} , an estimator $\tilde{\varphi}$ of $\varphi^* = \operatorname{argmin}_{\varphi \in \Phi} \mathbb{E}[d(Y, \varphi(X))^2]$ is available: in this case, we address the minimization problem in (5) via alternating 1) updating $\hat{f}(x) = [\hat{\varphi}(x)]$ by setting $\hat{\varphi}$ to the $\tilde{\varphi}$ fitting the data $(\tilde{y}_i, x_i), i = 1, \dots, n$ for current response realizations $\tilde{y}_i \in [y_i] \subset \mathcal{Y}$, and 2) optimally aligning the data, i.e. finding $\tilde{y}_i = \operatorname{argmin}_{y \in [y_i]} d(y, \tilde{\varphi}(x_i))$ for each y_i and a current estimator $\tilde{\varphi} \in \Phi$.

Alternating algorithms are natural in settings such as ours and corresponding estimation schemes have successfully been used for estimation of Fréchet means in different quotient space scenarios, including, for instance, conditional mean estimation for unlabeled networks (Calissano et al., 2022) or unconditional estimation of Procrustes means in shape analysis (Dryden and Mardia, 2016a), of elastic mean curves (Steyer et al., 2022), elastic mean shape (Srivastava and Klassen, 2016), and elastic full Procrustes mean shape estimation (Stöcker et al., 2022).

In practice it is necessary to compute numerical approximations $\tilde{\varphi} \in \Phi$ and $\tilde{y}_i = \tilde{g}_i \circ y_i$ for some $\tilde{g}_i \in G$, where true optima need not be unique or even exist in general. The algorithm iteratively reduces the loss in each step and returns a single $\hat{f} \in \Phi//G$ even in cases where the set of empirical conditional Fréchet means does not contain exactly one function. The resulting estimator \hat{f} is expected to give a good fit to the data even if technically there exists a $f \in \mathcal{F}$ with a (slightly) lower empirical loss. Such differences are likely to be small compared to the variability introduced by finite samples, and the practically relevant issue of multiple local minima can be addressed by testing different initial values.

2.1.2 Quotient geodesic regression and geodesics on the quotient space

For the case of a Riemannian manifold \mathcal{Y} , geodesic regression has been discussed by various authors (e.g., Fletcher, 2013) as natural generalization of simple linear regression on a single covariate $x \in \mathcal{X} \subset \mathbb{R}$ to curved spaces. In the context of manifolds, geodesics are typically defined as curves $c : (-\varepsilon, \varepsilon) \rightarrow \mathcal{Y}$ around some $\mu = c(0)$ with a constant velocity $\beta = \dot{c}(0)$ in the tangent space $T_\mu \mathcal{Y}$ at μ . Locally, they correspond to paths of shortest length. For general metric spaces, and in particular for a quotient metric space \mathcal{Y}/G with some general G , “geodesics” commonly directly refer to shortest paths, due to the lack of a manifold structure. As such they are less tangible, do not offer the same parameterization in terms of “intercept” μ and “slope” β , and the set of geodesics in \mathcal{Y}/G does, in general, not yield a convenient model space \mathcal{F} .

The next lemma gives a characterization of the shortest paths and therefore geodesics on the quotient metric space $(\mathcal{Y}/G, d_G)$ if (\mathcal{Y}, d) is a length metric space, i.e. if the distance d coincides with the intrinsic metric, which is the infimum of the lengths of all paths from one point to another.

Lemma 2.9 (Shortest paths in quotient metric spaces). *Let (\mathcal{Y}, d) be a length metric space and G a group acting on \mathcal{Y} by isometries. Let $y_1, y_2 \in \mathcal{Y}$ and assume there is a $\tilde{g} \in G$ with $d_G([y_1], [y_2]) = d(y_1, \tilde{g} \circ y_2)$, in which case we call y_1 and $\tilde{g} \circ y_2$ aligned. Furthermore assume there is a shortest path $\gamma : [a, b] \rightarrow \mathcal{Y}$ with $\gamma(a) = y_1$ and $\gamma(b) = \tilde{g} \circ y_2$, i.e. γ is a continuous function connecting y_1 and $\tilde{g} \circ y_2$ with minimal length. Then $[\gamma]$ is a shortest path in $(\mathcal{Y}/G, d_G)$ between $[y_1]$ and $[y_2]$, where $[\gamma]$ is the projection of γ onto \mathcal{Y}/G , i.e. $[\gamma](t) = [\gamma(t)]$ for all $t \in [a, b]$.*

A proof of this statement based on the argumentation of Burago et al. (2001) can be found in the appendix. It shows that shortest paths in the quotient metric space, and therefore geodesics, are essentially a subset of those in the original space (\mathcal{Y}, d) , for which start y_1 and end point y_2 are aligned, that is $\operatorname{argmin}_{g \in G} d(y_1, g \circ y_2) = \operatorname{id}$.

Lemma 2.9 tells us how to compute shortest paths between two given points in the quotient space with respect to the quotient metric. Yet, finding the geodesic in \mathcal{Y}/G that minimizes the squared loss in (3) with respect to d_G is still not feasible in general settings, and there is even no numerical estimation algorithm available that would promise at least reasonable practical solutions. This is, in particular, the case in our motivating example of elastic regression.

As suitable alternative, we suggest *quotient geodesic regression* for the case where \mathcal{Y} carries a Riemannian manifold structure (or in particular a Hilbert space structure) that allows for geodesic (or linear) modeling Φ , and show that the resulting model space $\Phi//G$ in fact contains the geodesics in \mathcal{Y}/G . Moreover, Simulation 5 (Fig. 7) in Section 5 gives one illustrative example of a non-geodesic model $f \in \Phi//G$ that is likely desirable to also have included in the model space, a further argument for a larger model space in practical data scenarios.

Definition 2.10 (Quotient geodesic regression). *Referring to the setting of Definition 2.6, we call quotient regression on a single covariate X in $\mathcal{X} \subset \mathbb{R}$ for a response $[Y]$ in \mathcal{Y}/G quotient geodesic regression if Φ is the set of geodesics on \mathcal{Y} .*

Given the requirements on \mathcal{X} , \mathcal{Y} and (X, Y) in Lemma 2.2, quotient regression yields a consistent estimator \hat{f} for all true $f^* \in \mathcal{F} = \Phi//G$. Accordingly, in particular all true f^* that are geodesics in \mathcal{Y}/G , which form a subset in this space (see Lemma 2.9), can be consistently estimated by quotient geodesic regression, using the quotient $\Phi//G$ of the geodesics Φ in \mathcal{Y} as a larger model space than the geodesics in \mathcal{Y}/G .

2.1.3 Quotient linear models

In quotient geodesic regression, we considered the special case of simple regression with a single covariate $x \in \mathcal{X} \subset \mathbb{R}$. We now consider multiple regression with covariates $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^k$ as a basis for our main goal of elastic regression, and focus in the following on linear models. To facilitate a suitable linear structure, we consider the important special case where \mathcal{Y} is a Hilbert space, where geodesics are straight lines and the model space Φ can be chosen as (a linear subspace of) the space of affine linear functions $\mathcal{X} \rightarrow \mathcal{Y}$. In Section 2.1.4, we will then also briefly discuss extensions to the more general case where \mathcal{Y} is a Riemannian manifold.

Definition 2.11 (Quotient linear regression with multiple scalar covariates). *Let $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,k})^\top \in \mathcal{X} \subset \mathbb{R}^k$, $i = 1, \dots, n$ be realizations of a random vector $\mathbf{X} = (X_1, \dots, X_k)^\top$ and let $[y_1], \dots, [y_n] \in \mathcal{Y}/G$ be realizations of a random variable $[Y]$ taking values in \mathcal{Y}/G , where $(\mathcal{Y}, \|\cdot\|_{\mathcal{Y}})$ is a Hilbert space and G a group acting on \mathcal{Y} by isometries. Then the quotient linear regression model is a quotient regression over a model space Φ of affine linear functions $\mathcal{X} \rightarrow \mathcal{Y}$ given by $(k+1)$ parameters in a subspace $\mathcal{B} \subset \mathcal{Y}$, that is $\mathcal{F} = \Phi//G$ with elements*

$$f : \mathbb{R}^k \rightarrow \mathcal{Y}/G; \quad \mathbf{x} = (x_1, \dots, x_k)^\top \mapsto \left[\beta_0 + \sum_{j=1}^k \beta_j x_j \right],$$

where we assume $\mathcal{E}([Y]|X_1 = x_1, \dots, X_k = x_k) = f(\mathbf{x}) = [\beta_0 + \sum_{j=1}^k \beta_j x_j]$ and the coefficients $\beta_0, \dots, \beta_k \in \mathcal{B} \subset \mathcal{Y}$ are estimated as

$$(\hat{\beta}_0, \dots, \hat{\beta}_k) = \operatorname{argmin}_{\beta_0, \dots, \beta_k \in \mathcal{B}} \sum_{i=1}^n d_G([y_i], [\beta_0 + \sum_{j=1}^k \beta_j x_{i,j}])^2 = \operatorname{argmin}_{\beta_0, \dots, \beta_k \in \mathcal{B}} \sum_{i=1}^n \inf_{g_i \in G} \left\| \beta_0 + \sum_{j=1}^k \beta_j x_{i,j} - g_i \circ y_i \right\|_{\mathcal{Y}}^2. \quad (6)$$

Thus, the estimated regression function becomes $\hat{f}(\mathbf{x}) = [\hat{\beta}_0 + \sum_{j=1}^k \hat{\beta}_j x_j]$.

For a quotient over a linear space, this generalizes the definition of the univariate quotient geodesic model 2.10 since for $k = 1$ and $\mathcal{B} = \mathcal{Y}$ the set of constant speed geodesics coincides with the set of affine linear functions Φ and therefore $\mathcal{F} = \Phi//G = \{f : \mathbb{R} \rightarrow \mathcal{Y}/G, x_1 \mapsto [\beta_0 + \beta_1 x_1]\}$ is the set of projections of constant speed geodesics.

The following corollary shows that the model space $\Phi//G$ of quotient linear regression includes geodesics on \mathcal{Y}/G not only in coordinate directions but also in any direction in the covariate space that is a convex linear combination of coordinate directions. We proof this statement in the appendix via showing that the set of elements which are aligned to one point form a convex cone (Lemma A.3).

Corollary 2.12. *Let $(\mathcal{Y}, \|\cdot\|_{\mathcal{Y}})$ be a Hilbert space and G act on \mathcal{Y} by isometries. Let $f : [0, 1]^k \rightarrow \mathcal{Y}/G, (x_1, \dots, x_k)^T \mapsto [\beta_0 + \sum_{j=1}^k x_j \beta_j]$ with $\beta_0, \beta_1, \dots, \beta_k \in \mathcal{Y}$ be such that $\beta_0 + \beta_j$ is aligned to β_0 for all $j = 1, \dots, k$. Then $f|_{x_j} : [0, 1] \rightarrow \mathcal{Y}/G, x_j \mapsto [\beta_0 + x_j \beta_j]$ is a constant speed geodesic for all $j = 1, \dots, k$ due to Lemma 2.9. Furthermore, let $\lambda_1, \dots, \lambda_k \in [0, 1]$ with $\sum_{j=1}^k \lambda_j = 1$. Then*

$$\tilde{f} : [0, 1] \rightarrow \mathcal{Y}/G, x \mapsto \left[\beta_0 + x \sum_{j=1}^k \lambda_j \beta_j \right]$$

is a constant speed geodesic in \mathcal{Y}/G between $[\beta_0]$ and $[\beta_0 + \sum_{j=1}^k \lambda_j \beta_j]$.

This generalizes geodesics to the multiple covariate setting as well as possible given the lack of a linear space structure for \mathcal{Y}/G . Such a quotient linear model has been suggested by Calissano et al. (2022) for the special case of \mathcal{Y} being the set of networks and G being the permutation group on the set of nodes. Our construction shows that their model is an example of a general class of models, which can be defined for the quotient of an arbitrary Hilbert space by a group which acts on \mathcal{Y} by isometries, and points out the inherent connection to other such cases.

In practice, the coefficients β_j will usually be modeled within a suitable finite-dimensional subspace $\mathcal{B} \subset \mathcal{Y}$, such that also $\Phi \cong \mathcal{B}^{k+1}$ will be finite-dimensional. While $\Phi//G$ then no longer necessarily contains the geodesics on \mathcal{Y}/G precisely, it may still yield good approximations to them. That the model space $\Phi \subseteq C(\mathcal{X}, \mathcal{Y})$ is a finite dimensional subspace allows us to conclude that the regression model is non-empty under weaker assumptions than in Lemma 2.8.

Theorem 2.13 (Existence in finite dimensional model spaces). *Let \mathcal{Y} be a Hilbert space, $\mathcal{X} \subset \mathbb{R}^k$ compact and $\Phi \subseteq C(\mathcal{X}, \mathcal{Y})$ a finite dimensional subspace. If $[Y]$ is bounded and $\operatorname{supp}(X) = \mathcal{X}$, there is a minimizer of $\Psi(\varphi) = \mathbb{E}(d_G([Y], [\varphi(X)])^2)$ in Φ .*

A proof of this statement can be found in the appendix. It shows that for any finite dimensional model space Ψ we can expect $f^* \neq \emptyset$, i.e. that the quotient regression model f^* in Definition 2.6 is not the empty set.

2.1.4 Side-remark on quotient regression over a Riemannian manifold

While for a single covariate geodesic regression is the canonical generalization of simple linear regression to a Riemannian manifold \mathcal{Y} , transfer of multiple linear regression to curved spaces is somewhat less straight-forward. Yet, a still natural option is given by generalized linear model (glm) type intrinsic regression (Zhu et al., 2009; Cornea et al., 2017) with a ‘‘Riemannian Log-link’’, i.e. with the model space Φ consisting of functions $\varphi : \mathbf{x} \mapsto \operatorname{Exp}_{\beta_0}(\beta_1 x_1 + \dots + \beta_k x_k)$ with intercept $\beta_0 \in \mathcal{Y}$, coefficients $\beta_1, \dots, \beta_k \in T_{\beta_0} \mathcal{Y}$ in the tangent space at β_0 , and the Riemannian exponential

map Exp at β_0 as response-function. The model models and estimates the conditional Fréchet mean with respect to the intrinsic Riemannian distance d and reduces to geodesic regression for $k = 1$. Quotient intrinsic regression over a Riemannian manifold can then be defined using $\mathcal{F} = \Phi//G$ with the above glm-type intrinsic Φ . Intrinsic regression on Kendall's shape space Σ^m of 2D landmark configurations $\mathbf{y} \in \mathbb{C}^m$ modulo translation, scale and rotation, discussed as an example by (Cornea et al., 2017), can, in fact, be considered a special case of quotient intrinsic regression with $\mathcal{Y} = \mathbb{S}^{2(m-1)}$ the sphere of dimension $2(m-1)$, Φ the model space of intrinsic regression on $\mathbb{S}^{2(m-1)}$, and the 2D rotations $G = \{\exp(\omega\sqrt{-1}) \mid \omega \in [-\pi, \pi)\}$ the isometric group action. In this case, $\mathcal{Y}/G = \Sigma^m$ carries itself a Riemannian manifold structure (of the complex projective space $\Sigma^m \cong \mathbb{C}P^{m-2}$). For shapes in higher dimensions, \mathcal{Y}/G does not carry a manifold structure anymore (Huckemann et al., 2010), but an analogous quotient intrinsic regression model could also be formulated. Additionally, an intrinsic regression model of the 2D form/size-and-shape space of \mathbf{y} modulo translation and rotation (Stöcker et al., 2023) with $\mathcal{Y} = \mathbb{C}^{m-1}$ yields another example of quotient linear regression. Hence, intrinsic regression on manifolds does not only yield a further, more general, underlying model space Φ for quotient regression, but also further motivation for the quotient (linear) model approach, since in special cases intrinsic regression models on manifolds present specially tailored quotient regression models.

2.2 Elastic regression for curves via quotient linear models in the SRV framework

In this subsection we will develop quotient regression for the particular case of curves modulo re-parametrization (and translation) in order to obtain an elastic regression model for curves. To achieve that the re-parameterization group Γ acts by isometries, we will not consider the quotient space regression model for the curves \mathbf{y} directly, but for their SRV transformation. Considering SRV transforms in the Hilbert space \mathbb{L}_2 of square integrable functions $q : [0, 1] \rightarrow \mathbb{R}^d$ induces a suitable metric on the space of absolutely continuous curves \mathcal{A} modulo translation.

Lemma 2.14 (SRV transformation (Srivastava and Klassen, 2016)). *The SRV transformation Q defined via*

$$Q(\mathbf{y})(t) = \begin{cases} \frac{\dot{\mathbf{y}}(t)}{\sqrt{\|\dot{\mathbf{y}}(t)\|}} & \text{if } \dot{\mathbf{y}}(t) \neq 0 \\ 0 & \text{if } \dot{\mathbf{y}}(t) = 0, \end{cases}$$

gives a one-to-one correspondence between the absolutely continuous curves \mathcal{A} modulo translation and the Hilbert space \mathbb{L}_2 , on which $\Gamma = \{\gamma : [0, 1] \rightarrow [0, 1] \mid \gamma \text{ monotonically increasing, onto and differentiable}\}$ acts by isometries.

More precisely, the action of Γ on the SRV transformed curves becomes $\Gamma \times \mathbb{L}_2 \rightarrow \mathbb{L}_2$, $(\gamma, \mathbf{q}) = (\mathbf{q} \circ \gamma)\sqrt{\dot{\gamma}}$, which is by isometries since $\|(\mathbf{q}_1 \circ \gamma)\sqrt{\dot{\gamma}} - (\mathbf{q}_2 \circ \gamma)\sqrt{\dot{\gamma}}\|_{\mathbb{L}_2}^2 = \int_0^1 (\mathbf{q}_1(\gamma(t)) - \mathbf{q}_2(\gamma(t)))^2 \dot{\gamma}(t) dt = \int_0^1 (\mathbf{q}_1(t) - \mathbf{q}_2(t))^2 dt = \|\mathbf{q}_1 - \mathbf{q}_2\|_{\mathbb{L}_2}^2$ for all $\gamma \in \Gamma$. That means we can define an elastic distance d on \mathcal{A}/Γ modulo translation as the quotient metric (d_G in Definition 2.4) on \mathbb{L}_2/Γ .

Definition 2.15 (Elastic distance (Srivastava and Klassen, 2016)). *Let $[\mathbf{y}_1], [\mathbf{y}_2]$ be equivalence classes in \mathcal{A}/Γ modulo translation. Then the elastic distance*

$$d([\mathbf{y}_1], [\mathbf{y}_2]) = \inf_{\gamma_1, \gamma_2 \in \Gamma} \|Q(\mathbf{y}_1 \circ \gamma_1) - Q(\mathbf{y}_2 \circ \gamma_2)\|_{\mathbb{L}_2}, \quad (7)$$

is a proper metric. Here $\|\mathbf{q}\|_{\mathbb{L}_2} = (\int_0^1 \|\mathbf{q}(t)\|^2 dt)^{1/2}$, $\mathbf{q} \in \mathbb{L}_2$, denotes the usual \mathbb{L}_2 norm.

Thus, we can define a quotient regression model for SRV curves modulo re-parametrization as in Subsection 2.1. We formulate a regression model for the elastic curves themselves using the inverse of the SRV transformation Q , which is given via $Q^{-1}(\mathbf{q})(t) = \int_0^t \mathbf{q}(s)\|\mathbf{q}(s)\| ds$ for all $\mathbf{q} \in \mathbb{L}_2$.

Definition 2.16 (Quotient SRV-linear regression for elastic curves). *Let $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,k})^\top \in \mathbb{R}^k$, $i = 1, \dots, n$ be realizations of a random vector $\mathbf{X} = (X_1, \dots, X_k)^\top$ and $\mathbf{q}_1, \dots, \mathbf{q}_n \in \mathbb{L}_2$ be SRV transformations of realizations of a random variable $[\mathbf{Y}]$ taking values in \mathcal{A}/Γ , where \mathcal{A} is the set of absolutely continuous curves from $[0, 1]$ to \mathbb{R}^d and*

Γ the set of monotonically increasing, onto and differentiable re-parametrizations. On curve level, the quotient linear regression model then becomes

$$f(\mathbf{x}) = f(x_1, \dots, x_k) = \mathcal{E}([Y]|X_1 = x_1, \dots, X_k = x_k) = [Q^{-1}(\varphi(\mathbf{x}))]$$

with linear predictor

$$\varphi(\mathbf{x}) = \beta_0 + \sum_{j=1}^k \beta_j x_j$$

on SRV-level. The coefficients $\beta_0, \dots, \beta_k \in \mathbb{L}_2$ of the regression function are estimated as

$$\operatorname{argmin}_{\beta_0, \dots, \beta_k \in \mathbb{L}_2} \sum_{i=1}^n \inf_{\gamma_i \in \Gamma} \left\| \beta_0 + \sum_{j=1}^k \beta_j x_{i,j} - (\mathbf{q}_i \circ \gamma_i) \sqrt{\gamma_i} \right\|_{\mathbb{L}_2}^2.$$

We further assume that the parameters lie in a spline space, that is $\beta_j(t) = \sum_{m=1}^M \xi_{j,m} B_m(t)$, $j = 1, \dots, k$, where $\{B_m, m = 1, \dots, M\}$ is a spline basis (e.g. linear B-splines) and $\xi_{j,m} \in \mathbb{R}^d$ for all $j = 1, \dots, k$ and $m = 1, \dots, M$. We showed identifiability modulo warping of splines from several spline spaces in Steyer et al. (2022).

For SRV-transforms $Q(Y)$ this model directly corresponds to a quotient linear model (Definition 2.11, with original space $\mathcal{Y} = \mathbb{L}_2$ and the respective isometric group action Γ implied by re-parameterization of a curve \mathbf{y} for its SRV transform $\mathbf{q} = Q(\mathbf{y})$). As such, it enjoys consistency in the sense of Corollary 2.7 and, using the finite-dimensional spline space for modeling, also existence of a Fréchet mean, i.e. $f^* \neq \emptyset$, as we showed in Theorem 2.13 in a more general setting. Due to Lemma 2.14 we can equivalently understand the model on curve level.

The minimization needed to estimate this quotient regression model for elastic curves is tackled via alternating between fitting a function-on-scalar model in each of the d dimensions for fixed γ_i , and updating the optimal re-parametrizations $\gamma_i, i = 1, \dots, n$ for fixed β s, see Algorithm 1 below. The two alternated steps are generic in the sense that suitable warping and L2 fitting steps can be combined that are tailored to the situation at hand (e.g. densely vs. sparsely observed curves). In our own implementation in the R-package `elasdics` (Steyer, 2022), since the data $\mathbf{q}_i, i = 1, \dots, n$ are SRV transformations of usually discretely observed curves, we use our methods specifically developed in Steyer et al. (2022) for potentially sparse settings for both steps. That is we replace \mathbf{q}_i by $\check{\mathbf{q}}_i$, the SRV transformation of the polygon $\check{\mathbf{y}}_i$ which is constructed via connecting the observed points linearly and choosing a constant speed parameterization. Note that this parameterization does not play a role for our model itself but only provides a suitable initial value. Also note that the relevant error made in this approximation, i.e. the difference between the polygon $\check{\mathbf{q}}_i$ and the unobserved curves \mathbf{q}_i , is the one at the SRV level. Accordingly, relatively densely observed points drawn with error at the curve level cause large errors at the SRV level (since the polygonal approximation corresponds to computing derivatives via finite differences). In this case it can be advantageous to coarsen the observed points first or to smooth them by a spline approximation on curve level.

Note that the spline model assumption is not compatible to a geodesic model assumption. Although geodesic lines are contained in the quotient space regression model assumption as shown in Lemma 2.9, geodesics between two spline curves do in general not lie in a spline space (see Subsection A.7), since aligning one spline curve to another does in general not result in a spline curve. Thus, a model can not be a geodesic model and a spline model at the same time, but we can use a spline model to approximate a geodesic model.

2.3 Extensions to closed curves

Since the space of SRV curves belonging to closed curves, $\{\mathbf{p} \in \mathbb{L}_2 \mid \int_0^1 \mathbf{p}(t) \|\mathbf{p}(t)\| dt = \mathbf{0} \in \mathbb{R}^d\}$, does not form a linear subspace in \mathbb{L}_2 , regression of closed curves cannot be treated analogously to that of open curves. While in principle it would be possible to consider the space of closed curves as a submanifold of \mathbb{L}_2 and then define the

Algorithm 1: Quotient SRV-linear regression for elastic open curves

Input: data pairs $(\mathbf{x}_i, \check{\mathbf{q}}_i)$, $i = 1, \dots, n$, where $\check{\mathbf{q}}_i$ are the SRV transformations of observed polygons $\check{\mathbf{y}}_i$ and $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,k}) \in \mathbb{R}^k$, $i = 1, \dots, n$ are observed covariates; convergence tolerance $\epsilon > 0$

Compute initial estimate $\hat{\boldsymbol{\beta}}_{0,new}, \dots, \hat{\boldsymbol{\beta}}_{k,new} = \operatorname{argmin}_{\boldsymbol{\beta}_0, \dots, \boldsymbol{\beta}_k \in \mathbb{L}_2} \sum_{i=1}^n \|\boldsymbol{\beta}_0 + \sum_{j=1}^k \boldsymbol{\beta}_j x_{i,j} - \check{\mathbf{q}}_i\|_{\mathbb{L}_2}^2$;

Set $\hat{\boldsymbol{\beta}}_{j,old} = \operatorname{Inf} \quad \forall j = 0, \dots, k$;

while $\max_{j=0, \dots, k} \|\hat{\boldsymbol{\beta}}_{j,old} - \hat{\boldsymbol{\beta}}_{j,new}\|_{\mathbb{L}_2}^2 > \epsilon$ **do**

$\hat{\boldsymbol{\beta}}_{j,old} = \hat{\boldsymbol{\beta}}_{j,new} \quad \forall j = 0, \dots, k$;

$\gamma_i = \operatorname{argmin}_{\gamma} \left\| \hat{\boldsymbol{\beta}}_{0,old} + \sum_{j=1}^k \hat{\boldsymbol{\beta}}_{j,old} x_{i,j} - (\check{\mathbf{q}}_i \circ \gamma) \sqrt{\gamma} \right\|_{\mathbb{L}_2}^2, \quad \forall i = 1, \dots, n; \quad // \text{warping step}$

$\hat{\boldsymbol{\beta}}_{0,new}, \dots, \hat{\boldsymbol{\beta}}_{k,new} = \operatorname{arginf}_{\boldsymbol{\beta}_0, \dots, \boldsymbol{\beta}_k \in \mathbb{L}_2} \sum_{i=1}^n \left\| \boldsymbol{\beta}_0 + \sum_{j=1}^k \boldsymbol{\beta}_j x_{i,j} - (\check{\mathbf{q}}_i \circ \gamma_i) \sqrt{\gamma_i} \right\|_{\mathbb{L}_2}^2$
// \mathbb{L}_2 spline fit via least-squares

return $\hat{\boldsymbol{\beta}}_j = \hat{\boldsymbol{\beta}}_{j,new} \quad \forall j = 0, \dots, k$

quotient regression model on this submanifold modulo warping, to the best of our knowledge there are no methods to compute minimizing geodesics on this submanifold. (Srivastava and Klassen (2016) provide algorithms for numerical computation of geodesics between two closed curves – extending this to finding a minimizing geodesic through a sample of curves is, however, not straightforward). For this reason, we do not focus on closed curves here. However, as closed curves often appear naturally in practical applications, we describe at least a heuristic method for the regression of closed curves based on quotient regression for open curves. This method is also implemented in the R-package `elasdics` (Steyer, 2022).

Specifically, we treat the curves as open curves in the \mathbb{L}_2 fitting step, but restrict the splines we use for modeling their SRV transforms to be closed (which is necessary but not sufficient for closedness of the modeled curves, ensuring matching derivatives at starting and end points). Then we close the predictions via projecting them onto the space of derivatives belonging to closed curves: Since we model the SRV transform \mathbf{p} as a spline and therefore bounded curve, the corresponding derivative $\mathbf{p} \|\mathbf{p}\|$ is also bounded and therefore in \mathbb{L}_2 . Hence we can consider the space $\{\mathbf{p} \|\mathbf{p}\| \in \mathbb{L}_2 \mid \int_0^1 \mathbf{p}(t) \|\mathbf{p}(t)\| dt = 0\}$, which is a linear subspace of the Hilbert space \mathbb{L}_2 , and compute the orthogonal projection of $\mathbf{p} \|\mathbf{p}\|$ onto this space as $\mathbf{p} \|\mathbf{p}\| - \int_0^1 \mathbf{p}(s) \|\mathbf{p}(s)\| ds$. Thus, the prediction on curve level becomes $t \mapsto \int_0^t \mathbf{p}(s) \|\mathbf{p}(s)\| ds - t \cdot \int_0^1 \mathbf{p}(s) \|\mathbf{p}(s)\| ds$, which is a closed curve. We use these closed predictions in the iterative algorithm 1 to replace the $\hat{\boldsymbol{\beta}}_{j,old}$ when aligning the observations in each iteration (warping step). See Algorithm 2 in the Appendix for details.

3 Alternative regression approaches

Although there are so far no direct competitors available to our quotient regression for curves modulo re-parametrization, we discuss in the following different approaches that we consider natural alternatives. Comparison to these alternatives may be relevant beyond our specific focus as they exemplify a) pre-alignment as natural alternative to quotient regression with responses in any quotient metric space, b) statistical modeling on curve level with only alignment based on SRV transforms, and c) usage of a generic approach for metric spaces without using the quotient structure. The first three alternatives we give are new proposals reflecting combinations of a) and b), while for c) Fréchet regression in Subsection 3.3 constitutes an existing general approach, which has to be adapted to and implemented for our setting, and for which we give a novel concrete implementation for the elastic regression case. All methods discussed here will then be used as comparison methods to benchmark our quotient regression approach in simulations in Section 5.

3.1 Regression after pre-alignment

For elastic regression as for general quotient metric spaces \mathcal{Y}/G where $(\mathcal{Y}, \|\cdot\|_{\mathcal{Y}})$ is a Hilbert space, an obvious competitor of the quotient linear model is to fit a linear model on the original space \mathcal{Y} after once pre-aligning the data $y_i, i = 1, \dots, n$ to its (marginal) Fréchet mean $\mu_0 = \operatorname{arginf}_{\mu \in \mathcal{Y}} \sum_{i=1}^n d_G([y_i], [\mu])^2$. Here, we consider the model with predictor $f : x \mapsto [\varphi(x)]$ and the estimator $\hat{\varphi}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_k x_k$ given by

$$\hat{\beta} = (\hat{\beta}_0, \dots, \hat{\beta}_k)^\top = \operatorname{argmin}_{\beta_0, \dots, \beta_k \in \mathcal{Y}} \sum_{i=1}^n \left\| \beta_0 + \sum_{j=1}^k \beta_j x_{i,j} - g_i^* \circ y_i \right\|_{\mathcal{Y}}^2$$

where $g_i^* = \operatorname{argmin}_{g \in G} \|\mu_0 - g \circ y_i\|_{\mathcal{Y}}$. Here we assume that there exists an optimal alignment to the mean for all y_i 's. The minimiser $\hat{\beta}$ can be computed as $\hat{\beta} = (\Xi^\top \Xi)^{-1} \Xi^\top (\mathbf{g}^* \circ \mathbf{y})$, where $\Xi \in \mathbb{R}^{n \times k}$ is the design matrix and $\mathbf{g}^* \circ \mathbf{y} = (g_1^* \circ y_1, \dots, g_n^* \circ y_n)^\top \in \otimes_{i=1}^n \mathcal{Y}$.

Although the model space Φ also consists of affine linear functions in \mathcal{Y} , this is not an intrinsic regression, i.e. we do not truly consider its projection to the quotient $\Phi//G$ as model space on \mathcal{Y}/G here. That means no attempt is made to minimize the empirical risk (6) with respect to the quotient space distance d_G and therefore, this risk will always be greater than or equal to that for the quotient space regression model.

In the specific case that we want to model curves with respect to the elastic distance (7), this means computing a linear model for the SRV transformed curves in \mathbb{L}_2 after pre-aligning the corresponding data curves to the elastic mean. That is

$$(\hat{\beta}_0, \dots, \hat{\beta}_k) = \operatorname{argmin}_{\beta_0, \dots, \beta_k \in \mathbb{L}_2} \sum_{i=1}^n \left\| \beta_0 + \sum_{j=1}^k \beta_j x_{i,j} - (\mathbf{q}_i \circ \gamma_i) \sqrt{\gamma_i} \right\|_{\mathbb{L}_2}^2$$

with $\gamma_i = \operatorname{argmin}_{\gamma \in \Gamma} \|\mu_0 - (\mathbf{q}_i \circ \gamma) \sqrt{\gamma}\|_{\mathbb{L}_2}$ and μ_0 is the SRV transformation of the elastic mean curve. (Guo et al., 2020) propose a similar procedure, where they then use the principal component scores of the pre-aligned SRV curves in a simple regression model. In contrast, we use splines to model the β s and the alignment methods developed in Steyer et al. (2022) to enable fitting of irregularly and/or sparsely observed curves and to allow better comparison with our quotient regression model for elastic curves (Definition 2.16). We refer to this procedure as 'pre-align, srv fit' in the following.

3.2 Alternative procedures with fit on curve level

Considering pre-alignment of the data curves $\mathbf{y}_1, \dots, \mathbf{y}_n$ with SRV transformations $\mathbf{q}_1, \dots, \mathbf{q}_n$ to their elastic mean curve a pre-processing step, it might also deem natural to compute the regression model on curve level instead of on SRV level. We call this approach 'pre-align, curve fit'. Here, the fitted predictor is given by $\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_k x_k$ with

$$(\hat{\beta}_0, \dots, \hat{\beta}_k) = \operatorname{argmin}_{\beta_0, \dots, \beta_k \in \mathbb{L}_2} \sum_{i=1}^n \left\| \beta_0 + \sum_{j=1}^k \beta_j x_{i,j} - \mathbf{y}_i \circ \gamma_i^* \right\|_{\mathbb{L}_2}^2$$

where $\gamma_i^* = \operatorname{argmin}_{\gamma \in \Gamma} \|\mu_0 - (\mathbf{q}_i \circ \gamma) \sqrt{\gamma}\|_{\mathbb{L}_2}$ and μ_0 is again the SRV transformation of the elastic mean curve. This is tempting in particular if we want to fit closed curves since, on curve level, closed curves can be modeled without further modifications using a closed spline basis for the model coefficients β_0, \dots, β_k .

We further consider a heuristic procedure in which we alternate between optimal alignment and regression fit as in the quotient regression approach, but fit the linear model on curve level rather than on SRV level ('iterate align, curve fit'). This is not a suitable method for fitting the quotient regression model with respect to the elastic distance, because the

elastic distance becomes the usual \mathbb{L}_2 metric only for SRV transforms. Fitting the linear model on curve instead of on SRV level will not return a minimizer of the squared elastic distances to the data curves. In fact, there is no risk function that this algorithm aims to minimize, and the procedure is thus only defined by the iterative algorithm rather than being the fitting algorithm of a regression model.

Moreover, both procedures with linear model fit on curve level do not include geodesics with respect to the elastic distance in their model space, i.e., they are not suitable to generalize linear regression in this sense.

3.3 Fréchet regression

So far we considered models that exploit the linear space structure of either the space on SRV or on curve level to define regression models for curves with respect to the elastic distance. In contrast, Petersen and Müller (2019) developed a regression model they call Fréchet regression for random objects lying in arbitrary metric spaces with covariates in \mathbb{R}^k , which does not rely on any linear structure. They achieve this by noting that in standard linear regression, the regression function can be viewed as a function mapping the input $\mathbf{x} \in \mathbb{R}^k$ to a weighted mean of the y_i , where only the weights depend on \mathbf{x} . Their Fréchet regression model then extends standard linear regression by using the same weights with an arbitrary metric instead of the Euclidean distance, i.e. using a weighted Fréchet mean. Although this implicitly defines a regression model for arbitrary metric spaces, without explicit model equation however, details and complexity of the estimation depend on the specific space considered. Petersen and Müller (2019) discuss in their paper the case of probability distributions equipped with the Wasserstein metric as well as the case of covariance matrices. For both cases, there is an implementation in the R package `frechet` (Chen et al., 2020). To the best of our knowledge, the case of curves with respect to the elastic distance has not yet been considered, so we describe below how we estimate the Fréchet regression model in this case.

For observed curves with SRV transforms $\mathbf{q}_1, \dots, \mathbf{q}_n \in \mathbb{L}_2$, the predictor f for an input vector $\mathbf{x} \in \mathbb{R}^k$ is given by

$$f(\mathbf{x}) = \operatorname{argmin}_{\mathbf{p} \in \mathbb{L}_2} \sum_{i=1}^n s(\mathbf{x}_i, \mathbf{x}) \inf_{\gamma_i \in \Gamma} \|\mathbf{p} - (\mathbf{q}_i \circ \gamma_i) \sqrt{\gamma_i}\|_{\mathbb{L}_2}^2,$$

via a point-wise optimization function where the weights (Petersen and Müller, 2019) are given as $s(\mathbf{x}_i, \mathbf{x}) = 1 + (\mathbf{x}_i - \bar{\mathbf{x}})^\top \hat{\Sigma}^{-1} (\mathbf{x} - \bar{\mathbf{x}})$. Here $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$ is the mean of the observed covariates and $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$ their empirical covariance matrix. Thus, for a given input value $\mathbf{x} \in \mathbb{R}^k$, the conditional mean response curve is computed as a weighted Fréchet mean with respect to the elastic distance using weights $s(\mathbf{x}_i, \mathbf{x})$. For the particular case of the space of SRV curves with the elastic metric, we propose to consider the observed polygons $\check{\mathbf{y}}_1, \dots, \check{\mathbf{y}}_n$ with SRV transformations $\check{\mathbf{q}}_1, \dots, \check{\mathbf{q}}_n$ as we do for our quotient space regression model to handle discretely observed curves. Then we estimate the weighted Fréchet mean via alternating between updating the optimal re-parametrizations $\gamma_i \in \Gamma$ as $\operatorname{argmin}_{\gamma_i \in \Gamma} \|\mathbf{p} - (\check{\mathbf{q}}_i \circ \gamma_i) \sqrt{\gamma_i}\|_{\mathbb{L}_2}$ for a given $\mathbf{p} \in \mathbb{L}_2$, using our alignment methods developed in Steyer et al. (2022) to align discretely observed curves to a model based curve, and computing the weighted \mathbb{L}_2 -mean $\operatorname{argmin}_{\mathbf{p}} \sum_{i=1}^n s(\mathbf{x}_i, \mathbf{x}) \|\mathbf{p} - (\check{\mathbf{q}}_i \circ \gamma_i) \sqrt{\gamma_i}\|_{\mathbb{L}_2}^2$ for given alignments γ_i . For the mean estimation step we propose to use splines, as we do for the quotient regression model. Details are given as Algorithm 3 in the Appendix.

One disadvantage of this approach is that the regression function is not given by a set of parameters, such as slopes and intercepts. In fact, for every given input vector $\mathbf{x} \in \mathbb{R}^k$, the value of the regression function has to be estimated separately as a weighted mean. This makes interpretation of the model more challenging and estimation more time consuming. One advantage in the SRV context is that handling closed curves is straightforward, as we can compute closed (weighted) Fréchet means using results in Steyer et al. (2022).

3.4 Differences in curve alignment implied by the different approaches

To gain an understanding of the differences between the proposed approaches, we compare how the observed curves are aligned during the fitting process and discuss the implications of these differences in specific data scenarios. When fitting the quotient regression model, we align the observed curves \check{y}_i to the model based predictions $Q^{-1}(\hat{\beta}_0 + \sum_{j=1}^k \hat{\beta}_j x_{i,j})$ for all $i = 1, \dots, n$. This means that each observed \check{y}_i is aligned to a model-based curve that is expected to have similar features as the observation. Likewise, in the 'iterate align, curve fit' approach, the observed \check{y}_i is aligned to its associated prediction.

In contrast, the pre-alignment methods 'pre-align, srv fit' and 'pre-align, curve fit' align the curves to the elastic mean, hence may not properly align certain features of the curves if these features occur in specific directions of \mathbf{x} that are missing in the mean curve. Similarly, in the fitting algorithm for the Fréchet regression model (Algorithm 3) the observed curves \check{y}_i are aligned to the model prediction for each considered new value of \mathbf{x} , which is usually different from \mathbf{x}_i . Accordingly, we also expect less convincing results for this model in situations where certain features of the curves occur only for some values of \mathbf{x} but not for others.

Overall, we expect all five methods to provide satisfactory results in scenarios where all observed curves have similar features, and that the quotient regression model outperforms the fits after pre-alignment as well as the Fréchet regression model when some features are missing in the elastic mean curve respectively some of the curves. For the 'iterate align, curve fit' approach, the behavior is more difficult to anticipate, as its iterative procedure optimizes no loss function. We will investigate these expectations for model performance in simulations in Section 5. Besides that Section 5 will also cover simulations on methods of inference in quotient regression, which we describe beforehand in the next section.

4 Inference and model selection

4.1 A generalized coefficient of determination

Both Fréchet regression and the quotient regression are defined as empirical risk minimization problems in one way or another. Petersen and Müller (2019) generalize the coefficient of determination R^2 to models with values y_1, \dots, y_n , $n \in \mathbb{N}$ in metric spaces (\mathcal{Y}, d) . For an estimated model equation \hat{f} their Fréchet coefficient of determination is given as

$$\tilde{R}^2 = 1 - \frac{\sum_{i=1}^n d(y_i, \hat{f}(\mathbf{x}_i))^2}{\sum_{i=1}^n d(y_i, \hat{\mu}_0)^2}$$

where $\hat{\mu}_0 = \operatorname{argmin}_{\mu \in \mathcal{Y}} \sum_{i=1}^n d(y_i, \mu)^2$ is the Fréchet mean of the data. Note that if constant functions are contained in the model space in which \hat{f} is estimated, we have $\tilde{R}^2 \in [0, 1]$ as for R^2 in standard linear regression. In this case testing the global null hypothesis of no effect, that is $f(\mathbf{x}) \equiv \mu_0$ constant, is equivalent to testing $H_0 : \tilde{R}^2 = 0$. The distribution of the test statistic \tilde{R}^2 under H_0 is available via permutation re-sampling of the data, i.e randomly permuting the labels of the response variable y_i while keeping the covariates \mathbf{x}_i fixed. They further suggest to use an adjusted coefficient of determination $\tilde{R}_{\text{adj}}^2 = 1 - (1 - \tilde{R}^2) \frac{n-1}{n-k-1}$ for model selection, where k accounts for the number of covariates $\mathbf{x}_i = (x_{i1}, \dots, x_{ik})^T$ in the model.

4.2 Distance-based bootstrap confidence regions

To obtain confidence regions for the predicted curves we propose to bootstrap the data $(\mathbf{x}_i, \check{y}_i)$, $i = 1, \dots, n$ to obtain an approximate sample of the model predictions, $\hat{y}_1, \dots, \hat{y}_{N_{\text{boot}}}$, for a given \mathbf{x} . From this we construct a $(1 - \alpha)$ -confidence region as a generalized convex hull (Edelsbrunner et al., 1983, α -shapes), of the (centered, i.e we subtract the center of mass for each predicted curve) $\lceil (1 - \alpha)N_{\text{boot}} \rceil$ closest curves to the bootstrap mean with respect to the elastic distance. Note that when the bootstrapped curves form a relatively dense set, directly plotting the $(1 - \alpha)$ closest curves gives a good and simple visual approximation to plotting the generalized convex hull in practice.

4.3 Bootstrap confidence regions based on spline coefficients

Inference as described above can be conducted for approaches without a parametric model equation, such as Fréchet regression, and parametric models, such as the quotient regression model, which provide estimates for intercept and slope parameters. However, since our quotient linear model for elastic curves is a parametric model, we are not only interested in the global null hypothesis of none of the covariates having an effect, but also want to assess the relevance of individual parameters. We propose to test individual hypotheses by bootstrapping the data $(\mathbf{x}_i, \check{\mathbf{y}}_i)$, $i = 1, \dots, n$ to obtain an approximate sample from the distribution of the estimated model parameters $\hat{\boldsymbol{\beta}}_0, \dots, \hat{\boldsymbol{\beta}}_k$. Confidence regions for the parameters can then be constructed from this sample and used to decide whether a particular parameter, for instance $\boldsymbol{\beta}_j = \mathbf{0}$ corresponding to no effect, is plausible given the observed data, as detailed below.

Our proposed representation of the coefficient functions $\boldsymbol{\beta}_j(t) = \sum_{m=1}^M \boldsymbol{\xi}_{j,m} B_m(t)$, $t \in [0, 1]$, $j = 1, \dots, k$ has the additional advantage that using a linear combination of spline basis functions $B_m(t)$, $m = 1, \dots, M$ with local support, such as B-splines, also allows to test local individual hypotheses on subintervals of $[0, 1]$, i.e. to test where a given covariate affects the response curve. We have shown in Steyer et al. (2022) that linear splines on SRV level (among other splines) are identifiable via their spline coefficients modulo parametrization, and that the mapping between the spline coefficients and the elastic curves is a homeomorphism. We can thus use the variation in the spline coefficients as representative of that in the estimated effects and construct alternative confidence regions as outlined in the following. Note that this alternative to Section 4.2 is, however, only recommended when estimates are sufficiently concentrated, as we will briefly discuss in Section 4.4.

We construct a $(1 - \alpha)$ -confidence region for $\boldsymbol{\beta}_j$ based on the bootstrapped spline coefficients $\boldsymbol{\xi}_{j,m}^{(b)}$, $b = 1, \dots, N_{boot}$ as the d -dimensional ellipse

$$C_{j,m,\alpha} = \{ \boldsymbol{\xi} \in \mathbb{R}^d \mid (\boldsymbol{\xi} - \bar{\boldsymbol{\xi}}_{j,m})^T \hat{\boldsymbol{\Sigma}}_{j,m}^{-1} (\boldsymbol{\xi} - \bar{\boldsymbol{\xi}}_{j,m}) \leq c_{j,1-\alpha} \},$$

where $\bar{\boldsymbol{\xi}}_{j,m} = \frac{1}{N_{boot}} \sum_{b=1}^{N_{boot}} \boldsymbol{\xi}_{j,m}^{(b)}$ is the bootstrap mean, $\hat{\boldsymbol{\Sigma}}_{j,m} = \frac{1}{N_{boot}-1} \sum_{b=1}^{N_{boot}} (\boldsymbol{\xi}_{j,m}^{(b)} - \bar{\boldsymbol{\xi}}_{j,m})(\boldsymbol{\xi}_{j,m}^{(b)} - \bar{\boldsymbol{\xi}}_{j,m})^T$ is the bootstrap sample covariance and $c_{j,m,1-\alpha}$ the empirical $(1 - \alpha)$ -quantile of the studentized bootstrap sample $\{(\boldsymbol{\xi}_{j,m}^{(b)} - \bar{\boldsymbol{\xi}}_{j,m})^T \hat{\boldsymbol{\Sigma}}_{j,m}^{-1} (\boldsymbol{\xi}_{j,m}^{(b)} - \bar{\boldsymbol{\xi}}_{j,m}) \mid b = 1, \dots, N_{boot}\}$ for all $j = 1, \dots, k$. From this confidence regions for the coefficients $\boldsymbol{\xi}_{j,m}$ on can proceed to construct pointwise confidence regions for the corresponding effect functions $\boldsymbol{\beta}_j$. Moreover, $C_{j,m,\alpha}$ can also be used to test the local individual hypothesis $H_{0,j,m} : \boldsymbol{\xi}_{j,m} = \mathbf{0}$ by checking for overlap with $\mathbf{0}$. Using these confidence regions for the single spline coefficients, we construct a joint $(1 - \alpha)$ -confidence region for the matrix of spline coefficients $\boldsymbol{\xi}_j = (\boldsymbol{\xi}_{j,1}, \dots, \boldsymbol{\xi}_{j,M})^T \in \mathbb{R}^{M \times d}$ corresponding to the effect function $\boldsymbol{\beta}_j$ as $C_{j,\alpha} = \times_{m=1}^M C_{j,m,\frac{\alpha}{M}}$, where $\frac{\alpha}{M}$ is a Bonferroni-type correction of the confidence level. Hence $P(\boldsymbol{\xi}_j \in C_{j,\alpha}) = P(\bigcap_{m=1}^M \{\boldsymbol{\xi}_{j,m} \in C_{j,m,\frac{\alpha}{M}}\}) = 1 - P(\bigcup_{m=1}^M \{\boldsymbol{\xi}_{j,m} \notin C_{j,m,\frac{\alpha}{M}}\}) \geq 1 - \sum_{m=1}^M P(\{\boldsymbol{\xi}_{j,m} \notin C_{j,m,\frac{\alpha}{M}}\}) \geq 1 - \alpha$, if $C_{j,m,\frac{\alpha}{M}}$ is a valid confidence region, i.e. fulfills $P(\{\boldsymbol{\xi}_{j,m} \notin C_{j,m,\frac{\alpha}{M}}\}) \leq \frac{\alpha}{M}$.

The constructed confidence region can be utilized to test the individual hypothesis $H_{0,j} : \boldsymbol{\beta}_j = \mathbf{0}$. This is done by rejecting $H_{0,j}$ if and only if $\mathbf{0} \notin C_{j,m}$, which is equivalent to $\bar{\boldsymbol{\xi}}_{j,m}^T \hat{\boldsymbol{\Sigma}}_{j,m}^{-1} \bar{\boldsymbol{\xi}}_{j,m} \geq c_{j,1-\frac{\alpha}{M}}$ for at least one $m = 1, \dots, M$. We thus use $\max\{\bar{\boldsymbol{\xi}}_{j,m}^T \hat{\boldsymbol{\Sigma}}_{j,m}^{-1} \bar{\boldsymbol{\xi}}_{j,m} \mid m = 1, \dots, M\}$ as a test statistic. Since the resulting test relies on the local representation property of the spline coefficients for the effect functions and, as a bootstrap method, also on the interchangeability of the data generating distribution with the empirical distribution, we examine the validity and power of the test in a simulation in the following subsection.

4.4 Distance vs. spline coefficient based confidence regions

The idea of the spline coefficient based confidence regions proposed in Section 4.3 is based on the assumption that the distribution of the $\hat{\boldsymbol{\beta}}_j$, or alternatively the bootstrap samples $\hat{\boldsymbol{\beta}}_j^{(b)}$, is reflected well by an elliptical distribution of the re-

spective spline coefficients $\hat{\xi}_j^{(b)}$. Despite identifiability of the used piece-wise linear splines modulo re-parameterization (Steyer et al., 2022), this does not necessarily have to be the case. In particular, if two estimators $\hat{\beta}_j^{(1)}$ and $\hat{\beta}_j^{(2)}$ differ too much, different curve alignment may result in the m -th spline coefficients $\hat{\xi}_{j,m}^{(1)}$ and $\hat{\xi}_{j,m}^{(2)}$ of each of them corresponding to different segments of the curves, which might occur especially when we estimate very flexible curves with many basis functions relative to the sample size. In these cases, inference based on the spline coefficients will lead to a loss in power due to the added variability of the parameterization, and the distance-based methods described in Section 4.2 should be used. Conversely, if the estimators $\hat{\beta}_j^{(b)}$ are sufficiently concentrated, spline coefficient based methods allow for local investigation and might yield more power since they make use of elliptical confidence regions rather than depending on the distance to the bootstrap mean only.

5 Simulations

We first compare in simulations the quotient regression model with the alternative procedures presented in Section 3. Then, in the second part of this section, we examine the test for the parameters of the quotient regression model based on the bootstrapped spline coefficients.

5.1 Comparison of model performance

We compare the quotient linear model to the procedures described in Section 3. To this end, we choose three simulation scenarios for each of which we add errors of different magnitude and draw a varying number of points per curve. The predictive performance is then determined on an independent test set drawn according to the same principle, using the mean squared (elastic) distance (MSE) of the new observations to their predicted curves. Evaluation on a test set rather than on a true underlying model is necessary because quotient regression as well as Fréchet regression are defined as risk minimization problems and no distribution is available that would allow us to draw random curves with a specific conditional Fréchet mean structure. With the auxiliary sampling scheme used instead, we may specify a template model but there is no precise ‘true’ model explicitly available that we can compare the model estimates to. Each of the $3 \times 4 = 12$ simulations is then repeated 100 times to obtain a stable estimate of the MSE.

sce- nario	sim	sd	$\kappa_i \in$	MSE					Average run time in seconds				
				quotient space regres- sion	pre align, SRV fit	iterate align, curve fit	pre align, curve fit	Fréchet regres- sion	quotient space regres- sion	pre align, SRV fit	iterate align, curve fit	pre align, curve fit	Fréchet regres- sion
1	1	0.4	[15, 20]	0.57	0.66	0.65	0.71	0.59	12	5	9	5	56
1	2	0.8	[15, 20]	0.88	0.95	0.96	1.00	0.89	13	5	10	5	67
1	3	0.4	[30, 40]	0.32	0.39	0.37	0.44	0.33	83	28	54	24	528
1	4	0.8	[30, 40]	0.74	0.82	0.80	0.85	0.76	51	18	34	17	338
2	5	0.2	[15, 20]	0.35	0.59	0.38	0.54	0.37	14	14	25	14	105
2	6	0.4	[15, 20]	0.43	0.67	0.46	0.62	0.45	4	4	8	4	33
2	7	0.2	[30, 40]	0.19	0.41	0.22	0.37	0.22	16	11	28	10	141
2	8	0.4	[30, 40]	0.31	0.55	0.35	0.50	0.34	16	11	31	11	195
3	9	0.1	[15, 20]	0.78	0.89	0.81	0.93	0.84	32	90	44	82	1922
3	10	0.2	[15, 20]	1.37	1.49	1.39	1.52	1.41	38	102	58	97	1769
3	11	0.1	[30, 40]	0.95	1.06	0.95	1.08	0.99	14	47	22	47	707
3	12	0.2	[30, 40]	2.80	2.96	2.79	2.95	2.81	13	48	20	48	528

Table 1: Mean squared elastic distance (MSE) estimated out of sample (smallest per row in bold) and average run time of one estimation for the five methods in three different scenarios with a varying error magnitude (sd) and number of points drawn per curve, where the number of points κ_i is drawn uniformly on a given range (15 to 20 or 30 to 40 points per curve). This gives a total of 12 simulations (sim).

The three different simulation scenarios differ regarding which curves are used as models for $x = -1$ and for $x = 1$ and whether the trajectory between them is modeled linearly on SRV or on curve level. For the first scenario (simulations

1-4, see Fig. 1 (left) for an example of simulation 1), we use similar fish shapes to model the curves for $x = -1$ and for $x = 1$, and consider the geodesic between them (i.e. linear on SRV level with curves aligned). This setting is meant to be advantageous to methods using pre-alignment to the mean curve (fish), which should give good alignment among all curves, and we expect that all five methods should be able to model this type of data well. In contrast, in the second scenario (simulations 5-8, see Fig. 1 (middle) for an example of simulation 5) we also consider a linear relationship of the covariate x with the SRV curves, but not a geodesic (i.e. no alignment with respect to the elastic distance between endpoints) between the curves for $x = -1$ (fish with open mouth) and $x = 1$ (fish with closed mouth). This seems natural since aligning the modeled curves here would not match the back end of the open mouth ($x = -1$) with the tip of the closed mouth ($x = 1$). In this setting we expect that pre-aligning the data to the elastic mean will not properly align the open/closed mouth of the fish and therefore the quotient linear model is beneficial. In the last simulation scenario (simulations 9-12, see Fig. 1 (right) for an example of simulation 11), we consider model misspecification in the sense that the effect of the covariate $x \in [-1, 1]$ is simulated linearly on curve instead of on SRV level. Additionally, in this setting we investigate the quality of our approach to modeling smooth, closed contours, here simulating closed quadratic spline curves.

To generate observations for the first scenario, we first obtain $n = 11$ smooth curves for $x = -1, -0.8, \dots, 0.8, 1$ as the convex combinations of the SRV transformed modeled curves for $x = -1$ and $x = 1$. Next we evaluate them on a regular grid of 51 points from which we compute 50 SRV vector via finite difference approximation of the derivative. After adding a Gaussian 1st order random walk error with standard deviation sd to these SRV vectors we back transform them to the curve level and select κ_i of the resulting 51 points, where κ_i is drawn uniformly on the given interval, to obtain sparse/irregular settings. Here we choose relatively small standard deviations sd for the additional noise compared to the effect, since we want to focus on demonstrating structural differences between possible effects on the curves and how the different methods handle those.

Each of the five regression models/procedures is fitted to these data assuming linear SRV splines with 11 knots for the quotient space regression model, the fit on SRV level after pre-alignment and the Fréchet regression model, and quadratic splines also with 11 knots for the models with fit on curve level. This results in the same model flexibility for all five models modulo translation. Since in this scenario, the modeled curves for $x = -1$ and $x = 1$ are approximately aligned, we expect all five methods to give reasonable results. This is confirmed visually in the model predictions (Fig. 1, left, and Fig. 6), but the MSEs in rows 1-4 of Tab. 1 reveal that the quotient space regression model performs best for this scenario and all combinations of sd and κ_i .

The data for the second scenario, i.e., simulations 5-8, are generated in the same way as the data for the first scenario, except that the shape of the modeled curves for $x = -1$ and $x = 1$ differs more and we do not consider the geodesic between them as the generating model. Since in this scenario the modeled curves have sharp edges around the mouth, we use constant splines on SRV-level corresponding to linear splines on curve level and 51 knots for all five procedures (see Fig. 1, middle, and Fig. 7 for an example of simulation 5). In this setting pre-aligning the data to the elastic mean (which also corresponds to the model prediction for $x = 0$ of the Fréchet regression model) will not properly align the open/closed mouth of the fish. Thus, a procedure that pre-aligns and then fits a model is not able to fit the open mouth of the fish for $x = -1$ (see Fig. 1, middle). Similarly, for the Fréchet model fit the open mouth appears too small, as well as the whole predicted curve for $x = -1$ and $x = 1$. This is the case since for fitting the Fréchet model, we also align fish with open and closed mouth, since for each new value of x , we align all data curves to the corresponding new prediction (cf. Algorithm 3). Hence in this setting only the quotient regression model gives visually satisfying results, which is also reflected in the MSEs of the five models (Tab. 1, simulations 5 to 8). Here the MSE is always the smallest for the quotient regression model followed by Fréchet regression and the heuristic procedure of iterating between alignment and curve-level fit. We expect this to be the case in general if features of the curves (as for example the open mouth of the fish) that occur in certain directions of x are missing in the mean curve.

For the last simulation scenario (simulations 9 to 12) we not only choose the model to be linear on curve instead of on SRV level and use closed quadratic spline curves here to generate smooth, closed contours, we also add the random

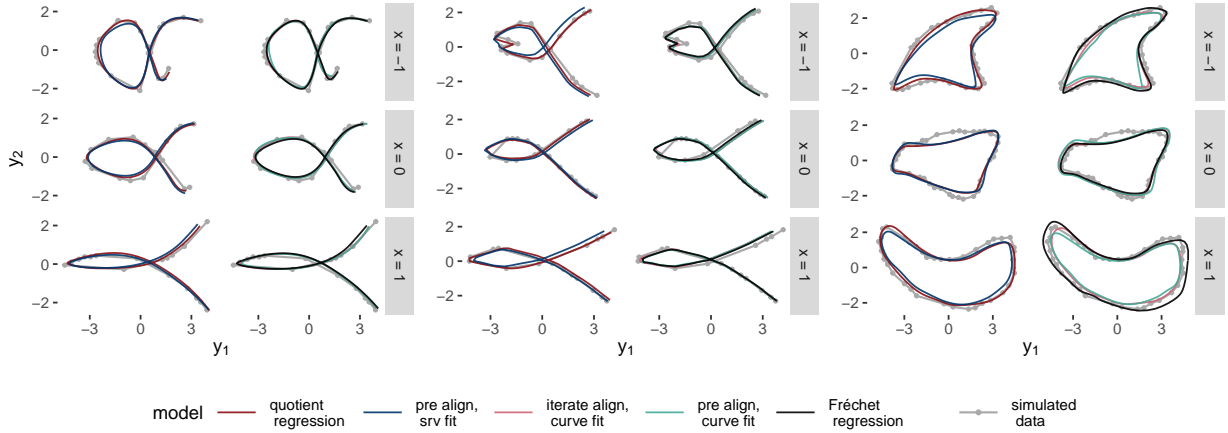


Figure 1: Predictions for $x = -1, 0, 1$ for one typical selected run of simulation 1 (left), simulation 5 (middle) and simulation 11 (right). The predictions for all x values of the same runs can be found in the appendix (Fig. 6, Fig. 7 and Fig. 8, respectively).

walk error with standard deviation sd directly to the κ_i selected points, and not to the observed SRV vectors. This leads to observed curves that suit a curve-level functional model better than an SRV-level model, both in terms of their relationship with the covariate and in terms of error structure. We choose this setting, which neither fits well with the quotient linear model nor with Fréchet regression, to demonstrate the robustness of our method and to validate the adapted algorithm for closed curves (Algorithm 2 in the Appendix). For the quotient linear model and the pre-align, SRV fit procedure we use closed linear splines with 21 knots on SRV-level and the procedure for closing the splines described in Subsection 2.3. For the procedures with fit on curve level we use quadratic closed splines with 21 knots and for the Fréchet regression model we use linear SRV splines with 21 knots and the algorithm for estimating closed mean curves of Steyer et al. (2022) adapted for weighted mean estimation (Algorithm 3 in the Appendix). See Fig. 1, right, and Fig. 8 in the appendix for an example of simulation 11. Even in this unfavorable setting the quotient linear model performs best in three out of four simulations. Only in the case of $\kappa_i \in [30, 40]$ points per curve and $sd = 0.2$, the procedure where we iterate between alignment and curve-level fit performs slightly better in terms of the MSE (Tab. 1). This can be explained by the fact that in this case the points are observed relatively densely and therefore errors at the curve level cause large errors at the SRV level (since we calculate the derivative via finite differences).

Visually, the quotient regression model and iterating between alignment and curve-level fit gives satisfying results, while if we fit a model after pre-alignment, the predicted curves for $x = -1$ and $x = 1$ appear too small (see Fig. 1, right). This can again be explained by the fact that alignment to the mean does not automatically result in good alignment among the curves. This is similarly problematic for Fréchet regression. Here, the prediction for $x = 1$ appears too large and the prediction for $x = -1$ is a bit too bulky on the left.

Overall, in the 12 simulations, the quotient regression model performed best in terms of the MSE among the five estimation methods considered, followed by Fréchet regression and the alternation between curve level fitting and alignment. Also, the average time required for one computation is relatively small for the quotient regression model compared to the other methods, especially for the more complex simulations 5 to 12 (Tab. 1), while it naturally takes somewhat longer than methods with pre-alignment only in most scenarios. The increased run times for methods with pre-alignment in scenarios with closed curves (simulation 9-12) stem from the fact that they involve unconditional elastic mean computation explicitly optimizing for closed mean curves (Steyer et al., 2022) whereas quotient regression utilizes the simplified approach described in Section 2.3. In part, this also explains the long run times of Fréchet regression in these scenarios, building on an adapted version of this unconditional elastic mean computation. Fréchet regression, however, also generally takes longer than the other methods, since optimization must be performed for

each value of \mathbf{x} separately; here for our small $n = 11$ and single covariate we used all observed covariates $x = -1, -0.8, \dots, 0.8, 1$. This means that for this model the computation time increases not only with the number n of observed curves but also with the number of predictions for covariate combinations desired.

5.2 Inference based on spline coefficients

Another advantage of the quotient regression model over Fréchet regression is that it yields parameter estimates for each covariate. These are useful not only for interpretation but also for model inference. In this simulation, we investigate and validate the bootstrap based tests described in Section 4 for the slope parameters of the quotient regression model. In particular, we focus on the more difficult case of the test based on the associated spline coefficients, which additionally allows to investigate local properties of the slope parameters.

For this purpose, we generate SRV curves as linear splines with 6 equidistant knots as a function of two covariates. To see how the test behaves with stronger and weaker effects, a strong effect $\tilde{\beta}_1$ is used for the association with x_1 and a weaker, local effect $\tilde{\beta}_2$ is assumed for the relation with x_2 , with $\tilde{\beta}_2(t) = 0$ for $t \geq 0.4$ (cf. Fig. 2). In addition, we assume that there is a third covariate x_3 , which is independent of the observed curves.

For the simulation, we first draw samples of sample size $n \in \{10, 30, 60\}$ of the covariates $x_{ji} \sim \text{Unif}(-1, 1)$ for $j = 1, 2, 3$ and $i = 1, \dots, n$. Then, similar as in the previous subsection, for $i = 1, \dots, n$ we randomly select 10 to 15 points on the curve $Q^{-1}(\tilde{\beta}_0 + \tilde{\beta}_1 x_{1i} + \tilde{\beta}_2 x_{2i})$. Note, that this procedure will not generate observations from the quotient regression model with parameters $\tilde{\beta}_0, \tilde{\beta}_1$ and $\tilde{\beta}_2$, as the sampling of the points on the curve generates a not further defined error in the quotient space. Since the quotient regression model is defined only as a minimization problem and there is no generating probability distribution available to sample curves from this model for given model parameters but we have to rely on the described auxiliary sampling scheme. In general, this implies that the model will be misspecified, i.e. $\mathcal{E}([Y]|x_1, x_2, x_3) \neq [Q^{-1}(\tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2)]$ for the quotient regression model and the data generated as described

above. As a consequence, if we estimate $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) = \underset{(\beta_0, \beta_1, \beta_2, \beta_3) \in \mathbb{L}_2}{\operatorname{argmin}} \sum_{i=1}^n d(\mathbf{y}_i, \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i})^2$

with respect to the elastic distance, the parameters $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$ will be different to $\tilde{\beta}_0, \tilde{\beta}_1$ and $\tilde{\beta}_2$ even if $n \rightarrow \infty$. However, $\hat{\beta}_3 \rightarrow \mathbf{0}$ if $n \rightarrow \infty$ holds, since x_3 and the curves are assumed to be independent. For the test of the slope parameters, this means that rejections of $H_{01} : \beta_1 = 0$ and $H_{02} : \beta_2 = 0$ correspond to the test's power, while those of $H_{03} : \beta_3 = \mathbf{0}$ should keep the type one error rate here specified as $\alpha = 0.05$. Looking at the tests for the individual spline coefficients $\xi_{2,m}$, $m = 1, \dots, 6$ of β_2 , we expect the null hypotheses $\xi_{2,1} = \mathbf{0}$ and $\xi_{2,2} = \mathbf{0}$ to be rejected, but because of the above argument, the other spline coefficients are not guaranteed to be zero.

To obtain an estimate of the rejection probability for the tests of the coefficients being zero given the sample size $n \in \{10, 30, 60\}$ and the number of bootstrap repetitions $N_{boot} \in \{100, 500, 1000\}$, we draw 1000 times a sample consisting of curves $\mathbf{y}_1, \dots, \mathbf{y}_n$ with covariates x_{1i}, x_{2i}, x_{3i} , $i = 1, \dots, n$ as described above. Next, we draw bootstrap replicates $\mathbf{y}_1^{(b)}, \dots, \mathbf{y}_n^{(b)}$, $b = 1, \dots, N_{boot}$, from the sample and reject the null hypothesis $H_{0j} : \beta_j = \mathbf{0}$ if $\bar{\xi}_{j,m}^T \hat{\Sigma}_{j,m}^{-1} \bar{\xi}_{j,m} \geq c_{j,1-\frac{\alpha}{M}}$, where $c_{j,1-\frac{\alpha}{M}}$ is the $1 - \frac{\alpha}{M}$ percentile, for any of the spline coefficients $\xi_{j,m}$, $m = 1, \dots, M = 6$ of β_j (as described in more detail in Section 4). The estimated rejection probability (Tab. 2) then is the relative proportion of the 1000 repetitions in which the null hypothesis is rejected.

For the data constellation described above, table 2 indicates that the rejection probability of H_{03} keeps the α level of 5% if the number of bootstrap replications is sufficiently large, i.e. at least about 500-1000. In this setup, the weak effect β_2 is found to be significant in 67% and 97% of the cases and the strong effect β_1 even in 100% of the cases for $n = 30$ and 60, respectively. To see if the distinction of zero and non-zero effects is also possible for parts of the curves, we consider in Fig. 2 (right) the rejection probabilities for the tests of the individual spline coefficients.

The plot indicates that the null hypotheses for coefficients of β_1 are mostly rejected (rejection probabilities 0.15, 1.00, 1.00, 1.00, 0.95, and 1.00 for coefficients 1-6, respectively) while for β_2 the coefficients changing the lower part of the

n	N_{boot}	$H_{01} : \beta_1 = \mathbf{0}$	$H_{02} : \beta_2 = \mathbf{0}$	$H_{03} : \beta_3 = \mathbf{0}$
10	100	0.66	0.07	0.02
10	500	0.42	0.01	0.00
10	1000	0.38	0.01	0.00
30	100	1.00	0.76	0.12
30	500	1.00	0.67	0.06
30	1000	1.00	0.67	0.05
60	100	1.00	0.97	0.10
60	500	1.00	0.96	0.05
60	1000	1.00	0.96	0.04

Table 2: Estimated rejection probability of the null hypothesis $H_{0j} : \beta_j = \mathbf{0}$, $j = 1, 2, 3$, for a sample of size n and B Bootstrap replications.

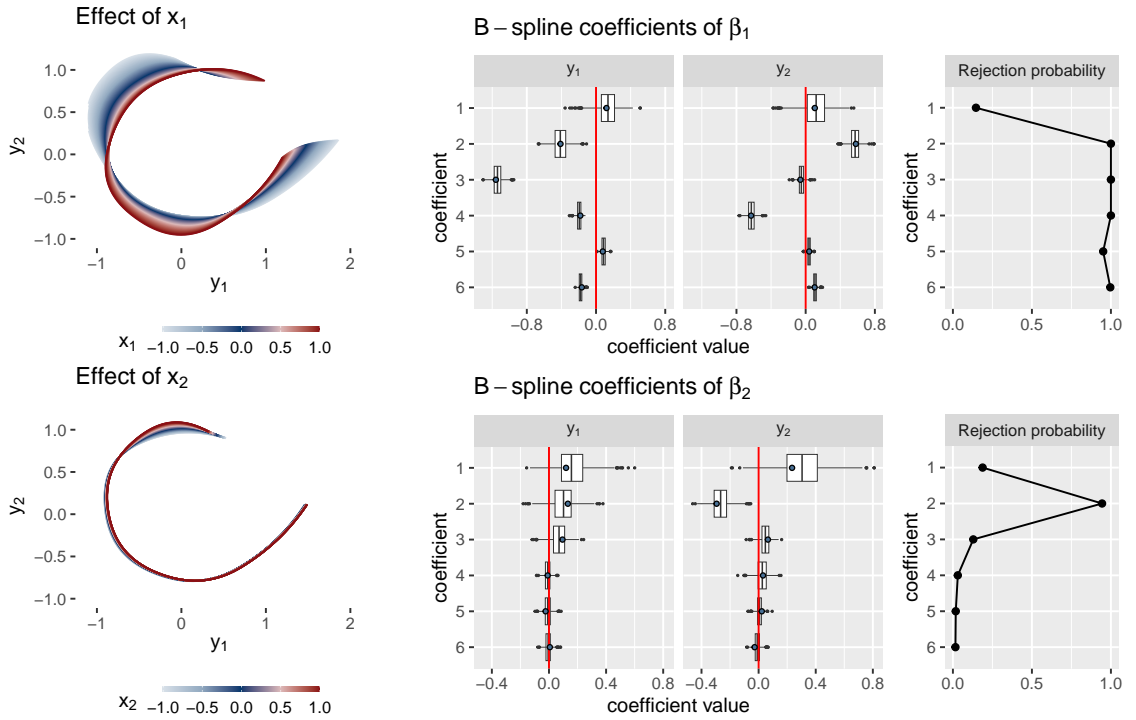


Figure 2: Left: Effect of x_1 given $x_2 = 0$ (top) and of x_2 given $x_1 = 0$ (bottom) estimated on a large simulated sample ($n = 500$) as an approximation of the closest model to the true model in our model space. Middle: Distribution of bootstrapped coefficients $\xi_{j,m}^{(b)}$, $m = 1, \dots, 6$, $j = 1$ (top) and $j = 2$ (bottom) over the 1000 repetitions for the setting with $n = 60$ and $N_{boot} = 1000$; the coefficients of the effect estimated on the large sample are displayed as blue dots. Coefficients $\xi_{j,1}, \dots, \xi_{j,6}$ correspond to regions on the curves from the top anti-clockwise to the right. Right: Rejection probabilities for the tests of the individual spline coefficients.

curves, ξ_4 , ξ_5 and ξ_6 , are mostly not rejected (rejection probabilities 0.19, 0.94, 0.13, 0.03, 0.02 and 0.02 for coefficients 1-6, respectively), which is consistent with the absent visible effect (Fig. 2, left). The distribution of the bootstrapped coefficients $\xi_{j,m}^{(b)}$ (Fig. 2, middle) essentially scatters around the estimated optimal parameters for large sample size ($n = 500$), which we use as an approximation of the closest model to the true model in our model space. This indicates good identifiability of the regression model via its spline coefficients.

We further repeat the simulation for $n = 60$ and $B = 1000$ with 11 instead of 6 knots still using linear SRV splines for data generation and for modeling to check that the tests are also valid for more complex curves with more spline coefficients. Here we observe a rejection probability of 100% for H_{01} and H_{02} and of 6% for H_{03} . The high rejection

probability also for the smaller effect β_2 probably results from the fact that we did not succeed in simulating a local effect (see Fig. 9 in the appendix). However, the coefficients appear to be well identified here as well. To keep the significance level of $\alpha = 0.05$ for H_{03} exactly, more observations would be necessary for this larger number of spline coefficients.

Overall, we conclude from this simulation that it is possible to test the significance of the slope parameters for the quotient space regression model using the corresponding spline coefficients. As this requires that the model is well identified by the spline coefficients, the sample size should be large enough to ensure that the spline coefficients of the different bootstrap model estimates represent equivalent parts of the curves. In particular, for a larger number of spline coefficients allowing flexible modeling of curves, a relatively large number of observations is needed to a) maintain the α level in the bootstrap setting and b) obtain reasonable power. We also discuss the choice of test based or not based on spline coefficients in the context of the application in the next section.

6 Investigating the effect of age and Alzheimer’s disease on hippocampus outlines via elastic regression

Hippocampal volume loss is associated with both Alzheimer’s disease and normal aging (Henneman et al., 2009). Moreover, Frisoni et al. (2008) showed that these covariates affect the hippocampal surface locally using parametric surface mesh models. The surface mesh model, however, depends on a meaningful parametrization of the shape. In contrast, we investigate local effects on the hippocampal volume by modeling the shape of the hippocampus. However, it’s essential to note that when we refer to ‘shape’ we do not mean the classical shape spaces that consider point configurations modulo rotation and scaling. Instead, our approach focuses on curves modulo reparametrization and translation, using the two-dimensional outlines in a quotient linear model that defines an elastic model of the outlines modulo re-parametrization without dependence on any chosen parametrization.

6.1 Data acquisition and preparation

Data used in the preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer’s disease (AD). In addition to the MRI images, ADNI also provides semi-automated segmentations of the hippocampus created using a high-dimensional brain mapping program SNT, which was commercially available from Medtronic Surgical Navigation Technologies (Louisville, CO). For more details on this procedure and a comparison with manual segmentation of the hippocampal volumes, see Hsu et al. (2002). For our analysis, we use all available hippocampal masks of the 101 Alzheimer’s disease (AD) patients and 138 controls (CN) obtained from the MRI images of the first scanning session. To apply our quotient regression model to the hippocampus data, we need to extract two-dimensional outlines (Fig. 3, right) from the three-dimensional hippocampal masks (Fig. 3, left). To do this, we perform the following steps for the left and right hippocampus separately. First, each hippocampus is rotated around the left-right axis using principal component analysis so that its first principal component lies in the horizontal plane. Then we project the data onto the horizontal plane and use the function `ocontour` from the R-package “EBImage” (Pau et al., 2010) to extract a closed outline curve. After alignment to the overall mean, the outlines of the hippocampus are sliced at the tail in the same location to obtain meaningful open curves, since the hippocampus merges into the fornix at the tail, i.e. it is not anatomically closed. In the last step the number of points per curve is reduced to improve the computational efficiency of model estimation, via keeping only points whose time stamps are at least 0.015 apart after alignment to the overall mean.

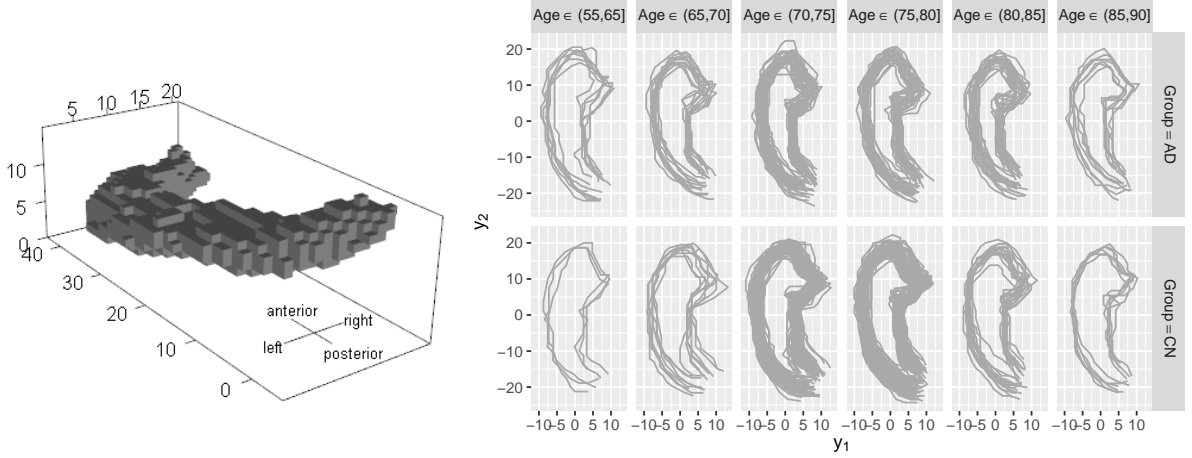


Figure 3: Left: Three-dimensional left hippocampal mask for one person. Right: Open outlines of the left hippocampus for different age groups separately for Alzheimer’s disease (AD) patients and control group (CN).

The result of this preprocessing are left and right hippocampal outlines of 101 Alzheimer’s disease (AD) patients and 138 control subjects (CN) with 33 to 48 points per curve. The explanatory variables considered are Age, Group = AD, CN and Sex = M, F of the subjects. These covariates are roughly balanced, the mean age for AD and CN is 76 years, ranging from 57 to 89 and 62 to 90 years, respectively, and about 49% of the subjects in both groups are female. Visual inspection of the outlines, taking into account the covariates Age and Group (Fig. 3, right), reveals no clear relationship with the shape of the hippocampus. This might be due to the large overall variance of the outlines.

6.2 Regression analysis of hippocampal shapes

To see how age, Alzheimer’s disease and the sex of a subject influence the shape of the hippocampus, we model the hippocampus outlines using the quotient linear model for elastic curves (Def. 2.16). Precisely, we assume

$$\mathcal{E}([Y]|x_{\text{Age}}, x_{\text{Group}}, x_{\text{Sex}}) = [Q^{-1}(\beta_0 + \beta_{\text{Age}}x_{\text{Age}} + \beta_{\text{Group}}x_{\text{Group}} + \beta_{\text{Sex}}x_{\text{Sex}})],$$

where the conditional Fréchet mean $\mathcal{E}([Y]|x_{\text{Age}}, x_{\text{Group}}, x_{\text{Sex}})$ of the hippocampal outlines is defined with respect to the elastic distance on the product space of elastic curves for the left and right hippocampus. That is $d(\mathbf{y}_1, \mathbf{y}_2) = \sqrt{d_{\text{left}}(\mathbf{y}_{1,\text{left}}, \mathbf{y}_{2,\text{left}})^2 + d_{\text{right}}(\mathbf{y}_{1,\text{right}}, \mathbf{y}_{2,\text{right}})^2}$ with $\mathbf{y}_i = (\mathbf{y}_{i,\text{left}}, \mathbf{y}_{i,\text{right}})$, $i = 1, 2$, where d_{left} and d_{right} are the separate elastic distances for the left and the right hippocampal curves, respectively. With this product space distance, the optimization problem defining the metric regression model $\operatorname{argmin}_{f \in \mathcal{F}} \sum_{i=1}^n d(\mathbf{y}_i, f(\mathbf{x}_i))^2$ becomes $\operatorname{argmin}_{f \in \mathcal{F}} \sum_{i=1}^n d_{\text{left}}(\mathbf{y}_{i,\text{left}}, f_{\text{left}}(\mathbf{x}_i))^2 + \sum_{i=1}^n d_{\text{right}}(\mathbf{y}_{i,\text{right}}, f_{\text{right}}(\mathbf{x}_i))^2$ with $f = (f_{\text{left}}, f_{\text{right}})$ and therefore can be solved separately for the left and right hippocampal shapes.

The parameters $\beta_0, \beta_{\text{Age}}, \beta_{\text{Group}}, \beta_{\text{Sex}} \in \mathbb{L}_2$ of this intrinsic metric regression model are estimated using linear spline functions with 21 equidistant knots for the left and the right hippocampus each. Since this leads to piecewise linear predictions on SRV level, the predicted outlines are smooth curves (Fig. 4). Linear effects on SRV level are visualized on curve level by varying one covariate at a time to illustrate effect directions via corresponding predictions.

As expected, we observe similar effects for left and right hippocampus and the hippocampal volume decreases with age and for AD patients. Moreover, the Sex effect appears to be small compared to Age and Group effect. Since age and Alzheimer’s disease appear at first glance to have a comparable effect on the hippocampus, the question arises to

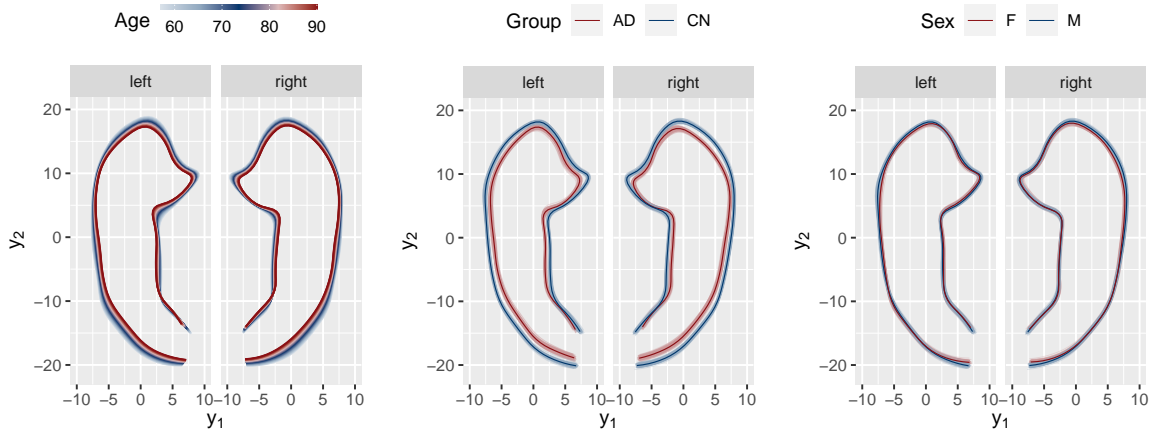


Figure 4: Effects displayed via model predictions with one varying covariate at a time. As a common reference, the remaining covariates are set to Age = mean(Age) = 76, Group = CN, and Sex = M. For the binary covariates Group and Sex, bootstrap predictions (lighter color) are added to the model predictions.

what extent the covariates Age and Group affect the shape of the hippocampus differently. To answer this, we use the linear structure underlying the quotient regression model and project $\hat{\beta}_{\text{Group}}$ onto $\hat{\beta}_{\text{Age}}$. The scalar projection of $\hat{\beta}_{\text{Group}}$ onto $\hat{\beta}_{\text{Age}}$ is 12.8 years, which means that having Alzheimer’s diseases shrinks the hippocampus about as much as 12.8 years of aging would do, but the angle between $\hat{\beta}_{\text{Group}}$ and $\hat{\beta}_{\text{Age}}$ is 47 degree, which means only about half of the Group effect shows in the same direction as the Age effect does. To visualize the remaining effect, we plot the prediction for a subject with Alzheimer’s disease alongside the prediction for a model where the Group effect is replaced by its linear projection on the Age effect (Fig. 5, left). This allows us to see which parts of the hippocampus are effected differently. While both age and Alzheimer’s disease reduce the volume at the ① label in Fig. 5 (left), the width of the hippocampal head, i.e., the distance between ② and ④, appears to be reduced substantially more by AD than by normal aging. In contrast, the distance between ① and ③ appears to be similarly affected by both covariates, although for the right hippocampus this distance might be smaller for AD patients compared to someone in the control group who is 12.8 years older.

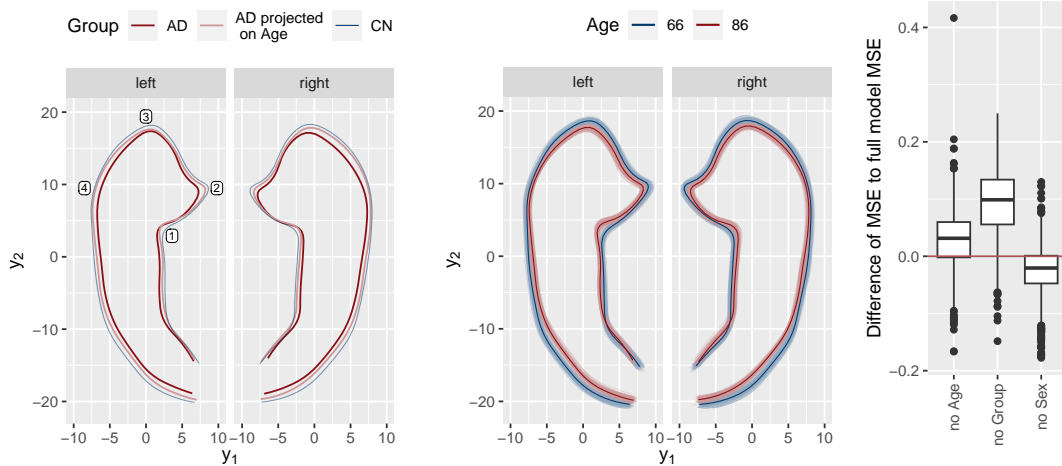


Figure 5: All non mentioned covariates are fixed to Age = mean(Age) = 76, Group = CN, and Sex = M. Left: Prediction for AD compared to the prediction for a model where the Group effect is replaced by its linear projection onto the Age effect. Labels indicate prominent features. Middle: Bootstrap predictions for Age effect. Right: Difference of the MSE of the models with an omitted covariate and the MSE of the full model computed on the out-of-bootstrap sample.

6.3 Model inference for the hippocampus regression model

To assess the significance of the estimated effects, we apply the model inference and selection tools described in Section 4 to this regression model for the hippocampal shapes. First, we test the global null hypothesis H_0 that none of the covariates Age, Group, and Sex has an effect by using the Fréchet coefficient of determination \tilde{R}^2 as a test statistic and approximating its distribution under H_0 through permutation sampling (number of samples = 500). We observe $\tilde{R}^2 = 0.033$, which is significantly different from the estimated mean under H_0 ($\overline{\tilde{R}^2} = 0.016$, p-value of the one-sided t-test $< 2.2 \times 10^{-16}$). Next we compare the adjusted coefficients of determination \tilde{R}_{adj}^2 for all possible sub-models to decide which covariates improve the model fit. Here, the largest value is obtained for the full model ($\tilde{R}_{\text{adj}}^2 = 0.021$, see Tab. 3), while all reported values for \tilde{R}^2 and \tilde{R}_{adj}^2 are relatively small. Thus, although explaining only a small proportion of the total variation, with a larger remaining fraction corresponding to individual variation among individuals, all variables considered can be deemed relevant based on this criterion.

To account for the variability of the model predictions, we draw 1000 bootstrap samples and estimate the model parameters on each. The distribution of the bootstrapped spline coefficients is shown in Fig. 10 and 11 in the appendix. The large variation in some of the bootstrap coefficients, much larger than for the bootstrap predictions in Fig. 4 and Fig. 5, middle, indicates that a test based on the coefficients might lose power due to warping variability, as discussed for a setting with many spline coefficients (here 42 2-dimensional coefficients per covariate for left and right in total) in Section 4.4. Therefore, we construct confidence regions directly for the predictions. To this end, for each prediction, we compute the bootstrap predictions for the same combinations of covariates as well as their distance from the original model prediction. We then construct the simultaneous 95% confidence region based on the closest 950 bootstrap predictions. The result for the binary effects Group and Sex is shown in Fig. 4. For the continuous covariate Age, we show the bootstrap predictions with confidence regions for Age = 66 and Age = 86 in Fig. 5, middle. Note that the resulting confidence regions not only include the variance of the effects, but also that of the intercept. It can be seen that the confidence regions constructed in this way clearly separate the Alzheimer’s group from the control group, whereas the confidence regions for Sex overlap in all parts of the hippocampus. For Age, there are parts (especially at the head of the hippocampus) where the regions are separated, as well as parts where no clear separation is found. This is consistent with the previously obtained results that the group effect is the most pronounced and the Sex effect is the least evident.

In addition, we evaluate the importance of the included covariates by comparing how much the model estimation error increases when we remove single covariates from the model. Here we look at the difference of the mean squared error (MSE) of the model with an omitted covariate and the MSE for the full model (Fig. 5, right), where we estimate the MSE for each bootstrap model on the out-of-bootstrap sample. That is, we evaluate the model prediction on the data that were not used for model estimation on the bootstrap sample (about 36.8% of the data). On average, we compute an out-of-bootstrap MSE of 13.34 for the full model and 13.37, 13.43 or 13.31 for a model without Age, Group or Sex effect, respectively. Omitting one effect increases the MSE for 740, 943 and 263 out of 1000 bootstrap samples for the Age, Group and Sex effect, respectively. Thus, based on these two variables improving the out-of-bootstrap prediction error in most of the samples, we would choose the model including Age and Group but no Sex effect.

6.4 Discussion of the hippocampus application

Overall, Alzheimer’s disease has the largest and most stable effect on the hippocampus among the covariates considered. The direction of this effect, i.e., the way the hippocampus shrinks, differs from normal aging. Although females appear to have slightly smaller hippocampi, this Sex effect is not clearly significant in our model. In further studies, it would be interesting to include in the analysis the mild cognitive impairment (MCI) group, for whom hippocampal masks are also available from ADNI. Since this group is known to be heterogeneous, with likely unknown subgroups, we did not include them in this study. In addition, it may be worthwhile to explore more complex model equations, for example,

including further covariates or interaction effects (e.g. of Age and Sex) and to examine if and how the effects differ between the left and right hippocampus.

7 Conclusion and Outlook

In this work, we developed an elastic regression approach, which we motivate as a special case of a general type of regression models we call quotient linear regression, for curves with respect to the elastic distance as response values and multiple scalar covariates. It allows modeling curves in two or more dimensions, e.g. outlines of anatomical features, while being invariant to their parameterization. Using the projections of affine linear functions as the model space for the SRV transformed curves permits iterative estimation of the model by alternating between alignment and estimation of a functional linear model. We provide an implementation of this algorithm in the R-package `elasdics` (Steyer, 2022).

To deal with sparsely and/or irregularly observed curves, we use splines to model the SRV curves. Since certain of these splines are identifiable modulo parameterization, inference based on the estimated spline coefficients is possible in these cases and also allows to investigate local effects. Here, however, as with our proposed inference methods based on distances and/or on the predicted curves, we rely on re-sampling methods such as bootstrap and permutation re-sampling. Further research will be needed to develop tests and confidence sets with formal guarantees.

Placing the proposed elastic regression model into the more general context of quotient (linear) regression, allows us to point out direct connections to similar approaches on other quotient spaces in literature and to present results on properties of the model space, consistency and existence of Fréchet mean estimation in a higher level of generality. Moreover, we also pave the way to for quotient regression beyond linear model spaces:

Using affine linear functions as underlying model space includes constant speed geodesics in the quotient space of curves modulo re-parameterization, but is somewhat more flexible. Using this larger space not only enables the estimation strategy described above, but we have also shown through examples (in the simulations in Section 5) that geodesic regression lines alone are not sufficient to model all changes to curves that naturally appear in practice, and a larger model space thus is beneficial. However, our proposed model space is still linear in the SRV space, which may be too restrictive for some real data applications. To allow more flexible smooth dependence of curves on covariates, quotient models could be extended to an additive linear regression model.

Another appealing direction for further research is to develop the quotient regression model for elastic shapes, i.e. curves modulo translation, scaling, rotation and parametrization, which goes beyond quotient spaces of a Hilbert space. Along the lines of Section 2.1.4, this space can be seen as the quotient of the sphere, which is a submanifold of \mathbb{L}_2 , on which the product of the rotation and re-parametrization groups acts by isometries. Such a model could be implemented building on 'generalized linear' regression models on the sphere as suggested in Stöcker et al. (2023) for planar (in-)elastic shapes and forms.

Acknowledgements

We gratefully acknowledge funding by grants GR 3793/3-1 'Flexible regression methods for curve and shape data' and GR 3793/5-1 'Combining geometry-aware statistical and deep learning for neuroimaging data' (project P7 in the research unit FOR 5363) from the German research foundation (DFG). We would like to thank Kerstin Ritter (Charité Berlin) for sharing her domain knowledge on Alzheimer's disease and the ADNI data.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association;

Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

- Martins Bruveris. Optimal reparametrizations in the square root velocity framework. *SIAM Journal on Mathematical Analysis*, 48(6):4335–4354, 2016.
- Dmitri Burago, Yuri Burago, and Sergei Ivanov. A course in metric geometry. *Graduate Studies in Math.*, 33, 01 2001.
- John Charles Burkill and Harry Burkill. *A second course in mathematical analysis*. Cambridge University Press, 1970.
- Anna Calissano, Aasa Feragen, and Simone Vantini. Graph-valued regression: Prediction of unlabelled networks in a non-euclidean graph space. *Journal of Multivariate Analysis*, 190:104950, 2022. ISSN 0047-259X. doi: <https://doi.org/10.1016/j.jmva.2022.104950>. URL <https://www.sciencedirect.com/science/article/pii/S0047259X22000021>.
- Anna Calissano, Aasa Feragen, and Simone Vantini. Populations of Unlabelled Networks: Graph Space Geometry and Generalized Geodesic Principal Components. *Biometrika*, 04 2023. ISSN 1464-3510. doi: 10.1093/biomet/asad024. URL <https://doi.org/10.1093/biomet/asad024>. asad024.
- Yaqing Chen, Alvaro Gajardo, Jianing Fan, Qixian Zhong, Paromita Dubey, Kyunghye Han, Satarupa Bhattacharjee, and Hans-Georg Müller. *frechet: Statistical Analysis for Random Objects and Non-Euclidean Data*, 2020. URL <https://CRAN.R-project.org/package=frechet>. R package version 0.2.0.
- Emil Cornea, Hongtu Zhu, Peter Kim, Joseph G. Ibrahim, and the Alzheimer’s Disease Neuroimaging Initiative. Regression models on Riemannian symmetric spaces. *Journal of the Royal Statistical Society: Series B*, 79(2):463–482, 2017.
- Ian L Dryden and Kanti V Mardia. *Statistical shape analysis: with applications in R*, volume 995. John Wiley & Sons, 2016a.
- I.L. Dryden and K.V. Mardia. *Statistical Shape Analysis: With Applications in R*. Wiley Series in Probability and Statistics. Wiley, 2016b. ISBN 9780470699621. URL <https://books.google.de/books?id=jGstCwAAQBAJ>.
- H. Edelsbrunner, D. Kirkpatrick, and R. Seidel. On the shape of a set of points in the plane. *IEEE Transactions on Information Theory*, 29(4):551–559, 1983. doi: 10.1109/TIT.1983.1056714.
- P Thomas Fletcher. Geodesic regression and the theory of least squares on Riemannian manifolds. *International journal of computer vision*, 105(2):171–185, 2013.
- Giovanni B Frisoni, Rossana Ganzola, Elisa Canu, Udo Rüb, Francesca B Pizzini, Franco Alessandrini, Giada Zoccatelli, Alberto Beltramello, Carlo Caltagirone, and Paul M Thompson. Mapping local hippocampal changes in alzheimer’s disease and normal ageing with mri at 3 tesla. *Brain*, 131(12):3266–3276, 2008.
- Mengmeng Guo, Jingyong Su, Li Sun, and Guofeng Cao. Statistical regression analysis of functional and shape data. *Journal of Applied Statistics*, 47(1):28–44, 2020. doi: 10.1080/02664763.2019.1669541. URL <https://doi.org/10.1080/02664763.2019.1669541>.
- Xiaoyang Guo, Aditi Basu Bal, Tom Needham, and Anuj Srivastava. Statistical shape analysis of brain arterial networks (BAN). *The Annals of Applied Statistics*, 16(2):1130 – 1150, 2022. doi: 10.1214/21-AOAS1536. URL <https://doi.org/10.1214/21-AOAS1536>.
- WJP Henneman, JD Sluimer, J Barnes, WM Van Der Flier, IC Sluimer, NC Fox, Ph Scheltens, H Vrenken, and F Barkhof. Hippocampal atrophy rates in alzheimer disease: added value over whole brain volume measures. *Neurology*, 72(11):999–1007, 2009.
- Yi Hong, Roland Kwitt, Nikhil Singh, Nuno Vasconcelos, and Marc Niethammer. Parametric regression on the grassmannian. *IEEE transactions on pattern analysis and machine intelligence*, 38(11):2284–2297, 2016.
- Yuan-Yu Hsu, Norbert Schuff, An-Tao Du, Kevin Mark, Xiaoping Zhu, Dawn Hardin, and Michael W Weiner. Comparison of automated and manual mri volumetry of hippocampus in normal aging and dementia. *Journal of Magnetic*

- Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 16(3): 305–310, 2002.
- 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.
- Stephan F. Huckemann. Intrinsic inference on the mean geodesic of planar shapes and tree discrimination by leaf growth. *The Annals of Statistics*, 39(2):1098–1124, 2011. ISSN 00905364, 21688966. URL <http://www.jstor.org/stable/29783668>.
- Shantanu H. Joshi, K. Narr, O. Phillips, K. Nuechterlein, R. Asarnow, A. Toga, and R. Woods. Statistical shape analysis of the corpus callosum in schizophrenia. *NeuroImage*, 64:547–559, 2013.
- Gregoire Pau, Florian Fuchs, Oleg Sklyar, Michael Boutros, and Wolfgang Huber. Ebimage—an r package for image processing with applications to cellular phenotypes. *Bioinformatics*, 26(7):979–981, 2010. doi: 10.1093/bioinformatics/btq046.
- Vera Pawlowsky-Glahn, Juan José Egozcue, and Raimon Tolosana-Delgado. *Modeling and analysis of compositional data*. John Wiley & Sons, 2015.
- Alexander Petersen and Hans-Georg Müller. Fréchet regression for random objects with euclidean predictors. *Annals of Statistics*, 47:691–719, 04 2019. doi: 10.1214/17-AOS1624.
- Ronald Carl Petersen, PS Aisen, Laurel A Beckett, MC Donohue, AC Gamst, Danielle J Harvey, CR Jack, WJ Jagust, LM Shaw, AW Toga, et al. Alzheimer’s disease neuroimaging initiative (adni): clinical characterization. *Neurology*, 74(3):201–209, 2010.
- J. O. Ramsay and C. J. Dalzell. Some tools for functional data analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(3):539–572, 1991. ISSN 00359246. URL <http://www.jstor.org/stable/2345586>.
- James O. Ramsay and Bernhard W. Silverman. *Functional Data Analysis*. Springer New York, 2005.
- Laura Sangalli, Piercesare Secchi, Simone Vantini, and Alessandro Veneziani. A case study in exploratory functional data analysis: Geometrical features of the internal carotid artery. *Journal of the American Statistical Association*, 104:37–48, 03 2009. doi: 10.1198/jasa.2009.0002.
- A. Srivastava and E.P. Klassen. *Functional and Shape Data Analysis*. Springer Series in Statistics. Springer New York, 2016. ISBN 9781493940202. URL <https://books.google.de/books?id=0cMwDQAAQBAJ>.
- Anuj Srivastava, Eric Klassen, Shantanu Joshi, and Ian Jermyn. Shape analysis of elastic curves in Euclidean spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(7):1415–1428, 09 2010. doi: 10.1109/TPAMI.2010.184.
- S.M. Srivastava. *A Course on Borel Sets*. Graduate Texts in Mathematics. Springer New York, 1998. ISBN 9780387984124. URL <https://books.google.de/books?id=FhYGYJtMwcUC>.
- Lisa Steyer. *elasdics: Elastic Analysis of Sparse, Dense and Irregular Curves*, 2022. URL <https://CRAN.R-project.org/package=elasdics>. R package version 1.1.1.
- Lisa Steyer, Almond Stöcker, and Sonja Greven. Elastic analysis of irregularly or sparsely sampled curves. *Biometrics*, 0:1–13, 2022.
- Almond Stöcker, Manuel Pfeuffer, Lisa Steyer, and Sonja Greven. Elastic full procrustes analysis of plane curves via hermitian covariance smoothing, 2022.
- Almond Stöcker, Lisa Steyer, and Sonja Greven. Functional additive models on manifolds of planar shapes and forms. *Journal of Computational and Graphical Statistics*, pages 1–24, 02 2023. doi: 10.1080/10618600.2023.2175687.
- James Derek Tucker, John R. Lewis, and Anuj Srivastava. Elastic functional principal component regression. *Stat. Anal. Data Min.*, 12:101–115, 2019.

- Karl Gerald van den Boogaart, Juan José Egozcue, and Vera Pawlowsky-Glahn. Bayes Hilbert spaces. *Australian & New Zealand Journal of Statistics*, 56(2):171–194, 2014.
- Jussi Väisälä. A proof of the mazur-ulam theorem. *The American Mathematical Monthly*, 110(7):633–635, 2003. ISSN 00029890, 19300972. URL <http://www.jstor.org/stable/3647749>.
- Hongtu Zhu, Yasheng Chen, Joseph G Ibrahim, Yimei Li, Colin Hall, and Weili Lin. Intrinsic regression models for positive-definite matrices with applications to diffusion tensor imaging. *Journal of the American Statistical Association*, 104(487):1203–1212, 2009.
- 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.

A Proofs and Computations

A.1 Proof of Lemma 2.2

Huckemann (2011) generalizes the notion of the Fréchet mean to the Fréchet ρ -mean:

Definition A.1 (Fréchet ρ -mean). *Let X, X_1, X_2, \dots be random elements mapping from a probability space $\Omega, \mathcal{A}, \mathcal{P}$ to a topological space Q . Let (P, d) be a topological space with distance d . For a continuous function $\rho : Q \times P \rightarrow [0, \infty]$ define the set of population Fréchet ρ -means of X in P by*

$$E^{(\rho)} = \operatorname{argmin}_{\mu \in P} \mathbb{E}(\rho(X, \mu)^2).$$

For $\omega \in \Omega$, denote by

$$E_n^{(\rho)}(\omega) = \operatorname{argmin}_{\mu \in P} \sum_{i=1}^n \rho(X_i(\omega), \mu)^2$$

the set of sample Fréchet ρ -means.

With this definition, the usual Fréchet mean is a special case where ρ is the distance function. Similarly as for Fréchet means, one can show that sample Fréchet ρ -means are consistent in the sense of Ziezold (1977), that is $\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} E_k^{(\rho)}(\omega) \subseteq E^{(\rho)}$ for almost all $\omega \in \Omega$.

Theorem A.2 (Huckemann (2011)). *Let $\rho : Q \times P \rightarrow [0, \infty[$ be a continuous function on the product of a topological space with a separable space with distance (P, d) . Then strong consistency holds in the Ziezold sense for the set of Fréchet ρ -means in P if:*

- (i) X has compact support, or if
- (ii) $\mathbb{E}(\rho(X, p)^2) < \infty$ for all $p \in P$ and ρ is uniformly continuous in the second argument.

Our statement on consistency is then a straightforward consequence using (ii), with $\mathcal{X} \times \mathcal{Y}$ taking the role of Q and $(\mathcal{F}, d_{\mathcal{F}})$ that of (P, d) .

Proof of Lemma 2.2. Define $\rho : \mathcal{X} \times \mathcal{Y} \times \mathcal{F} \rightarrow [0, \infty)$, $\rho(x, y, f) = d(y, f(x))$. This loss function ρ is continuous in the first two arguments since d as a metric and f are continuous. Furthermore, ρ is uniformly continuous in the last argument since for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$ and $f, \tilde{f} \in \mathcal{F}$ it holds that $d(y, f(x)) \leq d(y, \tilde{f}(x)) + d(\tilde{f}(x), f(x))$ via the triangle inequality and therefore because of symmetry: $|d(y, f(x)) - d(y, \tilde{f}(x))| \leq d(\tilde{f}(x), f(x)) \leq d_{\mathcal{F}}(f, \tilde{f})$. $C(\mathcal{X}, \mathcal{Y})$ is separable by the proof of Theorem 2.4.3 in Srivastava (1998) and therefore \mathcal{F} is separable as a subspace of a separable metric space. \square

A.2 Proof of Lemma 2.3 and Lemma 2.8

Proof. $C(\mathcal{X}, \mathcal{Y})$ is complete since \mathcal{Y} is complete (e.g., Burkill and Burkill, 1970, Theorem 3.45) and therefore \mathcal{F} in Lemma 2.3 and Φ in Lemma 2.8 complete as a closed subsets. Thus \mathcal{F} and Φ are compact since they are complete and totally bounded. Since $f \mapsto \mathbb{E}(d(Y, f(X)))$ and $\varphi \mapsto \mathbb{E}(d([Y], [\varphi(X)]))$ are continuous as a compositions of continuous functions, they attain their minimum on \mathcal{F} and Φ , respectively. \square

A.3 Proof of Lemma 2.5

Proof. i) Since \mathcal{Y} is separable, there is a countable, dense subset $\mathcal{Z} \subseteq \mathcal{Y}$. Let $[y] \in \mathcal{Y}/G$. Since \mathcal{Z} is dense in \mathcal{Y} , there is a sequence $(z_k)_{k \in \mathbb{N}} \subset \mathcal{Z}$ such that $\lim_{k \rightarrow \infty} z_k = y$. Therefore $\lim_{k \rightarrow \infty} d_G([z_k], [y]) \leq \lim_{k \rightarrow \infty} d(z_k, y) = 0$. Hence $\{[z] | z \in \mathcal{Z}\}$ is a countable, dense subset in \mathcal{Y}/G and \mathcal{Y}/G thus separable.

ii) Let $([y_k])_{k \in \mathbb{N}} \subset \mathcal{Y}/G$ be a Cauchy sequence. W.l.o.g. assume $d_G([y_k], [y_{k+1}]) < \frac{1}{2^k}$ for all $k \in \mathbb{N}$, otherwise consider a subsequence (and $([y_k])_{k \in \mathbb{N}}$ as a Cauchy sequence will converge to the same limit if it exists). We construct a sequence $(g_k)_{k \in \mathbb{N}} \subseteq G$ such that $(g_1 \circ \dots \circ g_{k-1} \circ y_k)_{k \in \mathbb{N}}$ is a Cauchy sequence in \mathcal{Y} . To do so set $g_1 = \text{id}$ and for $k \geq 2$ assume we already picked g_1, \dots, g_{k-1} . Then choose g_k such that

$$d(g_1 \circ \dots \circ g_{k-1} \circ y_k, g_1 \circ \dots \circ g_k \circ y_{k+1}) < \frac{1}{2^{k-1}}.$$

This is possible since g_1, \dots, g_{k-1} are isometries and therefore

$$\inf_{g \in G} d(g_1 \circ \dots \circ g_{k-1} \circ y_k, g_1 \circ \dots \circ g_{k-1} \circ g \circ y_{k+1}) = \inf_{g \in G} d(y_k, g \circ y_{k+1}) = d_G([y_k], [y_{k+1}]) < \frac{1}{2^k}.$$

Thus, $(g_1 \circ \dots \circ g_{k-1} \circ y_k)_{k \in \mathbb{N}}$ is a Cauchy sequence and converges to a $y \in \mathcal{Y}$, since \mathcal{Y} is complete. Hence

$$\begin{aligned} d_G([y_k], [y]) &= d_G([g_1 \circ \dots \circ g_{k-1} \circ y_k], [y]) \\ &= \inf_{g \in G} d(g_1 \circ \dots \circ g_{k-1} \circ y_k, g \circ y) \leq d(g_1 \circ \dots \circ g_{k-1} \circ y_k, y) \xrightarrow{k \rightarrow \infty} 0. \end{aligned}$$

As $([y_k])_{k \in \mathbb{N}}$ has a limit in \mathcal{Y}/G , \mathcal{Y}/G is complete. □

A.4 Proof of Theorem 2.13

Proof. We first show that Ψ is coercive, that is $\Psi(\varphi) \rightarrow \infty$ if $\|\varphi\|_{\Phi} \rightarrow \infty$. To do so note that

$$\begin{aligned} d_G([Y], [\varphi(X)]) &= \inf_{g \in G} \|Y - g \circ \varphi(X)\| \geq \|g \circ \varphi(X)\|_{\mathcal{Y}} - \|Y\|_{\mathcal{Y}} \\ &\geq \|\varphi(X)\|_{\mathcal{Y}} - C_1 \end{aligned}$$

for some $C_1 \in \mathbb{R}$, where the first inequality is due to the triangle inequality and the second due to the assumption that $[Y]$ is bounded. Therefore,

$$\Psi(\varphi) = \mathbb{E}(d_G([Y], [\varphi(X)])^2) \geq \mathbb{E}((\|\varphi(X)\|_{\mathcal{Y}} - C_1)^2) \geq (\mathbb{E}(\|\varphi(X)\|_{\mathcal{Y}}) - C_1)^2$$

due to Jensen's inequality since $x \mapsto (x - C_1)^2$ is convex. Note that $\mathbb{E}(\|\varphi(X)\|_{\mathcal{Y}})$ defines a norm on Φ since $\text{supp}(X) = \mathcal{X}$ and all $\varphi \in \Phi$ are continuous. Since all norms are equivalent on Φ (finite dimensional vector space) this means $\mathbb{E}(\|\varphi(X)\|_{\mathcal{Y}}) \rightarrow \infty$ if $\|\varphi\|_{\Phi} \rightarrow \infty$ and therefore $\Psi(\varphi) \rightarrow \infty$ if $\|\varphi\|_{\Phi} \rightarrow \infty$.

Since Ψ is continuous as a composition of continuous functions and coercive it attains its minimum. This is a standard argument that we repeat here for the sake of completeness. Pick a $\varphi_0 \in \Phi$. Since Ψ is coercive, there is a $C_2 \in \mathbb{R}$ such that $\Psi(\varphi) \geq \Psi(\varphi_0) + 1$ if $\|\varphi\|_{\infty} \geq C_2$. Hence $\inf_{\|\varphi\|_{\infty} \leq C_2} \Psi(\varphi) = \inf_{\varphi \in \Phi} \Psi(\varphi)$. Since $\{\varphi \in \Phi \mid \|\varphi\|_{\infty} \leq C_2\}$ is a closed and bounded subset of a finite dimensional vector space, it is compact (Heine-Borel) and therefore Ψ attains its minimum on $\{\varphi \in \Phi \mid \|\varphi\|_{\infty} \leq C_2\}$, which is also a global minimizer. □

A.5 Proof of Lemma 2.9

Proof. Since (\mathcal{Y}, d) is a length metric space and γ a shortest path, for the length $l(\gamma)$ of γ it holds by definition that

$$l(\gamma) = \sup_{a=t_0 < t_1 < \dots < t_n = b} \sum_{i=0}^{n-1} d(\gamma(t_i), \gamma(t_{i+1})) = d(y_1, \tilde{g} \circ y_2),$$

where $t_0, t_1, \dots, t_n, n \in \mathbb{N}$ is a partition of $[a, b]$. The length of $[\gamma]$ is bounded from above by the length of γ as

$$\begin{aligned} l([\gamma]) &= \sup_{a=t_0 < t_1 < \dots < t_n = b} \sum_{i=0}^{n-1} d_G([\gamma(t_i)], [\gamma(t_{i+1})]) \\ &= \sup_{a=t_0 < t_1 < \dots < t_n = b} \sum_{i=0}^{n-1} \inf_{g \in G} d(\gamma(t_i), g \circ \gamma(t_{i+1})) \\ &\stackrel{g = \text{id}}{\leq} \sup_{a=t_0 < t_1 < \dots < t_n = b} \sum_{i=0}^{n-1} d(\gamma(t_i), \gamma(t_{i+1})) = l(\gamma) \end{aligned}$$

for all partitions $a = t_0 < t_1 < \dots < t_n = b, n \in \mathbb{N}$. To see that $l([\gamma]) = l(\gamma)$ choose the trivial partition $a = t_0 < t_1 = b$ and observe

$$l([\gamma]) \geq d_G([\gamma(t_0)], [\gamma(t_1)]) = \inf_{g \in G} d(\gamma(a), g \circ \gamma(b)) = d(y_1, \tilde{g} \circ y_2) = l(\gamma). \quad (8)$$

Thus, $[\gamma]$ is a shortest path in \mathcal{Y}/G as $l([\gamma]) = d_G([\gamma_1], [\gamma_2])$ and $l([\tilde{\gamma}]) \geq d_G([\gamma_1], [\gamma_2])$ as in (8) for all other paths $\tilde{\gamma}$. \square

A.6 Proof of Corollary 2.12

This corollary is an immediate consequence of the following lemma.

Lemma A.3. *Let $(\mathcal{Y}, \langle \cdot, \cdot \rangle)$ be a real inner product space and G act on \mathcal{Y} by isometries with $g \circ 0 = 0$ for all $g \in G$. Let $y_1, y_2 \in \mathcal{Y}$ be aligned to $y_0 \in \mathcal{Y}$ and $\alpha_1, \alpha_2 \geq 0$. Then $\alpha_1 y_1 + \alpha_2 y_2$ is aligned to y_0 , which means the set of elements which are aligned to y_0 is a convex cone.*

Proof. Let $y \in \mathcal{Y}$ be aligned to $y_0 \in \mathcal{Y}$, that is $\|y_0 - y\| = \inf_{g \in G} \|y_0 - g \circ y\|$. This is equivalent with $\|y\|^2 = \langle y, y \rangle$ to

$$\begin{aligned} \|y_0 - y\|^2 &= \inf_{g \in G} \langle y_0 - g \circ y, y_0 - g \circ y \rangle \\ &= \|y_0\|^2 + \inf_{g \in G} \{ \langle g \circ y, g \circ y \rangle - 2 \langle y_0, g \circ y \rangle \} \\ &= \|y_0\|^2 + \|y\|^2 - 2 \sup_{g \in G} \langle y_0, g \circ y \rangle \quad (g \text{ isometry}) \\ &= \|y_0 - y\|^2 + 2 \langle y_0, y \rangle - 2 \sup_{g \in G} \langle y_0, g \circ y \rangle. \end{aligned}$$

Hence, $y \in \mathcal{Y}$ being aligned to $y_0 \in \mathcal{Y}$ is equivalent to $\sup_{g \in G} \langle y_0, g \circ y \rangle = \langle y_0, y \rangle$.

Let $g \in G$, thus g is a bijective isometry on a real vector space with $g \circ 0 = 0$, which means g is linear by the Mazur-Ulam theorem (Väisälä, 2003). Hence, for $y_1, y_2 \in \mathcal{Y}$ being aligned to $y_0 \in \mathcal{Y}$ and $\alpha_1, \alpha_2 > 0$ it holds that

$$\begin{aligned} \sup_{g \in G} \langle y_0, g \circ (\alpha_1 y_1 + \alpha_2 y_2) \rangle &= \sup_{g \in G} [\alpha_1 \langle y_0, g \circ y_1 \rangle + \alpha_2 \langle y_0, g \circ y_2 \rangle] \\ &\leq \alpha_1 \sup_{g \in G} \langle y_0, g \circ y_1 \rangle + \alpha_2 \sup_{g \in G} \langle y_0, g \circ y_2 \rangle \quad (\alpha_1, \alpha_2 \geq 0) \\ &= \alpha_1 \langle y_0, y_1 \rangle + \alpha_2 \langle y_0, y_2 \rangle \quad (y_1, y_2 \text{ aligned to } y_0) \\ &= \langle y_0, \alpha_1 y_1 + \alpha_2 y_2 \rangle \end{aligned}$$

On the other hand, we observe $\sup_{g \in G} \langle y_0, g \circ (\alpha_1 y_1 + \alpha_2 y_2) \rangle \geq \langle y_0, \alpha_1 y_1 + \alpha_2 y_2 \rangle$ if we take $g = \text{id}$ and therefore $\alpha_1 y_1 + \alpha_2 y_2$ is aligned to y_0 . \square

Proof of the Corollary. $\beta_0 + \sum_{j=1}^k \lambda_j \beta_j = \sum_{j=1}^k \lambda_j (y_0 + \beta_j)$ is aligned to β_0 according to Lemma A.3 and, thus, \tilde{f} is a shortest path in \mathcal{Y}/G as $x \mapsto \beta_0 + x \sum_{j=1}^k \lambda_j \beta_j$ is linear and therefore a shortest path in the inner product space \mathcal{Y} (Lemma 2.9). \square

A.7 Geodesics in \mathbb{L}_2/Γ cannot be modeled with splines

We construct a counterexample that shows that geodesics between two spline curves in \mathbb{L}_2/Γ do not necessarily lie in a spline space. Even though we only show this for a specific example, we expect this to be true for most spline curves and that the geodesic will only actually lie in a spline space in exceptional cases.

Consider the linear SRV spline curves $q_1(t) = \begin{pmatrix} 1 \\ 2t+1 \end{pmatrix}$ and $q_2(t) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, $t \in [0, 1]$. Since q_2 is constant, the optimal warping γ of q_2 to q_1 is given via

$$\dot{\gamma}(t) = \frac{\langle q_1(t), \begin{pmatrix} 0 \\ 1 \end{pmatrix} \rangle_+}{\int_0^1 \langle q_1(t), \begin{pmatrix} 0 \\ 1 \end{pmatrix} \rangle_+ dt} = \frac{2t+1}{\int_0^1 2t+1 dt} = t+0.5, \quad (9)$$

using the formula derived in the online supplement B.1 of Steyer et al. (2022). Here $\langle \cdot, \cdot \rangle_+$ denotes the positive part of the scalar product. Hence $q_2(\gamma(t))\sqrt{\dot{\gamma}(t)} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \sqrt{t+0.5}$ is optimally aligned to q_1 . This means the geodesic between $[q_1]$ and $[q_2]$ in \mathbb{L}_2/Γ is given by

$$\begin{aligned} \xi : [0, 1] &\rightarrow \mathbb{L}_2/\Gamma \\ x &\mapsto \left[(1-x) \begin{pmatrix} 1 \\ 2t+1 \end{pmatrix} + x \begin{pmatrix} 0 \\ 1 \end{pmatrix} \sqrt{t+0.5} \right]. \end{aligned}$$

Thus, ξ lies in a spline space only if $\xi(0.5)$ contains a spline, i.e there is a warping function $\tilde{\gamma} : [0, 1] \rightarrow [0, 1]$ such that

I. $0.5\sqrt{\tilde{\gamma}}$ and

II. $(\tilde{\gamma} + 0.5)\sqrt{\tilde{\gamma}} + 0.5\sqrt{(\tilde{\gamma} + 0.5)\tilde{\gamma}}$

are splines of some degree m . From I. we conclude that $\tilde{\gamma}$ is a spline of degree $2m+1$, $m \in \mathbb{N}_0$. But this means $(\tilde{\gamma}(t) + 0.5)\tilde{\gamma}(t)$ is a piecewise polynomial with degree $4m+1$, hence its square root $\sqrt{(\tilde{\gamma} + 0.5)\tilde{\gamma}(t)}$ cannot be piecewise polynomial. This contradicts the assumption that II. is a spline.

A.8 Additional algorithms

Algorithm 2: Quotient space regression for elastic closed curves

Input: data pairs $(\mathbf{x}_i, \check{\mathbf{q}}_i)$, $i = 1, \dots, n$ where $\check{\mathbf{q}}_i$ are the SRV transformations of observed polygons and

$\mathbf{x}_i = (x_{i,1}, \dots, x_{i,k})$, $i = 1, \dots, n$ are observed covariates; convergence tolerance $\epsilon > 0$

Compute initial estimate $\hat{\boldsymbol{\beta}}_{0,new}, \dots, \hat{\boldsymbol{\beta}}_{k,new} = \underset{\boldsymbol{\beta}_0, \dots, \boldsymbol{\beta}_k \in \mathbb{L}_2}{\operatorname{arginf}} \sum_{i=1}^n \|\boldsymbol{\beta}_0 + \sum_{j=1}^k \boldsymbol{\beta}_j x_{i,j} - \check{\mathbf{q}}_i\|_{\mathbb{L}_2}^2$;

Set $\hat{\boldsymbol{\beta}}_{j,old} = \hat{\boldsymbol{\beta}}_{j,new} \quad \forall j = 0, \dots, k$;

while $\max_{j=0, \dots, k} \|\hat{\boldsymbol{\beta}}_{j,old} - \hat{\boldsymbol{\beta}}_{j,new}\| > \epsilon$ **do**

$\hat{\boldsymbol{\beta}}_{j,old} = \hat{\boldsymbol{\beta}}_{j,new} \quad \forall j = 0, \dots, k$;

for $i \in 1, \dots, n$ **do**

$\mathbf{p}_i = \hat{\boldsymbol{\beta}}_{0,old} + \sum_{j=1}^k \hat{\boldsymbol{\beta}}_{j,old} x_{i,j}$; // compute predicted SRV curves

$\mathbf{v}_i(t) = \mathbf{p}_i \|\mathbf{p}_i\| - \int_0^1 \mathbf{p}_i(s) \|\mathbf{p}_i(s)\| ds$; // compute derivative of closed predicted curves

$\gamma_i = \underset{\gamma}{\operatorname{arginf}} \left\| \frac{\mathbf{v}_i}{\sqrt{\|\mathbf{v}_i\|}} - (\mathbf{q}_i \circ \gamma) \sqrt{\gamma} \right\|_{\mathbb{L}_2}^2$; // warping step

$\hat{\boldsymbol{\beta}}_{0,new}, \dots, \hat{\boldsymbol{\beta}}_{k,new} = \underset{\boldsymbol{\beta}_0, \dots, \boldsymbol{\beta}_k \in \mathbb{L}_2}{\operatorname{arginf}} \sum_{i=1}^n \left\| \boldsymbol{\beta}_0 + \sum_{j=1}^k \boldsymbol{\beta}_j x_{i,j} - (\check{\mathbf{q}}_i \circ \gamma_i) \sqrt{\gamma_i} \right\|_{\mathbb{L}_2}^2$
 // \mathbb{L}_2 spline fit via least-squares

return $\hat{\boldsymbol{\beta}}_j = \hat{\boldsymbol{\beta}}_{j,new} \quad \forall j = 0, \dots, k$

Algorithm 3: Fréchet regression for elastic curves

Input: data pairs $(\mathbf{x}_i, \mathbf{q}_i)$, $i = 1, \dots, n$ where \mathbf{q}_i are the SRV transformations of observed curves and

$\mathbf{x}_i = (x_{i,1}, \dots, x_{i,k})$, $i = 1, \dots, n$ are observed covariates with mean $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$ and empirical

covariance matrix $\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$; new covariate values $\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_N$; convergence tolerance $\epsilon > 0$

Compute initial mean $\bar{\mathbf{p}}_{l,new} = \underset{\bar{\mathbf{p}}}{\operatorname{arginf}} \sum_{i=1}^n \|\bar{\mathbf{p}} - \mathbf{q}_i\|_{\mathbb{L}_2}^2$, $l = 1, \dots, N$;

for $l = 1, \dots, N$ **do**

$s_{l,i} = s(\mathbf{x}_i, \mathbf{x}_l) = 1 + (\mathbf{x}_i - \bar{\mathbf{x}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x}_l - \bar{\mathbf{x}}) \quad \forall i = 1, \dots, n$; // compute weights

 Set $\bar{\mathbf{p}}_{l,old} = \bar{\mathbf{p}}_{l,new}$;

while $\|\bar{\mathbf{p}}_{l,old} - \bar{\mathbf{p}}_{l,new}\| > \epsilon$ **do**

$\bar{\mathbf{p}}_{l,old} = \bar{\mathbf{p}}_{l,new}$;

$\gamma_i = \underset{\gamma}{\operatorname{arginf}} \left\| \bar{\mathbf{p}}_{l,old} - (\mathbf{q}_i \circ \gamma) \sqrt{\gamma} \right\|_{\mathbb{L}_2}^2$, $\forall i = 1, \dots, n$; // warping step

$\bar{\mathbf{p}}_{l,new} = \underset{\bar{\mathbf{p}}}{\operatorname{arginf}} \sum_{i=1}^n s_{l,i} \left\| \bar{\mathbf{p}} - (\mathbf{q}_i \circ \gamma_i) \sqrt{\gamma_i} \right\|_{\mathbb{L}_2}^2$ // weighted \mathbb{L}_2 spline fit

return $\bar{\mathbf{p}}_l = \bar{\mathbf{p}}_{l,new}$ for all $l = 1, \dots, N$

For the the weighted \mathbb{L}_2 spline fit step note that the average of the weights fulfills

$$\frac{1}{n} \sum_{i=1}^n s_{l,i} = 1 + \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x}_l - \bar{\mathbf{x}}) = 1 + \left(\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i - \bar{\mathbf{x}} \right)^T \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x}_l - \bar{\mathbf{x}}) = 1$$

and therefore the weighted \mathbb{L}_2 mean in the spline fit step can be written as

$$\begin{aligned} \operatorname{arginf}_{\bar{\mathbf{p}}} \sum_{i=1}^n s_{l,i} \left\| \bar{\mathbf{p}} - (\mathbf{q}_i \circ \gamma_i) \sqrt{\dot{\gamma}_i} \right\|_{\mathbb{L}_2}^2 &= \operatorname{arginf}_{\bar{\mathbf{p}}} \sum_{i=1}^n (s_{l,i} \|\bar{\mathbf{p}}\|^2 - 2s_{l,i} \langle \bar{\mathbf{p}}, (\mathbf{q}_i \circ \gamma_i) \sqrt{\dot{\gamma}_i} \rangle_{\mathbb{L}_2}) \\ &= \operatorname{arginf}_{\bar{\mathbf{p}}} \sum_{i=1}^n (\|\bar{\mathbf{p}}\|^2 - 2\langle \bar{\mathbf{p}}, s_{l,i} (\mathbf{q}_i \circ \gamma_i) \sqrt{\dot{\gamma}_i} \rangle_{\mathbb{L}_2}) \\ &= \operatorname{arginf}_{\bar{\mathbf{p}}} \sum_{i=1}^n \left\| \bar{\mathbf{p}} - s_{l,i} (\mathbf{q}_i \circ \gamma_i) \sqrt{\dot{\gamma}_i} \right\|_{\mathbb{L}_2}^2. \end{aligned}$$

This means we can find a solution for the weighted least-squares via fitting a spline to the pseudo data $s_{l,i} (\mathbf{q}_i \circ \gamma_i) \sqrt{\dot{\gamma}_i}$, $i = 1, \dots, n$. I.e, we can use the \mathbb{L}_2 spline fit step described in Web Appendix A of Steyer et al. (2022) to also fit the weighted \mathbb{L}_2 mean. In particular this allows us to compute closed weighted means and therefore perform Fréchet regression for closed elastic curves.

B Additional plots and tables

B.1 Additional simulation results

We show here the prediction for all covariates $x = -1, 0.8, \dots, 0.8, 1$ for the simulation runs we picked as an example in Section 5. All runs of simulations 1-4 show similar results as displayed in Fig. 6 but differ in the number of observed points and the added noise to the observations. Likewise, Fig. 7 shows an example of the second scenario (Simulations 5-8) and Fig. 8 an example of the third scenario (Simulations 9-12).

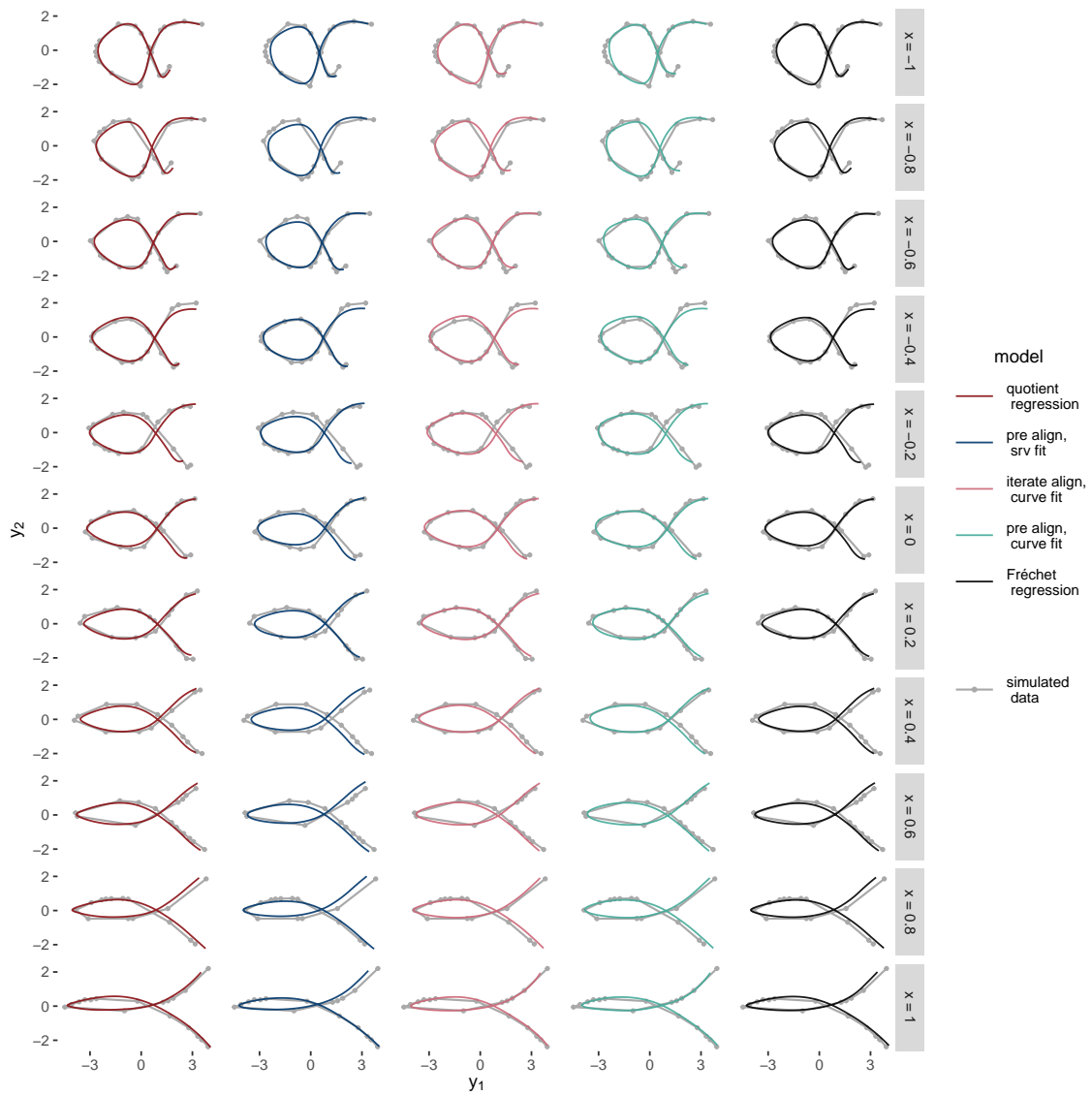


Figure 6: Predictions for an exemplarily selected run of simulation 1.

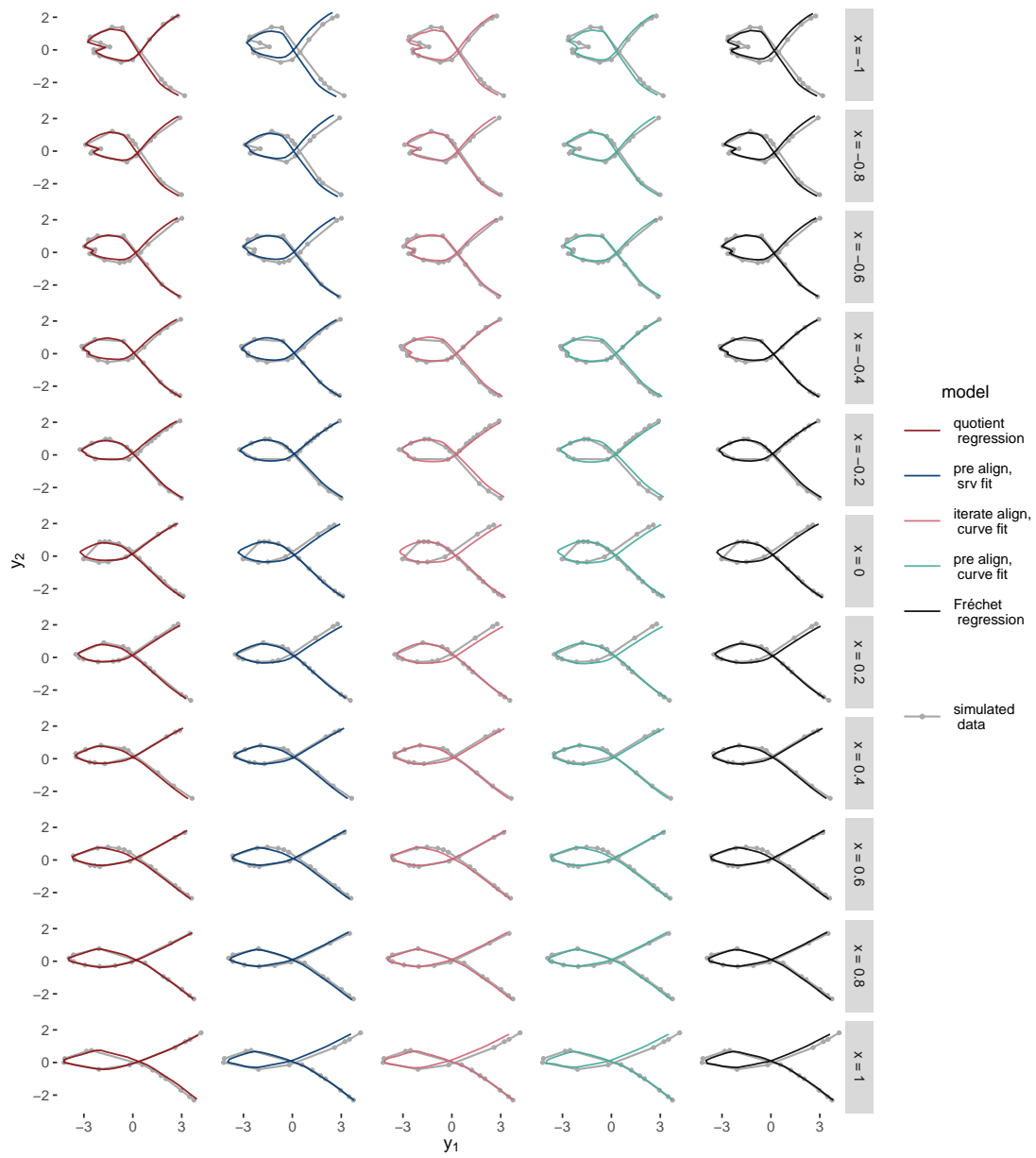


Figure 7: Predictions for an exemplarily selected run of simulation 5.

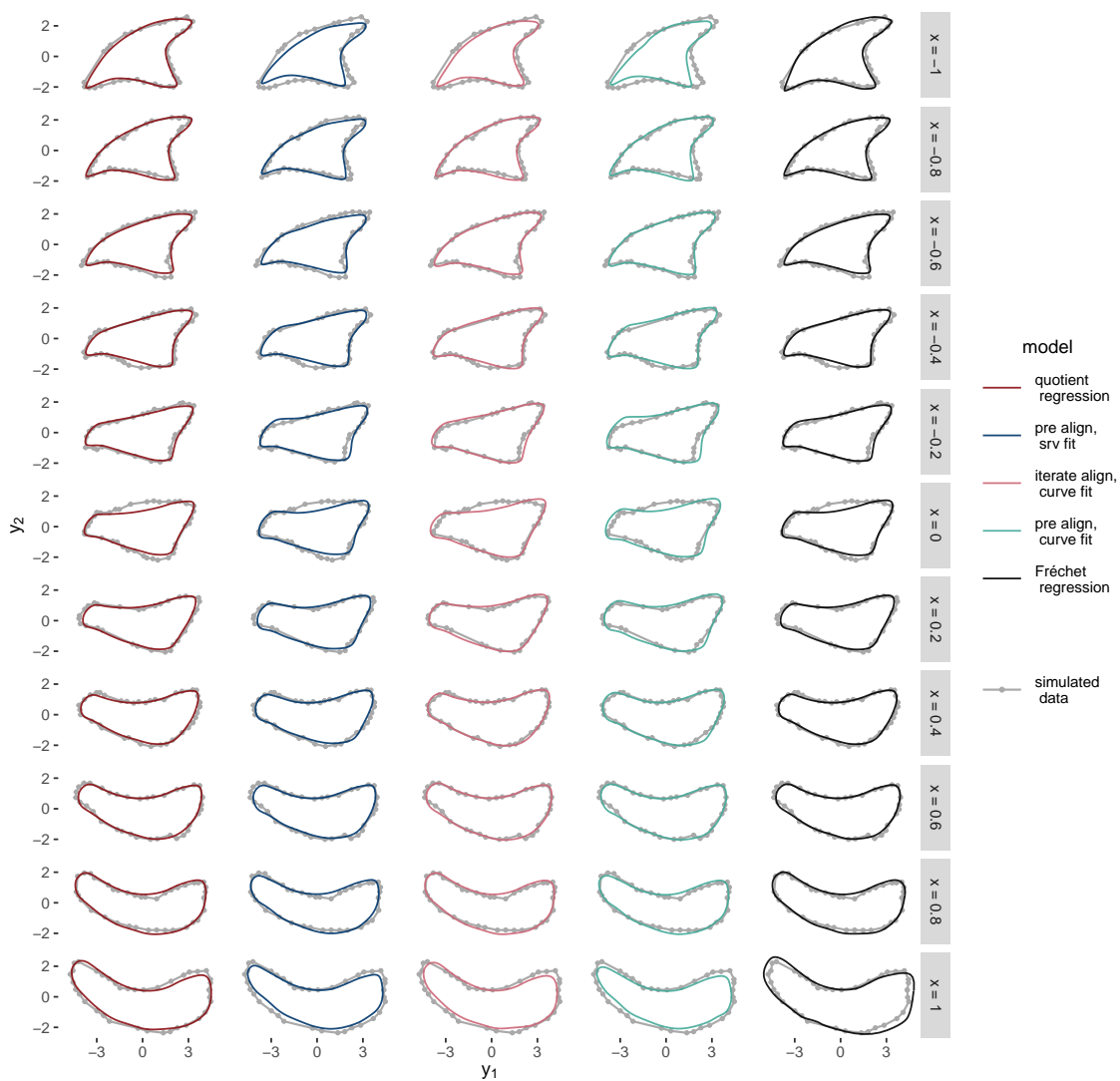


Figure 8: Predictions for an exemplarily selected run of simulation 11.

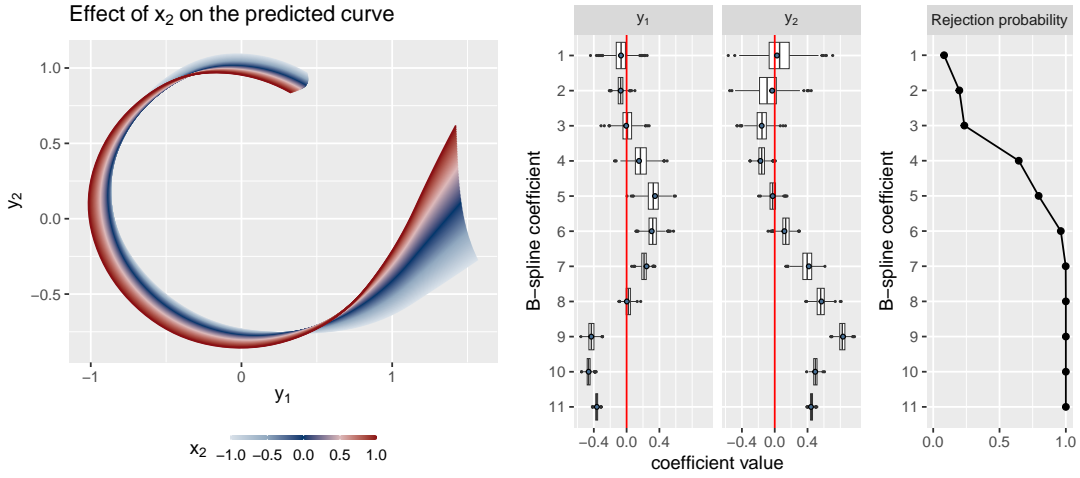


Figure 9: Results for the simulation for $n = 60$ and $N_{boot} = 1000$ using 11 instead of 6 spline coefficients per dimension. Left: Effect of x_2 given $x_1 = 0$ estimated on a large simulated sample ($n = 500$) as an approximation of the closest model to the true model in our model space. Note that the curves are translation invariant, hence the effect appears to be small in the left part of the curve. Middle: Distribution of mean bootstrapped coefficients $\bar{\xi}_{2,m}$, $m = 1, \dots, 12$ over the 1000 repetitions; the coefficients of the effect estimated on the large sample are displayed as blue dots. Coefficients $\xi_{j,1}, \dots, \xi_{j,11}$ correspond to regions on the curves from the top anti-clockwise to the right. Right: Rejection probabilities for the tests of the individual spline coefficients.

B.2 Additional application results

full	no Age	no Group	no Sex	only Group	only Sex	only Age
0.021	0.014	0.010	0.019	0.007	-0.001	0.011

Table 3: Adjusted coefficients of determination \bar{R}_{adj}^2 for all possible sub-models.

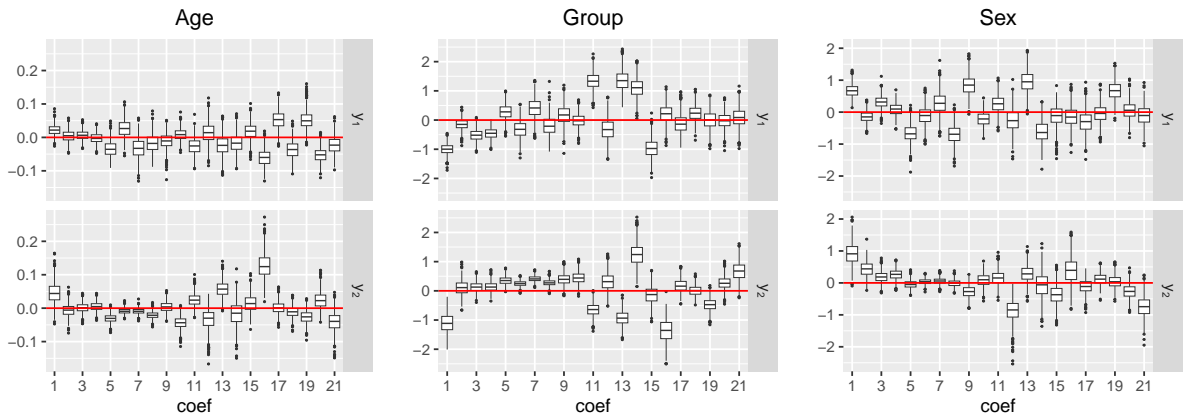


Figure 10: Distribution of the bootstrapped coefficients $\xi_{j,m}^b \in \mathbb{R}^2$, $m = 1, \dots, 21$, $j \in \{\text{Age, Group, Sex}\}$, $b = 1, \dots, 1000$ of the left hippocampus.

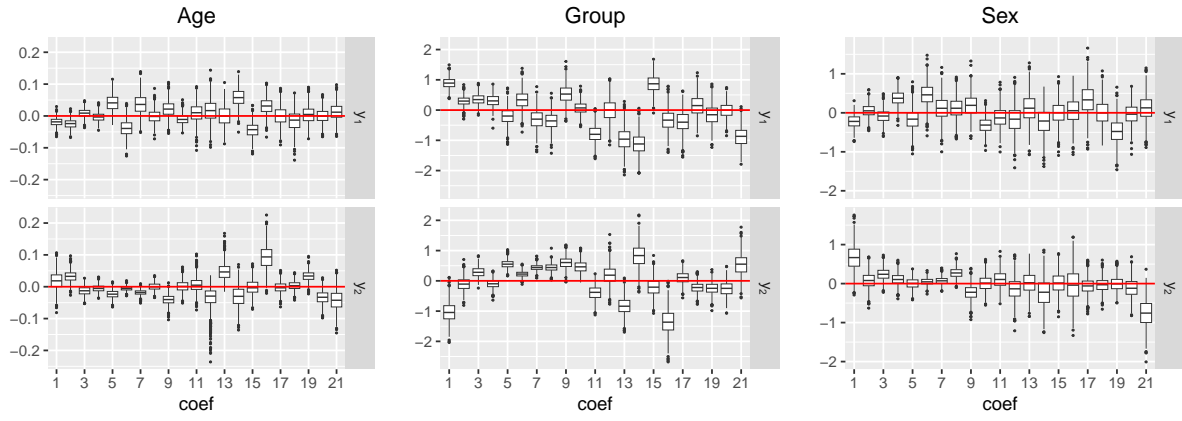


Figure 11: Distribution of the bootstrapped coefficients $\xi_{j,m}^b \in \mathbb{R}^2$, $m = 1, \dots, 21$, $j \in \{\text{Age, Group, Sex}\}$, $b = 1, \dots, 1000$ of the right hippocampus.

5. Paper IV: Functional Additive Models on Manifolds of Planar Shapes and Forms

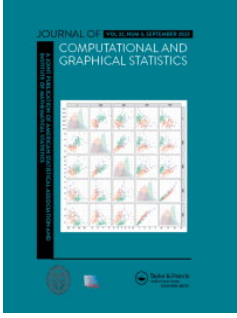
In addition to Paper III, Paper IV focuses on regression for functional planar shapes as response objects, i.e., on equivalence classes of functions taking values in \mathbb{R}^2 with respect to translation, rotation, and rescaling, but not reparameterization. This implies that the response space has a Riemannian manifold structure (see Subsection 1.3.1), which means that the conditional mean shape can be modeled by a geodesic response function, with residuals and distances determined by the shape geometry, as discussed in Subsection 1.2.3. To demonstrate the effectiveness of these methods, illustrations are provided from a morphological study of sheep bone shapes and from the analysis of cell shapes generated in biophysical simulations.

Contributing article:

Stöcker, A., Steyer, L. and Greven, S. (2023). Functional Additive Models on Manifolds of Planar Shapes and Forms. *Journal of Computational and Graphical Statistics*, to appear, 1-24. DOI: 10.1080/10618600.2023.2175687

Declaration on personal contributions:

Almond Stöcker conducted most of this research independently. Sonja Greven and the author supported the process with detailed advice and discussions. This work is also part of Almond Stöcker's dissertation.



Functional Additive Models on Manifolds of Planar Shapes and Forms

Almond Stöcker, Lisa Steyer & Sonja Greven

To cite this article: Almond Stöcker, Lisa Steyer & Sonja Greven (15 Mar 2023): Functional Additive Models on Manifolds of Planar Shapes and Forms, Journal of Computational and Graphical Statistics, DOI: [10.1080/10618600.2023.2175687](https://doi.org/10.1080/10618600.2023.2175687)

To link to this article: <https://doi.org/10.1080/10618600.2023.2175687>



© 2023 The Author(s). Published with license by Taylor & Francis Group, LLC.



[View supplementary material](#)



Published online: 15 Mar 2023.



[Submit your article to this journal](#)



Article views: 295



[View related articles](#)



[View Crossmark data](#)

Functional Additive Models on Manifolds of Planar Shapes and Forms

Almond Stöcker, Lisa Steyer, and Sonja Greven

School of Business and Economics, Humboldt-Universität zu Berlin, Berlin, Germany

ABSTRACT

The “shape” of a planar curve and/or landmark configuration is considered its equivalence class under translation, rotation, and scaling, its “form” its equivalence class under translation and rotation while scale is preserved. We extend generalized additive regression to models for such shapes/forms as responses respecting the resulting quotient geometry by employing the squared geodesic distance as loss function and a geodesic response function to map the additive predictor to the shape/form space. For fitting the model, we propose a Riemannian L_2 -Boosting algorithm well suited for a potentially large number of possibly parameter-intensive model terms, which also yields automated model selection. We provide novel intuitively interpretable visualizations for (even nonlinear) covariate effects in the shape/form space via suitable tensor-product factorization. The usefulness of the proposed framework is illustrated in an analysis of (a) astragalus shapes of wild and domesticated sheep and (b) cell forms generated in a biophysical model, as well as (c) in a realistic simulation study with response shapes and forms motivated from a dataset on bottle outlines. Supplementary materials for this article are available online.

ARTICLE HISTORY

Received November 2021
Accepted December 2022

KEYWORDS



Boosting; Functional regression; Shape analysis; Tensor-product model; Visualization


1. Introduction

In many imaging data problems, the coordinate system of recorded objects is arbitrary or explicitly not of interest. Statistical shape analysis (Dryden and Mardia 2016) addresses this point by identifying the ultimate object of analysis as the *shape* of an observation, reflecting its geometric properties invariant under translation, rotation and rescaling, or as its *form* (or *size-and-shape*) invariant under translation and rotation. This article establishes a flexible additive regression framework for modeling the shape or form of planar (potentially irregularly sampled) curves and/or landmark configurations in dependence on scalar covariates. A rich shape analysis literature has been developed for 2D or 3D landmark configurations—presenting for instance selected points of a bone or face—which are considered elements of Kendall’s shape space (see, e.g., Dryden and Mardia 2016). In many 2D scenarios, however, observed points describe a curve reflecting the outline of an object rather than dedicated landmarks (Adams, Rohlf, and Slice 2013). Considering outlines as images of (parameterized) curves shows a direct link to functional data analysis (FDA, Ramsay and Silverman 2005) and, in this context, we speak of functional shape/form data analysis. As in FDA, functional shape/form data can be observed on a common and often dense grid (*regular/dense* design) or on curve-specific often sparse grids (*irregular/sparse* design). While in the regular case, analysis often simplifies by treating curve evaluations as multivariate data, more general irregular designs gave rise to further developments in sparse FDA (e.g., Yao, Müller, and Wang 2005; Greven and Scheipl 2017), explicitly considering irregular measurements instead of pre-smoothing curves. To the best of our knowledge, we are the first to consider

irregular/sparse designs in the context of functional shape/form analysis.

Shapes and forms are examples of manifold data. Petersen and Müller (2019) propose “Fréchet regression” for random elements in general metric spaces, which requires estimation of a (potentially negatively) weighted Fréchet mean for each covariate combination. Their implicit rather than explicit model formulation renders model interpretation difficult. More explicit model formulations have been developed for the special case of a Riemannian geometry. Besides tangent space models (Kent et al. 2001), extrinsic models (Lin et al. 2017) and models based on unwrapping (Jupp and Kent 1987; Mallasto and Feragen 2018), a variety of manifold regression models have been designed based on the intrinsic Riemannian geometry. Starting from geodesic regression (Fletcher 2013), which extends linear regression to curved spaces, these include MANOVA (Huckemann, Hotz, and Munk 2010), polynomial regression (Hinkle, Fletcher, and Joshi 2014), smoothing splines (Kume, Dryden, and Le 2007), regression along geodesic paths with nonconstant speed (Hong et al. 2014), or kernel regression (Davis et al. 2010) and Kriging (Pigoli, Menafoglio, and Secchi 2016). However, mostly only one metric covariate or categorical covariates are considered, possibly in hierarchical model extensions for longitudinal data (Muralidharan and Fletcher 2012; Schiratti et al. 2017). By contrast, Zhu et al. (2009), Shi et al. (2009), and Kim et al. (2014) generalize geodesic regression to regression with multiple covariates focusing on Symmetric Positive-Definite (SPD) matrix responses. Cornea et al. (2017) develop a general Generalized Linear Model (GLM) analogue regression framework for responses in a symmetric manifold and apply it to shape analysis.

CONTACT Almond Stöcker  almond.stoecker@hu-berlin.de  School of Business and Economics, Humboldt-Universität zu Berlin, Berlin, Germany.

 Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

© 2023 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

Recently, Lin, Müller, and Park (2020) proposed a Lie group additive regression model for Riemannian manifolds focusing on SPD matrices rather than shapes.

In FDA, there is a much wider range of developed regression methods (see overviews in Morris 2015; Greven and Scheipl 2017). Among the most flexible models are Functional Additive Models (FAMs) for (univariate) functional responses (in contrast to FAMs with functional covariates (Ferraty et al. 2011)) with either semi- or nonparametric approaches to model (a) response functions and (b) smooth covariate effects. For (a), nonparametric approaches formulate estimation problems in infinite-dimensional model spaces to motivate finite-dimensional representations or effectively evaluate curves on grids (e.g., Jeon and Park 2020). Semi-parametric approaches directly employ finite expansions in spline bases (Brockhaus, Scheipl, and Greven 2015), Functional Principal Component (FPC) bases (Morris and Carroll 2006) or both (Scheipl, Staicu, and Greven 2015), as well as wavelets (Meyer et al. 2015), sometimes directly expanding functions to model on coefficients and sometimes expanding only predictions while keeping the raw measurements. Nonparametric approaches are formulated in infinite-dimensional model spaces and effectively evaluate curves on grids or apply pre-smoothing techniques (e.g., Jeon and Park 2020). For (b), again semiparametric penalized spline basis approaches are employed (Scheipl, Staicu, and Greven 2015; Brockhaus, Scheipl, and Greven 2015), or local linear/polynomial (Müller and Yao 2008; Jeon et al. 2022) or other nonparametric kernel-based approaches (Jeon and Park 2020; Jeon, Park, and Van Keilegom 2021). Semi- and nonparametric approaches come with different theoretical and practical advantages, but similarities such as regarding asymptotic behavior are also known from scalar nonparametric regression (Li and Ruppert 2008). Advantages of the semi-parametric approach summarized in Greven and Scheipl (2017) include its appropriateness for sparse irregular functional data and its modular extensibility to functional mixed models (Scheipl, Staicu, and Greven 2015; Meyer et al. 2015) and nonstandard response distributions (Brockhaus, Scheipl, and Greven 2015; Stöcker et al. 2021). For bivariate or multivariate functional responses, which are closest to functional shapes/forms but without invariances, Rosen and Thompson (2009), Zhu, Li, and Kong (2012), Olsen, Markussen, and Raket (2018) consider linear fixed effects of scalar covariates, the latter also allowing for warping. Zhu et al. (2017), Backenroth et al. (2018) consider one or more random effects for one grouping variable, linear fixed effects and common dense grids for all functions. Volkman et al. (2021) combine the FAM model class of Greven and Scheipl (2017) with multivariate FPC analysis (Happ and Greven 2018) to model multivariate (sparse) functional responses.

This article establishes an interpretable FAM framework for modeling the shape or form of planar (potentially irregularly sampled) curves and/or landmark configurations in dependence on scalar covariates, extending L_2 -Boosting (Bühlmann and Yu 2003; Brockhaus, Scheipl, and Greven 2015) to Riemannian manifolds for model estimation. The three major contributions of our regression framework are: (i) We introduce additive regression with shapes/forms of planar curves and/or landmarks as response, extending FAMs to nonlinear response spaces or, vice versa, extending GLM-type regression on manifolds for

landmark shapes both to functional shape manifolds and to include (nonlinear) additive model effects. (ii) We propose a novel Riemannian L_2 -Boosting algorithm for estimating regression models for this type of manifold response, and (iii) a visualization technique based on tensor-product factorization yielding intuitive interpretations even of multi-dimensional smooth covariate effects for practitioners. Although related tensor-product model transformations based on higher-order SVD have been used, e.g., in control engineering (Baranyi, Yam, and Várlaki 2013), we are not aware of any comparable application for visualization in FAMs or other statistical models for object data. Despite our focus on shapes and forms, transfer of the model, Riemannian L_2 -Boosting, and factorized visualization to other Riemannian manifold responses is intended in the generality of the formulation and the design of the provided R package `manifoldboost` (developer version on github.com/Almond-S/manifoldboost). The versatile applicability of the approach is illustrated in three different scenarios: an analysis of the shape of sheep astragali (ankle bones) represented by both regularly sampled curves and landmarks in dependence on categorical “demographic” variables; an analysis of the effects of different metric biophysical model parameters (including smooth interactions) on the form of (irregularly sampled) cell outlines generated from a cellular Potts model; and a simulation study with irregularly sampled functional shape and form responses generated from a dataset of different bottle outlines and including metric and categorical covariates.

In Section 2, we introduce the manifold geometry of irregular curves modulo translation, rotation and potentially rescaling, which underlies the intrinsic additive regression model formulated in Section 3. The Riemannian L^2 -Boosting algorithm is introduced in Section 4. Section 5 analyzes different data problems, modeling sheep bone shape responses (Section 5.1) and cell outlines (Section 5.2). Section 5.3 summarizes the results of simulation studies with functional shape and form responses. We conclude with a discussion in Section 6.

2. Geometry of Functional Shapes and Forms

Riemannian manifolds of planar shapes (and forms) are discussed in various textbooks at different levels of generality, in finite (Kendall et al. 1999; Dryden and Mardia 2016) or potentially infinite dimensions (Srivastava and Klassen 2016; Klingenberg 1995). Starting from the Hilbert space \mathcal{Y} of curve representatives y of a single shape or form observation, we successively characterize its quotient space geometry under translation, rotation and rescaling including the respective tangent spaces. Building on that, we introduce Riemannian exponential and logarithmic maps and parallel transports needed for model formulation and fitting, and the sample space of (irregularly observed) functional shapes/forms.

To make use of complex arithmetic, we identify the two-dimensional plane with the complex numbers, $\mathbb{R}^2 \cong \mathbb{C}$, and consider a planar curve to be a function $y : \mathbb{R} \supset \mathcal{T} \rightarrow \mathbb{C}$, element of a separable complex Hilbert space \mathcal{Y} with a complex inner product $\langle \cdot, \cdot \rangle$ and corresponding norm $\| \cdot \|$. This allows simple scalar expressions for the group actions of translation $\text{Trl} = \{y \xrightarrow{\text{Trl}_\gamma} y + \gamma t : \gamma \in \mathbb{C}\}$ with $t \in \mathcal{Y}$ canonically

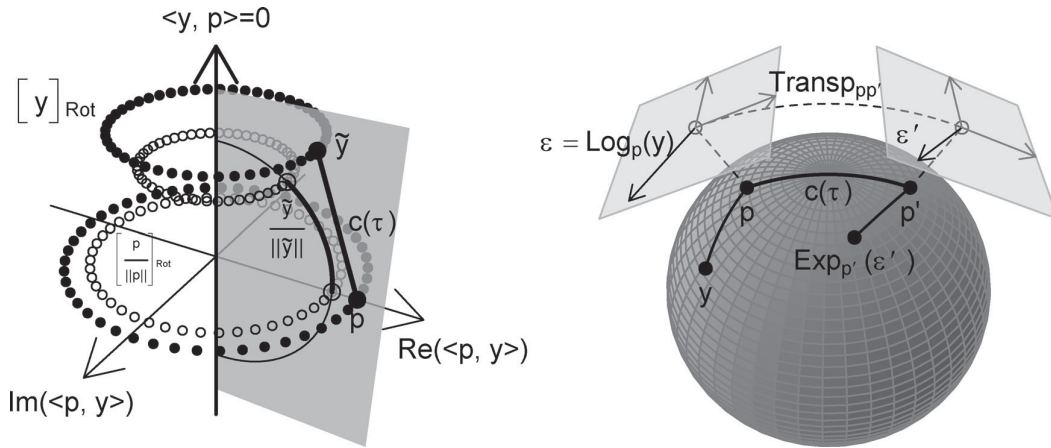


Figure 1. *Left:* Quotient space geometry: assuming p and y centered, translation invariance is not further considered in the plot; given pole representative p , we express $y = \frac{\text{Re}\langle p, y \rangle}{\|p\|^2} p + \frac{\text{Im}\langle p, y \rangle}{\|p\|^2} ip + (y - \frac{\langle p, y \rangle}{\|p\|^2} p) \in \mathcal{Y}$ in its coordinates in p and ip direction, subsuming all orthogonal directions in the third dimension. In this coordinate system, the rotation orbit $[y]_{\text{Rot}}$ corresponds to the dotted horizontal circle, and is identified with the aligned $\tilde{y} := \tilde{y}^{\text{Rot}}$ in the half-plane of p ; $[y]_{\text{Rot}} \times \text{Scl}$ is identified with the unit vector $\tilde{y}^{\text{Rot}} \times \text{Scl} = \frac{\tilde{y}}{\|\tilde{y}\|}$ projecting \tilde{y} onto the hemisphere depicted by the vertical semicircle. Form and shape distances between $[p]$ and $[y]$ correspond to the length of the geodesics $c(\tau)$ on the plane and sphere, respectively. *Right:* Geodesic line $c(\tau)$ between $p = c(0)$ and $p' = c(1)$, Log-map projecting y to $\epsilon \in T_p \mathcal{M}$, parallel transport $\text{Trans}_{pp'}$ forwarding ϵ to $\epsilon' \in T_{p'} \mathcal{M}$, and Exp-map projecting ϵ' onto \mathcal{M} visualized for a sphere. Tangent spaces, identified with subspaces of the ambient space, are depicted as *gray planes* above the respective poles. The parallel transport preserves all angles between tangent vectors and identifies $\dot{c}(0) \cong \dot{c}(1)$.

given by $f : t \mapsto \frac{1}{\|t-1\|}$ the real constant function of unit norm; rescaling $\text{Scl} = \{y \xrightarrow{\text{Scl}} \lambda \cdot (y - o_y) + o_y : \lambda \in \mathbb{R}^+\}$ around the centroid $o_y = \langle t, y \rangle / t$ (which we consider more natural than using o , the zero element of \mathcal{Y} , mostly chosen in the literature); and rotation $\text{Rot} = \{y \xrightarrow{\text{Rot}_\omega} u \cdot (y - o_y) + o_y : u \in \mathbb{S}^1\}$ around o_y with $\mathbb{S}^1 = \{u \in \mathbb{C} : |u| = 1\} = \{\exp(i\omega\sqrt{-1}) : \omega \in \mathbb{R}\}$ reflecting counterclockwise rotations by ω radian measure. Concatenation yields combined group actions G as direct products, such as the rigid motions $G = \text{Trl} \times \text{Rot} = \{\text{Trl}_\gamma \circ \text{Rot}_u : \gamma \in \mathbb{C}, u \in \mathbb{S}^1\} \cong \mathbb{C} \times \mathbb{S}^1$ (see Section S.1.1, supplementary materials for more details). The two real-valued component functions of y are identified with the real part $\text{Re}(y) : \mathcal{T} \rightarrow \mathbb{R}$ and imaginary part $\text{Im}(y) : \mathcal{T} \rightarrow \mathbb{R}$ of $y = \text{Re}(y) + \text{Im}(y) \sqrt{-1}$. While the complex setup is used for convenience, the real part of $\langle \cdot, \cdot \rangle$ constitutes an inner product $\text{Re}\langle y_1, y_2 \rangle = \langle \text{Re}(y_1), \text{Re}(y_2) \rangle + \langle \text{Im}(y_1), \text{Im}(y_2) \rangle$ for $y_1, y_2 \in \mathcal{Y}$ on the underlying real vector space of planar curves. Typically $\text{Re}(y)$, $\text{Im}(y)$ are assumed square-integrable with respect to a measure ν and we consider the canonical inner product $\langle y_1, y_2 \rangle = \int y_1^\dagger y_2 d\nu$ where y^\dagger denotes the conjugate transpose of y , that is, $y^\dagger(t) = \text{Re}(y)(t) - \text{Im}(y)(t) \sqrt{-1}$ is simply the complex conjugate, but for vectors $\mathbf{y} \in \mathbb{C}^k$, the vector \mathbf{y}^\dagger is also transposed. For curves, we typically assume ν to be the Lebesgue measure on $\mathcal{T} = [0, 1]$; for landmarks, a standard choice is the counting measure on $\mathcal{T} = \{1, \dots, k\}$.

The ultimate response object is given by the *orbit* $[y]_G = \{g(y) : g \in G\}$ (or short $[y]$) of $y \in \mathcal{Y}$, the equivalence class under the respective combined group actions G : with $G = \text{Trl} \times \text{Rot} \times \text{Scl}$, $[y] = [y]_{\text{Trl} \times \text{Rot} \times \text{Scl}} = \{\lambda u y + \gamma t : \lambda \in \mathbb{R}^+, u \in \mathbb{S}^1, \gamma \in \mathbb{C}\}$ is referred to as the *shape* of y and, for $G = \text{Trl} \times \text{Rot}$, $[y] = [y]_{\text{Trl} \times \text{Rot}} = \{u y + \gamma t : u \in \mathbb{S}^1, \gamma \in \mathbb{C}\}$ as its *form* or *size-and-shape*. $\mathcal{Y}/G = \{[y]_G : y \in \mathcal{Y}\}$ denotes the quotient space of \mathcal{Y} with respect to G . The description

of the Riemannian geometry of \mathcal{Y}/G involves, in particular, a description of the tangent spaces $T_{[y]} \mathcal{Y}/G$ at points $[y] \in \mathcal{Y}/G$, which can be considered local vector space approximations to \mathcal{Y}/G in a neighborhood of $[y]$. For a point q in a manifold \mathcal{M} the tangent vectors $\beta \in T_q \mathcal{M}$ can, i.e., be thought of as gradients $\dot{c}(0)$ of paths $c : \mathbb{R} \supset (-\delta, \delta) \rightarrow \mathcal{M}$ at 0 where they pass through $c(0) = q$. Besides their geometric meaning, they will also play an important role in the regression model, as additive model effects are formulated on tangent space level. Choosing suitable representatives $\tilde{y}^G \in [y]_G \subset \mathcal{Y}$ (or short \tilde{y}) of orbits $[y]_G$, we use an identification of tangent spaces with suitable linear subspaces $T_{[y]_G} \mathcal{Y}/G \subset \mathcal{Y}$.

Form geometry: Starting with translation as the simplest invariance, an orbit $[y]_{\text{Trl}}$ can be one-to-one identified with its centered representative $\tilde{y}^{\text{Trl}} = y - \langle y, t \rangle t$ yielding an identification $\mathcal{Y}/_{\text{Trl}} \cong \{y \in \mathcal{Y} : \langle y, t \rangle = 0\}$ with a linear subspace of \mathcal{Y} . Hence, also $T_{[y]} \mathcal{Y}/_{\text{Trl}} = \{y \in \mathcal{Y} : \langle y, t \rangle = 0\}$. For rotation, by contrast, we can only find local identifications with Hilbert subspaces (i.e., charts) around reference points $[p]_{\text{Trl}} \times \text{Rot}$ we refer to as “poles”. Moreover, we restrict to $y, p \in \mathcal{Y}^* = \mathcal{Y} \setminus [0]_{\text{Trl}}$ eliminating constant functions as degenerate special cases in the translation orbit of zero. For each $[y]_{\text{Trl}} \times \text{Rot}$ in an open neighborhood around $[p]_{\text{Trl}} \times \text{Rot}$ which can be chosen with $(\tilde{y}^{\text{Trl}}, \tilde{p}^{\text{Trl}}) \neq 0$, y can be uniquely rotation aligned to p , yielding a one-to-one identification of the form $[y]_{\text{Trl}} \times \text{Rot}$ with the aligned representative given by $\tilde{y}^{\text{Trl} \times \text{Rot}} = \frac{\langle \tilde{y}^{\text{Trl}}, \tilde{p}^{\text{Trl}} \rangle}{\|\tilde{y}^{\text{Trl}}, \tilde{p}^{\text{Trl}}\|} \tilde{y}^{\text{Trl}} = \underset{y' \in [y]_{\text{Trl}} \times \text{Rot}}{\text{argmin}} \|y' - p\|$ (compare Figure 1).

While $\tilde{y}^{\text{Trl} \times \text{Rot}}$ depends on p , we omit this in the notation for simplicity. All \tilde{y}^{Trl} rotation aligned to \tilde{p}^{Trl} lie on the hyper-plane determined by $\text{Im}\langle \tilde{y}^{\text{Trl}}, \tilde{p}^{\text{Trl}} \rangle = 0$ (Figure 1), which yields $T_{[p]} \mathcal{Y}^*_{/\text{Trl} \times \text{Rot}} = \{y \in \mathcal{Y} : \langle y, t \rangle = 0, \text{Im}\langle y, p \rangle = 0\}$ with normal vectors $\zeta^{(1)} = t, \zeta^{(2)} = \sqrt{-1} t, \zeta^{(3)} = \sqrt{-1} p$. Note that, despite the use of complex arithmetic, $T_{[p]} \mathcal{Y}^*_{/\text{Trl} \times \text{Rot}}$ is a real

vector space not closed under complex scalar multiplication. The geodesic distance of $[y]_{\text{Trl} \times \text{Rot}}$ to the pole $[p]_{\text{Trl} \times \text{Rot}}$ is given by $d([y]_{\text{Trl} \times \text{Rot}}, [p]_{\text{Trl} \times \text{Rot}}) = \|\tilde{y}^{\text{Trl} \times \text{Rot}} - \tilde{p}^{\text{Trl} \times \text{Rot}}\| = \operatorname{argmin}_{y' \in [y]_{\text{Trl} \times \text{Rot}}, p' \in [p]_{\text{Trl} \times \text{Rot}}} \|y' - p'\|$. It reflects the length of the shortest path (i.e., the geodesic) between the forms and the minimum distance between the orbits as sets.

Shape geometry: To account for scale invariance in shapes $[y]_{\text{Trl} \times \text{Rot} \times \text{Scl}}$, they are identified with normalized representatives $\tilde{y}^{\text{Trl} \times \text{Rot} \times \text{Scl}} = \frac{\tilde{y}^{\text{Trl} \times \text{Rot}}}{\|\tilde{y}^{\text{Trl} \times \text{Rot}}\|}$. Motivated by the normalization, we borrow the well-known geometry of the sphere $\mathbb{S} = \{y \in \mathcal{Y} : \|y\| = 1\}$, where $T_p\mathbb{S} = \{y \in \mathcal{Y} : \operatorname{Re}\langle y, p \rangle = 0\}$ is the tangent space at a point $p \in \mathbb{S}$ and geodesics are great circles. Together with translation and rotation invariance, the shape tangent space is then given by $T_{[p]} \mathcal{Y}_{\text{Trl} \times \text{Rot} \times \text{Scl}}^* = T_{[p]} \mathcal{Y}_{\text{Trl} \times \text{Rot}}^* \cap T_p\mathbb{S} = \{y \in \mathcal{Y} : \langle y, \ell \rangle = 0, \langle y, p \rangle = 0\}$ with normal vector $\zeta^{(4)} = p$ in addition to $\zeta^{(1)}, \zeta^{(2)}, \zeta^{(3)}$ above. The geodesic distance $d([p]_{\text{Trl} \times \text{Rot} \times \text{Scl}}, [y]_{\text{Trl} \times \text{Rot} \times \text{Scl}}) = \arccos |\langle \tilde{y}^{\text{Trl} \times \text{Rot} \times \text{Scl}}, \tilde{p}^{\text{Trl} \times \text{Rot} \times \text{Scl}} \rangle|$ corresponds to the arc-length between the representatives. This distance is often referred to as *Procrustes distance* in statistical shape analysis.

We may now define the maps needed for the regression model formulation. Let \tilde{y} and \tilde{p} be shape/form representatives of $[y]$ and $[p]$ rotation aligned to the shape/form pole representative p . Generalizing straight lines to a Riemannian manifold \mathcal{M} , geodesics $c : (-\delta, \delta) \rightarrow \mathcal{M}$ can be characterized by their “intercept” $c(0) \in \mathcal{M}$ and “slope” $\dot{c}(0) \in T_{c(0)}\mathcal{M}$. The *exponential map* $\operatorname{Exp}_q : T_q\mathcal{M} \rightarrow \mathcal{M}$ at a point $q \in \mathcal{M}$ is defined to map $\beta \mapsto c(1)$ for c the geodesic with $q = c(0)$ and $\beta = \dot{c}(0)$. It maps $\beta \in T_q\mathcal{M}$ to a point $\operatorname{Exp}_q(\beta) \in \mathcal{M}$ located $d(q, \operatorname{Exp}_q(\beta)) = \|\beta\|$ apart of the pole q in the direction of β . On the form space $\mathcal{Y}_{\text{Trl} \times \text{Rot}}$, the exponential map is simply given by $\operatorname{Exp}_{[p]_{\text{Trl} \times \text{Rot}}}(\beta) = [\tilde{p}^{\text{Trl} \times \text{Rot}} + \beta]_{\text{Trl} \times \text{Rot}}$. On the shape space $\mathcal{Y}_{\text{Trl} \times \text{Rot} \times \text{Scl}}$, identification with exponential maps on the sphere yields $\operatorname{Exp}_{[p]_G}(\beta) = \left[\cos(\|\beta\|) \tilde{p}^G + \sin(\|\beta\|) \frac{\beta}{\|\beta\|} \right]_G$ with $G = \text{Trl} \times \text{Rot} \times \text{Scl}$. In an open neighborhood \mathcal{U} , $q \in \mathcal{U} \subset \mathcal{M}$, Exp_q is invertible yielding the $\operatorname{Log}_q : \mathcal{U} \rightarrow T_q\mathcal{M}$ map from the manifold to the tangent space at q . For forms, it is given by $\operatorname{Log}_{[p]_{\text{Trl} \times \text{Rot}}}([y]_{\text{Trl} \times \text{Rot}}) = \tilde{y}^{\text{Trl} \times \text{Rot}} - \tilde{p}^{\text{Trl} \times \text{Rot}}$ and, for shapes, by $\operatorname{Log}_{[p]_G}([y]_G) = d([p]_G, [y]_G) \frac{\tilde{y}^G - (\tilde{p}^G, \tilde{y}^G) \tilde{p}^G}{\|\tilde{y}^G - (\tilde{p}^G, \tilde{y}^G) \tilde{p}^G\|}$ with $G = \text{Trl} \times \text{Rot} \times \text{Scl}$. Finally, $\operatorname{Transp}_{q,q'} : T_q\mathcal{M} \rightarrow T_{q'}\mathcal{M}$ parallel transports tangent vectors $\varepsilon \mapsto \varepsilon'$ isometrically along a geodesic $c(\tau)$ connecting q and $q' \in \mathcal{M}$ such that the slopes $\operatorname{Transp}_{q,q'}(\dot{c}(q)) = \dot{c}(q')$ are identified and all angles are preserved. For shapes, $\operatorname{Transp}_{[y]_G, [p]_G}(\varepsilon) = \varepsilon - \langle \varepsilon, \tilde{p}^G \rangle \frac{\tilde{y}^G + \tilde{p}^G}{1 + (\tilde{y}^G, \tilde{p}^G)}$, with $G = \text{Trl} \times \text{Rot} \times \text{Scl}$, takes the form of the parallel transport on a sphere replacing the real inner product with its complex analogue. For forms, it changes only the $\operatorname{Im}(\langle \varepsilon, \tilde{p} \rangle)$ coordinate orthogonal to the real \tilde{y} - \tilde{p} -plane as in the shape case, while the remainder of ε is left unchanged as in a linear space. This yields $\operatorname{Transp}_{[y]_G, [p]_G}(\varepsilon) = \varepsilon - \operatorname{Im}(\langle \tilde{p}^G / \|\tilde{p}^G\|, \varepsilon \rangle) \frac{\tilde{y}^G / \|\tilde{y}^G\| + \tilde{p}^G / \|\tilde{p}^G\|}{1 + (\tilde{y}^G / \|\tilde{y}^G\|, \tilde{p}^G / \|\tilde{p}^G\|)} \sqrt{-1}$, with $G = \text{Trl} \times \text{Rot}$, for form tangent vectors. While equivalent expressions for the parallel transport in the shape case can be found, for example,

in Dryden and Mardia (2016), Huckemann, Hotz, and Munk (2010), a corresponding derivation for the form case is given in Section S.1.2, supplementary materials including a discussion of the quotient space geometry in differential geometric terms.

Based on this understanding of the response space, we may now proceed to consider a sample of curves $y_1, \dots, y_n \in \mathcal{Y}$ representing orbits $[y_1], \dots, [y_n]$ with respect to group actions G . In the functional case, with the domain $\mathcal{T} = [0, 1]$, these curves are usually observed as evaluations $\mathbf{y}_i = (y_i(t_{i1}), \dots, y_i(t_{ik_i}))^\top$ on a finite grid $t_{i1} < \dots < t_{ik_i} \in \mathcal{T}$ which may differ between observations. In contrast to the *regular* case with common grids, this more general data structure is referred to as *irregular* functional shape/form data. To handle this setting, we replace the original inner product $\langle \cdot, \cdot \rangle$ on \mathcal{Y} by individual $\langle y_i, y'_i \rangle_i = \mathbf{y}_i^\top \mathbf{W}_i \mathbf{y}'_i$ providing inner products on the k_i -dimensional space $\mathcal{Y}_i = \mathbb{C}^{k_i}$ of evaluations y_i, y'_i on the same grid. The symmetric positive-definite weight matrix \mathbf{W}_i can be chosen to implement an approximation to integration w.r.t. the original measure ν with a numerical integration measure ν_i such as given by the trapezoidal rule. Alternatively, $\mathbf{W}_i = \frac{1}{k_i} \mathbf{I}_{k_i}$ with $k_i \times k_i$ identity matrix \mathbf{I}_{k_i} presents a canonical choice that is analog to the landmark case for $k_i \equiv k$. Moreover, data-driven \mathbf{W}_i could also be motivated from the covariance structure estimated for (potentially sparse) y_1, \dots, y_n along the lines of Yao, Müller, and Wang (2005), Stöcker et al. (2022). While this is beyond the scope of this article, potential procedures are sketched in Section S.7, supplementary materials. With the inner products given for $i = 1, \dots, n$, the sample space naturally arises as the Riemannian product $\mathcal{Y}_{1/G}^* \times \dots \times \mathcal{Y}_{n/G}^*$ of the orbit spaces, with the individual geometries constructed as described above.

3. Additive Regression on Riemannian Manifolds

Consider a data scenario with n observations of a random response covariate tuple (Y, \mathbf{X}) , where the realizations of Y are planar curves $y_i : \mathcal{T} \rightarrow \mathbb{C}$, $i = 1, \dots, n$, belonging to a Hilbert space \mathcal{Y} defined as above and potentially irregularly measured on individual grids $t_{i1} < \dots < t_{ik_i} \in \mathcal{T}$. The response object $[Y]$ is the equivalence class of Y with respect to translation, rotation and possibly scale and the sample $[y_1], \dots, [y_n]$ is equipped with the respective Riemannian manifold geometry introduced in the previous section. For $i = 1, \dots, n$, realizations $\mathbf{x}_i \in \mathcal{X}$ of a covariate vector \mathbf{X} in a covariate space \mathcal{X} are observed. \mathbf{X} can contain several categorical and/or metric covariates.

For regressing the mean of $[Y]$ on $\mathbf{X} = \mathbf{x}$, we model the shape/form $[\mu]$ of $\mu \in \mathcal{Y}$ as

$$[\mu] = \operatorname{Exp}_{[p]}(h(\mathbf{x})) = \operatorname{Exp}_{[p]} \left(\sum_{j=1}^J h_j(\mathbf{x}) \right), \quad (1)$$

with an additive predictor $h : \mathcal{X} \rightarrow T_{[p]} \mathcal{Y}_{1/G}^*$ acting in the tangent space at an “intercept” $[p] \in \mathcal{Y}_{1/G}^*$. Generalizing an additive model “ $Y = \mu + \varepsilon = p + h(\mathbf{x}) + \varepsilon$ ” in a linear space, we implicitly define $[\mu]$ as the conditional mean of $[Y]$ given $\mathbf{X} = \mathbf{x}$ by assuming zero-mean “residuals” ε . In their definition, we follow Cornea et al. (2017) but extend to the functional shape/form and additive case. We assume local linearized residuals $\varepsilon_{[\mu]} = \operatorname{Log}_{[\mu]}([Y])$ in $T_{[\mu]} \mathcal{Y}_{1/G}^*$ to have mean $\mathbb{E}(\varepsilon_{[\mu]}) =$

0, which corresponds to $\mathbb{E}(\varepsilon_{[\mu]}(t)) = 0$ for (ν -almost) all $t \in \mathcal{T}$. Here, we assume $[Y]$ is sufficiently close to $[\mu]$ with probability 1 such that $\text{Log}_{[\mu]}$ is well-defined, which is the case whenever $(\tilde{Y}, \tilde{\mu}) \neq 0$ for centered shape/form representatives \tilde{Y} and $\tilde{\mu}$, an unrestrictive and common assumption (compare also Cornea et al. 2017). However, residuals $\varepsilon_{[\mu]}$ for different $[\mu]$ belong to separate tangent spaces. To obtain a formulation in a common linear space instead, local residuals are mapped to residuals $\epsilon = \text{Transp}_{[\mu],[p]}(\varepsilon_{[\mu]})$ by parallel transporting them from $[\mu]$ to the common covariate independent pole $[p]$. After this isometric mapping into $T_{[p]}\mathcal{Y}_{/G}^*$, we can equivalently define the conditional mean $[\mu]$ via $\mathbb{E}(\epsilon) = 0$ for the transported residuals ϵ .

$\text{Exp}_{[p]}$ maps the additive predictor $h(\mathbf{x}) = \sum_{j=1}^J h_j(\mathbf{x}) \in T_{[p]}\mathcal{Y}_{/G}^*$ to the response space. It is analogous to a response function in GLMs but depends on $[p]$. Although covariate effects $h_j(\mathbf{x})$ often only depend on an individual covariate in \mathbf{x} for each j , they might also depend on covariate combinations in general to allow (smooth) interactions. While other response functions could be used, we restrict to the exponential map here, such that the model contains a geodesic model (Fletcher 2013)—the direct generalization of simple linear regression—as a special case for $h(\mathbf{x}) = \beta x_1$ with a single covariate x_1 and tangent vector β . Typically, it is assumed that h is centered such that $\mathbb{E}(h(\mathbf{X})) = 0$, and the pole $[p]$ is the overall mean of $[Y]$ defined, like the conditional mean, via residuals of mean zero.

3.1. Tensor-Product Effect Functions h_j

Scheipl, Staicu, and Greven (2015) and other authors employ Tensor-Product (TP) bases for functional additive model terms. This naturally extends to tangent space effects, which we model as

$$h_j(\mathbf{x}) = \sum_{r,l} \theta_j^{(r,l)} b_j^{(l)}(\mathbf{x}) \partial_r$$

with the TP basis given by the pair-wise products of m linearly independent tangent vectors $\partial_r \in T_{[p]}\mathcal{Y}_{/G}^*$, $r = 1, \dots, m$, and m_j basis functions $b_j^{(l)} : \mathcal{X} \rightarrow \mathbb{R}$, $l = 1, \dots, m_j$, for the j th covariate effect depending on one or more covariates. The real coefficients can be arranged as a matrix $\{\theta_j^{(r,l)}\}_{r,l} = \Theta_j \in \mathbb{R}^{m \times m_j}$. Also for infinite-dimensional $T_{[p]}\mathcal{Y}_{/G}^*$ and a general nonlinear dependence on x , a basis representation approach requires truncation to finite dimensions m and m_j in practice. Choosing the bases to capture the essential variability in the data, their size can be extended with increasing data size and computational resources.

While, in principle, the basis $\{\partial_r\}_r$ could also vary across effects $j = 1, \dots, J$, we assume a common basis for notational simplicity, which presents the typical choice. Due to the identification of $T_{[p]}\mathcal{Y}_{/G}^*$ with a subspace of the function space \mathcal{Y} , the $\{\partial_r\}_r$ may be specified using a function basis commonly used in additive models: Let $b_0^{(l)} : \mathcal{T} \rightarrow \mathbb{R}$, $l = 1, \dots, m_0$ be a basis of real functions, say a B-spline basis (other typical bases used in the literature include wavelet (Meyer et al. 2015) or FPC bases (Müller and Yao 2008)). Then we construct the tangent space basis as $\partial_r = \sum_{l=1}^{m_0} (z_p^{(l,r)} + z_p^{(m_0+l,r)} \sqrt{-1}) b_0^{(l)}$,

employing the same basis for the 1- and $\sqrt{-1}$ -dimension before transforming it with a basis transformation matrix $\mathbf{Z}_p = \{z_p^{(l,r)}\}_{l,r} \in \mathbb{R}^{2m_0 \times m}$ implementing the linear tangent space constraints $\text{Re}(\langle \partial_l, \zeta^{(r)} \rangle) = 0$ (or the empirical version) for all ∂_l and normal vectors $\zeta^{(1)}, \zeta^{(2)}, \zeta^{(3)}$ for forms and additionally $\zeta^{(4)}$ for shapes defining $T_{[p]}\mathcal{Y}_{/G}^*$ as described in Section 2. Thus, the tangent space basis dimension is $m = 2m_0 - 3$ for forms or $m = 2m_0 - 4$ for shapes (or could, in principle, be larger if the original basis already meets the constraints). For details on the construction of \mathbf{Z}_p see Section S.1.3, supplementary materials. For closed curves, we additionally choose \mathbf{Z}_p to enforce periodicity, that is, $\partial_r(t) = \partial_r(t + t_0)$ for some $t_0 \in \mathbb{R}$ (compare Hofner, Kneib, and Hothorn 2016).

Given the tangent space basis, we may now modularly specify the usual additive model basis functions $b_j^{(l)} : \mathcal{X} \rightarrow \mathbb{R}$, $l = 1, \dots, m_j$, for the j th covariate effect to obtain the full functional additive model “tool box” offered by, for example, Brockhaus, Scheipl, and Greven (2015). Typically, $b_j^{(l)}(\mathbf{x}) = b_j^{(l)}(z)$ depending on an individual covariate, say on z , in $\mathbf{x} = (\dots, z, \dots)^\top$. But for a single covariate also multiple different effects can be specified and a single interaction effect depends on multiple covariates. A linear effect—linear in the tangent space—of the form $h_j(\mathbf{x}) = \beta z$ of a scalar (typically centered) covariate z and $\beta \in T_{[p]}\mathcal{Y}_{/G}^*$ is simply implemented by a single function $b_j^{(1)}(\mathbf{x}) = z$. A smooth effect of the generic form $h_j(\mathbf{x})(t) = f(z, t)$ can be implemented by choosing, for example, a B-spline basis $b_j^{(1)}(z), \dots, b_j^{(m_j)}(z)$ (asymptotic properties of penalized B-splines and connections to kernel estimators are discussed, for example, by Wood, Pya, and Säfken (2016), Li and Ruppert (2008)). For a categorical covariate κ in \mathbf{x} , with effect $h_j(\mathbf{x}) : \{1, \dots, K\} \rightarrow T_{[p]}\mathcal{Y}_{/G}^*$, $\kappa \mapsto \beta_\kappa$, the basis $\mathbf{b}_j(\mathbf{x}) = (b_j^{(1)}(\kappa), \dots, b_j^{(m_j)}(\kappa))^\top$ maps $\kappa \mapsto \mathbf{e}_\kappa$ to a usual contrast vector \mathbf{e}_κ with the basis being of dimension $m_j = K - 1$ just as in standard linear models. Here, we typically use effect-encoding to obtain centered effects. Moreover, TP interactions of the model terms described above, as well as group-specific effects and smooth effects with additional constraints (Hofner, Kneib, and Hothorn 2016) can be specified in the model formula, relying on the `mboost` framework introduced by Hothorn et al. (2010), which also allows to define custom effect designs. For identification of an overall mean intercept $[p]$, sum-to-zero constraints yielding $\sum_{i=1}^n h_j(\mathbf{x}_i) = 0$ for observed covariates \mathbf{x}_i can be specified, and similar constraints can be used to distinguish linear from nonlinear effects and interactions from their marginal effects (Kneib, Hothorn, and Tutz 2009). Different quadratic penalties can be specified for the coefficients Θ_j , allowing to regularize high-dimensional effect bases and to balance effects of different complexity in the model fit (see, Section 4).

3.2. Tensor-Product Factorization

The multidimensional structure of the response objects makes it challenging to graphically illustrate and interpret additive model terms, in particular when it comes to nonlinear (interaction)

effects, or when effect sizes are visually small. To solve this problem, we suggest to rewrite estimated TP effects \hat{h}_j with estimated coefficient matrix $\hat{\Theta}_j$ as

$$\hat{h}_j(\mathbf{x}) = \sum_{r=1}^{m'_j} \xi_j^{(r)} \hat{h}_j^{(r)}(\mathbf{x})$$

factorized into $m'_j = \min(m_j, m_0)$ components consisting of covariate effects $\hat{h}_j^{(r)} : \mathcal{X} \rightarrow \mathbb{R}$, $r = 1, \dots, m'_j$, in corresponding orthonormal directions $\xi_j^{(r)} \in T_{[p]} \mathcal{Y}_{/G}^*$ with $\langle \xi_j^{(r)}, \xi_j^{(l)} \rangle = \mathbb{1}(r=l)$, that is, 1 if $r=l$ and 0 otherwise. Assuming $\mathbb{E}(b_j^{(l)}(\mathbf{X})^2) < \infty$, $l = 1, \dots, m_j$, for the underlying effect basis, the $\hat{h}_j^{(r)}$ are specified to achieve decreasing component variances $v_j^{(1)} \geq \dots \geq v_j^{(m'_j)} \geq 0$ given by $v_j^{(r)} = \mathbb{E}(\hat{h}_j^{(r)}(\mathbf{X})^2)$. In practice, the expectation over the covariates \mathbf{X} and the inner product $\langle \cdot, \cdot \rangle$ are replaced by empirical analogs (compare Corollary 3, supplementary materials). Due to orthonormality of the $\xi_j^{(r)}$, the component variances add up to the total predictor variance $\sum_{r=1}^{m'_j} v_j^{(r)} = v_j = \mathbb{E}(\langle \hat{h}_j(\mathbf{X}), \hat{h}_j(\mathbf{X}) \rangle)$. Moreover, the TP factorization is optimally concentrated in the first components in the sense that for any $l \leq m'_j$ there is no sequence of $\xi_*^{(r)} \in \mathcal{Y}$ and $\hat{h}_*^{(r)} : \mathcal{X} \rightarrow \mathbb{R}$, such that $\mathbb{E}(\|\hat{h}_j(\mathbf{X}) - \sum_{r=1}^l \xi_*^{(r)} \hat{h}_*^{(r)}(\mathbf{X})\|^2) < \mathbb{E}(\|\hat{h}_j(\mathbf{X}) - \sum_{r=1}^l \xi_j^{(r)} \hat{h}_j^{(r)}(\mathbf{X})\|^2)$, that is, the series of the first l components yields the best rank l approximation of \hat{h}_j . The factorization relies on SVD of (a transformed version of) the coefficient matrix $\hat{\Theta}_j$ and the fact that it is well-defined is a variant of the Eckart-Young-Mirsky theorem (proof in Section S.2, supplementary materials).

Particularly when large shares of the predictor variance are explained by the first component(s), the decomposition facilitates graphical illustration and interpretation: choosing a suitable constant $\tau \neq 0$, an effect direction $\xi_j^{(r)}$ can be visualized by plotting the pole representative p together with $\text{Exp}_p(\tau \xi_j^{(r)})$ on the level of curves, while accordingly rescaled $\frac{1}{\tau} \hat{h}_j^{(r)}(\mathbf{x})$ is displayed separately in a standard scalar effect plot. Adjusting τ offers an important degree of freedom for visualizing $\xi_j^{(r)}$ on an intuitively accessible scale while faithfully depicting $\xi_j^{(r)} \hat{h}_j^{(r)}(\mathbf{x})$. When based on the same τ , different covariate effects can be compared across the plots sharing the same scale. We suggest $\tau = \max_j \sqrt{v_j}$, the maximum total predictor standard deviation of an effect, as a good first choice.

Besides factorizing effects separately, it can also be helpful to apply TP factorization to the joint additive predictor, yielding

$$h(\mathbf{x}) = \sum_{r=1}^{m'} \xi^{(r)} \hat{h}^{(r)}(\mathbf{x}) = \sum_{r=1}^{m'} \xi^{(r)} \left(\hat{h}_1^{(r)}(\mathbf{x}) + \dots + \hat{h}_j^{(r)}(\mathbf{x}) \right),$$

with $m' = \min(\sum_j m_j, m)$ and again $\xi^{(r)} \in T_{[p]} \mathcal{Y}_{/G}^*$ orthonormal and the corresponding variance concentration in the first components, but now determined w.r.t. entire additive predictors $\hat{h}^{(r)} = \sum_{j=1}^J \hat{h}_j^{(r)}$ spanned by all covariate basis functions in

the predictor. In this representation, the first component yields a geodesic additive model approximation where the predictor moves along a geodesic line $c(\tau) = \text{Exp}_{[p]}(\xi^{(1)} \tau)$ with the signed distance $\tau \in \mathbb{R}$ from $[p]$, modeled by a scalar additive predictor $\hat{h}^{(1)}(\mathbf{x})$ composed of covariate effects analogous to the original model predictor. In Section 5, we illustrate its potential in three different scenarios.

4. Component-Wise Riemannian L_2 -Boosting

Component-wise gradient boosting (e.g., Hothorn et al. 2010) is a step-wise model fitting procedure accumulating predictors from smaller models, so called base-learners, to built an ensemble predictor aiming at minimizing a mean loss function. To this end, the base-learners are fit (via least squares) to the negative gradient of the loss function in each step and the best fitting base-learner is added to the current ensemble predictor. Due to its versatile applicability, inherent model selection, and slow over-fitting behavior, boosting has proven useful in various contexts (Mayr et al. 2014). Boosting with respect to the least squares loss function $\ell(y, \mu) = \frac{1}{2}(y - \mu)^2$, $y, \mu \in \mathbb{R}$, is typically referred to as L_2 -Boosting and simplifies to repeated refitting of residuals $\varepsilon = y - \mu = -\nabla_{\mu} \ell(y, \mu)$ corresponding to the negative gradient of the loss function. For L_2 -Boosting with a single learner, Bühlmann and Yu (2003) show how fast bias decay and slow variance increase over the boosting iterations suggest stopping the algorithm early before approaching the ordinary (penalized) least squares estimator. Lutz and Bühlmann (2006) prove consistency of component-wise L^2 -Boosting in a high-dimensional multivariate response linear regression setting and Stöcker et al. (2021) illustrate in extensive simulation studies how stopping the boosting algorithm early based on curve-wise cross-validation applies desired regularization when fitting (even highly autocorrelated) functional responses with parameter-intense additive model base-learners and, thus, leads to good estimates even in challenging scenarios.

When generalizing to least squares on Riemannian manifolds with the loss $\frac{1}{2} d^2([y], [\mu])$ given by the squared geodesic distance, the negative gradient $-\nabla_{[\mu]} \frac{1}{2} d^2([y], [\mu]) = \text{Log}_{[\mu]}([y]) = \varepsilon_{[\mu]}$ (compare e.g., Pennec 2006) corresponds to the local residuals $\varepsilon_{[\mu]}$ defined in Section 3. This analogy to L_2 -Boosting motivates the presented generalization where local residuals are further transported to residuals ε in a common linear space.

Consider the pole $[p]$ known and fixed for now. Assuming its existence, we aim to minimize the population mean loss

$$\sigma^2(h) = \mathbb{E} \left(d^2 \left([Y], \text{Exp}_{[p]}(h(\mathbf{X})) \right) \right)$$

with the point-wise minimizer $h^*(\mathbf{x}) = \underset{h: \mathcal{X} \rightarrow T_{[p]} \mathcal{Y}_{/G}^*}{\text{argmin}} \mathbb{E}(d^2([Y],$

$\text{Exp}_{[p]}(h(\mathbf{X})) \mid \mathbf{X} = \mathbf{x})$ minimizing the conditional expected squared distance. Fixing a covariate constellation $\mathbf{x} \in \mathcal{X}$, the prediction $[\mu] = \text{Exp}_{[p]}(h^*(\mathbf{x}))$ corresponds to the Fréchet mean (Karcher 1977) of $[Y]$ conditional on $\mathbf{X} = \mathbf{x}$. In a finite-dimensional context, Pennec (2006) show that $\mathbb{E}(\varepsilon_{[\mu]}) = 0$ for a Fréchet mean $[\mu]$ if residuals $\varepsilon_{[\mu]}$ are uniquely defined with probability one. This indicates the connection to our residual based model formulation in Section 3. We fit the model by

reducing the empirical mean loss $\hat{\sigma}^2(h) = \frac{1}{n} \sum_{i=1}^n d_i^2([y_i], \text{Exp}_{[p]}(h(\mathbf{x}_i)))$, where we replace the population mean by the sample mean and compute the geodesic distances d_i with respect to the inner products $\langle \cdot, \cdot \rangle_i$ defined for the respective evaluations of y_i .

A base-learner corresponds to a covariate effect $h_j(\mathbf{x}) = \sum_{r,l} \theta_j^{(r,l)} b_j^{(l)}(\mathbf{x}) \partial_r$, $\Theta_j = \{\theta_j^{(r,l)}\}_{r,l}$, which is repeatedly fit to the transported residuals $\epsilon_1, \dots, \epsilon_n$ by penalized least-squares (PLS) minimizing $\sum_{i=1}^n \|\epsilon_i - h_j(\mathbf{x}_i)\|_i^2 + \lambda_j \text{tr}(\Theta_j \mathbf{P}_j \Theta_j^\top) + \lambda \text{tr}(\Theta^\top \mathbf{P} \Theta)$. Via the penalty parameters $\lambda_j, \lambda \geq 0$ the effective degrees of freedom of the base-learners are controlled (Hofner et al. 2011) to achieve a balanced “fair” base-learner selection despite the typically large and varying number of coefficients involved in the TP effects. The symmetric penalty matrices $\mathbf{P}_j \in \mathbb{R}^{m_j \times m_j}$ and $\mathbf{P} \in \mathbb{R}^{m \times m}$ (imposing, e.g., a second-order difference penalty for B-splines in either direction) can equivalently be arranged as a $m_j m \times m_j m$ penalty matrix $\mathbf{R}_j = \lambda_j (\mathbf{P}_j \otimes \mathbf{I}_m) + \lambda (\mathbf{I}_{m_j} \otimes \mathbf{P})$ for the vectorized coefficients $\text{vec}(\Theta_j) = (\theta_j^{(1,1)}, \dots, \theta_j^{(m,1)}, \dots, \theta_j^{(m,m_j)})^\top$, where \otimes denotes the Kronecker product. The standard PLS estimator is then given by $\text{vec}(\hat{\Theta}_j) = (\Psi_j + \mathbf{R}_j)^{-1} \psi_j$ with $\Psi_j = \sum_{i=1}^n \left\{ \text{Re} \left(\langle b_j^{(l)}(\mathbf{x}_i) \partial_r, b_j^{(l')}(\mathbf{x}_i) \partial_{r'} \rangle_i \right) \right\}_{(r,l)=(1,1), \dots, (m,1), \dots, (m,m_j)} \in \mathbb{R}^{m_j \times m_j}$ and $\psi_j = \sum_{i=1}^n \left\{ \text{Re} \left(\langle b_j^{(l)}(\mathbf{x}_i) \partial_r, \epsilon_i \rangle_i \right) \right\}_{(r,l)=(1,1), \dots, (m,1), \dots, (m,m_j)} \in \mathbb{R}^{m_j}$. In a regular design, using the functional linear array model (Brockhaus, Scheipl, and Greven 2015) can save memory and computation time by avoiding construction of the complete matrices. The basis construction of $\{\partial_r\}_r$ via a transformation matrix \mathbf{Z}_p (Section 3.1) is reflected in the penalty by setting $\mathbf{P} = \mathbf{Z}_p^\top (\mathbf{I}_2 \otimes \mathbf{P}_0) \mathbf{Z}_p$ with \mathbf{P}_0 the penalty matrix for the un-transformed basis $\{b_0^{(r)}\}_r$.

In each iteration of the proposed Algorithm 1, the best-performing base-learner is added to the current ensemble additive predictor $h(\mathbf{x})$ after multiplying it with a step-length parameter $\eta \in (0, 1]$. Due to the additive model structure this corresponds to a coefficient update of the selected covariate effect. Accordingly, after repeated selection, the effective degrees of freedom of a covariate effect, in general, exceed the degrees specified for the base-learner. They are successively adjusted to the data. To avoid over-fitting, the algorithm is typically stopped early before reaching a minimum of the empirical mean loss. The stopping iteration is determined, for example, by resampling strategies such as bootstrapping or cross-validation on the level of shapes/forms.

The pole $[p]$ is, in fact, usually not a priori available. Instead we typically assume $[p] = \underset{q \in \mathcal{Y}^*}{\text{argmin}} \mathbb{E} (d^2([Y], [q]))$ is the overall Fréchet mean, also often referred to as *Riemannian center of mass* for Riemannian manifolds or as *Procrustes mean* in shape analysis (Dryden and Mardia 2016). Here, we estimate it as $[p] = \text{Exp}_{[p_0]}(h_0)$ in a preceding Riemannian L^2 -Boosting routine. The constant effect $h_0 \in T_{[p_0]} \mathcal{Y}_G^*$ in the intercept-only special case of our model is estimated with Algorithm 1 based on a preliminary pole $[p_0] \in \mathcal{Y}_G^*$. For shapes and forms, a good candidate for p_0 can be obtained as the standard functional

Algorithm 1: Component-wise Riemannian L^2 -Boosting

```

# Initialization:
Geometry      : specify geometry (shape/form)
                  and pole representative  $p$ 
Hyper-parameters: Step-length  $\eta \in (0, 1]$ , number of
                  boosting iterations
Base-learners :  $h_j(\mathbf{x})$  with penalty matrix  $\mathbf{R}_j$  and
                  initial coefficient matrix  $\Theta_j = \mathbf{0}$ 
for  $j = 1$  to  $J$  do      # Prepare penalized
least-squares (PLS)
    # set up  $m_j m \times m_j m$  matrix:  $\Psi_j \leftarrow \sum_{i=1}^n$ 
     $\left\{ \text{Re} \left( \langle b_j^{(l)}(\mathbf{x}_i) \partial_r, b_j^{(l')}( \mathbf{x}_i) \partial_{r'} \rangle_i \right) \right\}_{(r,l)=(1,1), \dots, (m,1), \dots, (m,m_j)}$ 
     $\left. \left. \left. \right\}_{(r',l')=(1,1), \dots, (m,1), \dots, (m,m_j)}$ 
end
repeat      # boosting steps
for  $i = 1, \dots, n$  do      # Compute current
transported residuals
     $[\mu_i] \leftarrow \text{Exp}_{[p]}(h(\mathbf{x}_i))$ 
     $\epsilon_{[\mu_i]} \leftarrow \text{Log}_{[\mu_i]}([y_i])$ 
     $\epsilon_i \leftarrow \text{Transp}_{[\mu_i], [p]}(\epsilon_{[\mu_i]})$ 
end
for  $j = 1, \dots, J$  do      # PLS fit to
residuals
    #  $m_j m$  vector:  $\psi_j \leftarrow$ 
     $\sum_{i=1}^n \left\{ \text{Re} \left( \langle b_j^{(l)}(\mathbf{x}_i) \partial_r, \epsilon_i \rangle_i \right) \right\}_{(r,l)=(1,1), \dots, (m,1), \dots, (m,m_j)}$ 
     $\hat{\Theta}_j = \{\hat{\theta}_j^{(r,l)}\}_{r,l} \leftarrow \text{Solve} \left($ 
     $(\Psi_j + \mathbf{R}_j) \text{vec}(\Theta) = \psi_j$ 
end
     $\hat{j} \leftarrow \underset{j \in \{1, \dots, J\}}{\text{argmin}} \sum_{i=1}^n \|\epsilon_i - \sum_{r,l} \hat{\theta}_j^{(r,l)} b_j^{(l)}(\mathbf{x}) \partial_r\|_i^2;$ 
    # Select base-learner
     $\Theta_j \leftarrow \Theta_j + \eta \hat{\Theta}_j;$       # Update selected
model coefficients
until Stopping criterion (e.g., minimal cross-validation
error)
    
```

mean of a reasonably well aligned sample $y_1, \dots, y_n \in \mathcal{Y}$ of representatives.

The proposed Riemannian L_2 -Boosting algorithm is available in the R (R Core Team 2018) package `manifoldboost` (github.com/Almond-S/manifoldboost). The implementation is based on the package `FDboost` (Brockhaus, Rügamer, and Greven 2020), which is in turn based on the model-based boosting package `mboost` (Hothorn et al. 2010).

5. Applications and Simulation

5.1. Shape Differences in Astragali of Wild and Domesticated Sheep

In a geometric morphometric study, Pöllath, Schafberg, and Peters (2019) investigate shapes of sheep astragali (ankle bones)

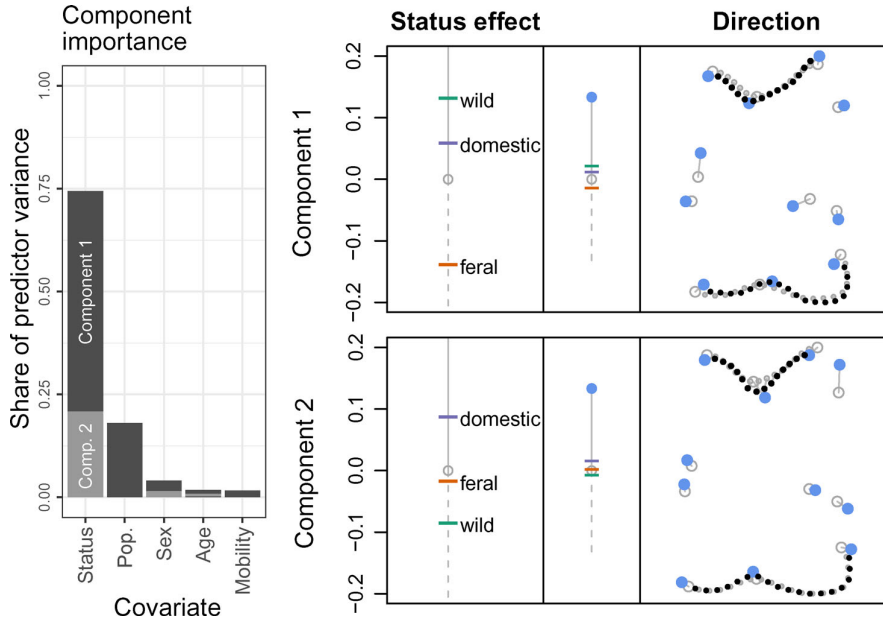


Figure 2. *Left:* Shares of different factorized covariate effects in the total predictor variance. *Right:* Factorized effect plots showing the two components of the status effect (rows): in the *right column*, the two first directions $\xi_1^{(1)}, \xi_1^{(2)} \in T_{[p]}\mathcal{Y}_{\text{Trl} \times \text{Rot} \times \text{Scl}}^*$ are visualized via line-segments originating at the overall mean shape (empty circles) and ending in the shape resulting from moving 1 unit into the target direction (solid circles; large: landmarks; small: semi-landmarks along the outline); in the *left column*, the status effect in the respective direction is depicted. As illustrated in the *middle plot*, an effect of 1 would correspond to the full extend of the direction shown to the right.

to understand the influence of different living conditions on the micromorphology of the skeleton. Based on a total of $n = 163$ shapes recorded by Pöllath, Schafberg, and Peters (2019), we model the astragalus shape in dependence on different variables, including domestication status (wild/feral/domesticated), sex (female/male/NA), age (juvenile/subadult/adult/NA), and mobility (confined/pastured/free) of the animals as categorical covariates. The sample comprises sheep of four different populations: Asiatic wild sheep (Field Museum, Chicago; Lay 1967; Zeder 2006), feral Soay sheep (British Natural History Museum, London; Clutton-Brock et al. 1990), and domestic sheep of the Karakul and Marsch breed (Museum of Livestock Sciences, Halle (Saale); Schafberg and Wussow 2010). Table S1 in Section S.3, supplementary materials shows the distribution of available covariates within the populations. Each sheep astragalus shape, $i = 1, \dots, n$, is represented by a configuration composed of 11 selected landmarks in a vector $\mathbf{y}_i^{\text{lm}} \in \mathbb{C}^{11}$ and two vectors of sliding semi-landmarks $\mathbf{y}_i^{\text{c}1} \in \mathbb{C}^{14}$ and $\mathbf{y}_i^{\text{c}2} \in \mathbb{C}^{18}$ evaluated along two outline curve segments, marked on a 2D image of the bone (dorsal view). Several example configurations are displayed in Figure S1, supplementary materials. In general, we could separately specify smooth function bases for the outline segments $y_i^{\text{c}1}$ and $y_i^{\text{c}2}$, respectively. Due to their systematic recording, we assume, however, that not only landmarks but also semi-landmarks are regularly observed on a fixed grid, and refrain from using smooth function bases for simplicity. Accordingly, shape configurations can directly be identified with their evaluation vectors $\mathbf{y}_i = (\mathbf{y}_i^{\text{lm}\top}, \mathbf{y}_i^{\text{c}1\top}, \mathbf{y}_i^{\text{c}2\top})^\top \in \mathbb{C}^{43} = \mathcal{Y}$, and the geometry of the response space $\mathcal{Y}_{\text{Trl} \times \text{Rot} \times \text{Scl}}^*$ widely corresponds to the classic Kendall's shape space geometry, with the difference that, considering landmarks more descriptive than single semi-landmarks, we choose a weighted inner product

$\langle \mathbf{y}_i, \mathbf{y}_i' \rangle = \mathbf{y}_i^\dagger \mathbf{W} \mathbf{y}_i'$ with diagonal weight matrix \mathbf{W} with diagonal $(\mathbf{1}_{11}^\top, \frac{3}{14} \mathbf{1}_{14}^\top, \frac{3}{18} \mathbf{1}_{18}^\top)^\top$ assigning the weight of three landmarks to each outline segment. We model the astragalus shapes $[\mathbf{y}_i] \in \mathcal{Y}_{\text{Trl} \times \text{Rot} \times \text{Scl}}^*$ as

$$[\boldsymbol{\mu}_i] = \text{Exp}_{[p]} \left(\boldsymbol{\beta}_{\text{status}_i} + \boldsymbol{\beta}_{\text{pop}_i} + \boldsymbol{\beta}_{\text{age}_i} + \boldsymbol{\beta}_{\text{sex}_i} + \boldsymbol{\beta}_{\text{mobility}_i} \right)$$

with the pole $[p] \in \mathcal{Y}_G^*$ specified as overall mean and the conditional mean $[\boldsymbol{\mu}_i] \in \mathcal{Y}_{\text{Trl} \times \text{Rot} \times \text{Scl}}^*$ depending on the effect coded covariate effects $x_{ij} \mapsto \boldsymbol{\beta}_{x_{ij}} \in T_{[p]}\mathcal{Y}_{\text{Trl} \times \text{Rot} \times \text{Scl}}^*$. For identifiability, the population and mobility effects are centered around the status effect, as we only have data on different populations/mobility levels for domesticated sheep. All base-learners are regularized to one degree of freedom by employing ridge penalties for the coefficients of the covariate bases $\{b_j^{(l)}\}_l$ while the coefficients of the response basis (the standard basis for \mathbb{C}^{43}) are left un-penalized. With a step-length of $\eta = 0.1$, 10-fold shape-wise cross-validation suggests early stopping after 89 boosting iterations. Due to the regular design, we can make use of the functional linear array model (Brockhaus, Scheipl, and Greven 2015) for saving computation time and memory, which lead to 8 sec of initial model fit followed by 47 sec of cross-validation. To interpret the categorical covariate effects, we rely on TP factorization (Figure 2). The first component of the status effect explains about 2/3 of the variance of the status effect and over 50% of the cumulative effect variance in the model. In that main direction, the effect of *feral* is not located between *wild* and *domestic*, as might be naively expected. By contrast, the second component of the effect seems to reflect the expected order and still explains a considerable amount of variance. Similar to Pöllath, Schafberg, and Peters (2019), we find little influence of

age, sex, and mobility on the astragalus shape. Yet, all covariates were selected by the boosting algorithm.

Visually, differences in estimated mean shapes are rather small, which is, in our experience, quite usual for shape data. With differences in size, rotation and translation excluded by definition, only comparably small variance remains in the observed shapes. Nonetheless, TP factorization provides accessible visualization of the effect directions and allows to partially order the effect levels in each direction.

5.2. Cellular Potts Model Parameter Effects on Cell Form

The stochastic biophysical model proposed by Thüroff et al. (2019), a Cellular Potts Model (CPM), simulates migration dynamics of cells (e.g., wound healing or metastasis) in two dimensions. The progression of simulated cells is the result of many consecutive local elementary events sampled with a Metropolis-algorithm according to a Hamiltonian. Different parameters controlling the Hamiltonian have to be calibrated to match real live cell properties (Schaffer 2021). Considering whole cells, parameter implications on the cell form are not obvious. To provide additional insights, we model the cell form in dependence on four CPM parameters considered particularly relevant: the bulk stiffness x_{i1} , membrane stiffness x_{i2} , substrate adhesion x_{i3} , and signaling radius x_{i4} are subsumed in a vector \mathbf{x}_i of metric covariates for $i = 1, \dots, n$. Corresponding sampled cell outlines y_i were provided by Sophia Schaffer in the context of Schaffer (2021), who ran underlying CPM simulations and extracted outlines. Deriving the intrinsic orientation of the cells from their movement trajectories, we parameterize $y_i : [0, 1] \rightarrow \mathbb{C}$, clockwise relative to arc-length such that $y_i(0) = y_i(1)$ points into the movement direction of the barycenter of the cell. With an average of $k = \frac{1}{n} \sum_{i=1}^n k_i \approx 43$ samples per curve (after sub-sampling preserving 95% of their inherent variation, as described in Volkman et al. 2021, supplement), the evaluation vectors $\mathbf{y}_i \in \mathbb{C}^{k_i}$ are equipped with an inner-product implementing trapezoidal rule integration weights. Example cell outlines are depicted in Figure S4, supplementary materials. The results shown below are based on cell samples obtained from 30 different CPM parameter configurations. For each configuration, 33 out of 10.000 Monte Carlo samples were extracted as approximately independent. This yields a dataset of $n = 990 = 30 \times 33$ cell outlines.

As positioning of the irregularly sampled cell outlines y_i , $i = 1, \dots, n$, in the coordinate system is arbitrary, we model the cell forms $[y_i] \in \mathcal{Y}_{\text{Tri} + \text{Rot}}^*$. Their estimated overall form mean $[p]$ serves as pole in the additive model

$$\begin{aligned} [\mu_i] &= \text{Exp}_{[p]}(h(\mathbf{x}_i)) \\ &= \text{Exp}_{[p]} \left(\sum_j \beta_j x_{ij} + \sum_j f_j(x_{ij}) + \sum_{j \neq j'} f_{jj'}(x_{ij}, x_{ij'}) \right) \end{aligned}$$

where the conditional form mean $[\mu_i]$ is modeled in dependence on tangent-space linear effects with coefficients $\beta_j \in T_{[p]} \mathcal{Y}_{\text{Tri} + \text{Rot}}$ and nonlinear smooth effects f_j for covariate $j = 1, \dots, 4$, as well as smooth interaction effects $f_{jj'}$ for each pair of covariates $j \neq j'$. All involved (effect) functions are modeled via a cyclic cubic P-spline basis $\{b_0^{(r)}\}_r$ with 7 (inner) knots and a ridge penalty, and quadratic P-splines with 4 knots for the

covariates x_{ij} equipped with a second-order difference penalty for the f_j and ridge penalties for interactions. Covariate effects are mean centered and interaction effects $f_{jj'}(x_j, x_{j'})$ are centered around their marginal effects $f_j(x_j), f_{j'}(x_{j'})$, which are in turn centered around the linear effects $\beta_j x_j$ and $\beta_{j'} x_{j'}$, respectively. Resulting predictor terms involve 69 (linear effect) to 1173 (interaction) basis coefficients but are penalized to a common degree of freedom of 2 to ensure a fair base-learner selection. We fit the model with a step-size of $\eta = 0.25$ and stop after 2000 boosting iterations observing no further meaningful risk reduction, since no need for early-stopping is indicated by 10-fold form-wise cross-validation. Due to the increased number of data points and coefficients, the irregular design, and the increased number of iterations, the model fit takes considerably longer than in Section 5.1, with about 50 initial minutes followed by 8 hr of cross-validation. However, as usual in boosting, model updates are large in the beginning and only marginal in later iterations, such that fits after 1000 or 500 iterations would already yield very similar results.

Observing that the most relevant components point into similar directions, we jointly factorize the predictor as $\hat{h}(\mathbf{x}_i) = \sum_r \xi^{(r)} \hat{h}^{(r)}(\mathbf{x}_i)$ with TP factorization. The first component explains about 93% of the total predictor variance (Figure S3, supplementary materials), indicating that, post-hoc, a good share of the model can be reduced to the geodesic model $[\hat{\mu}_i] = \text{Exp}_{[p]}(\xi^{(1)} \hat{h}^{(1)}(\mathbf{x}_i))$ illustrated in Figure 3. A positive effect in the direction $\xi^{(1)}$ makes cells larger and more keratocyte / croissant shaped, a negative effect—pointing into the opposite direction—makes them smaller and more mesenchymal shaped / elongated. The bulk stiffness x_{i1} turns out to present the most important driving factor behind the cell form, explaining over 75% of the cumulative variance of the effects (Figure S2, supplementary materials). Around 80% of its effect are explained by the linear term reflecting gradual shrinkage at the side of the cells with increasing bulk stiffness.

5.3. Realistic Shape and form Simulation Studies

To evaluate the proposed approach, we conduct simulation studies for both shape and form regression for irregular curves. We compare sample sizes $n \in \{54, 162\}$ and average grid sizes $k = \frac{1}{n} \sum_{i=1}^n k_i \in \{40, 100\}$ as well as an extreme case with $k_i = 3$ for each curve but $n = 720$, that is, where only random triangles are observed (yet, with known parameterization over $[0, 1]$). We additionally investigate the influence of nuisance effects and compare different inner product weights. While important results are summarized in the following, comprehensive visualizations can be found in Section S.5, supplementary materials.

Simulation design: We simulate models of the form $[\mu] = \text{Exp}_{[p]}(\beta_\kappa + f_1(z_1))$ with overall mean $[p]$, a binary effect with levels $\kappa \in \{0, 1\}$ and a smooth effect of $z_1 \in [-60, 60]$. We choose a cyclic cubic B-spline basis with 27 knots for $T_{[p]} \mathcal{Y}_{\text{G}}$, placing them irregularly at 1/27-quantiles of unit-speed parameterization time-points of the curves. Cubic B-splines with four regularly placed knots are used for covariates in smooth effects. True models are based on the `bot` dataset from R package `Momocs` (Bonhomme et al. 2014) comprising outlines of 20 beer ($\kappa = 0$) and 20 whiskey ($\kappa = 1$) bottles of different brands.

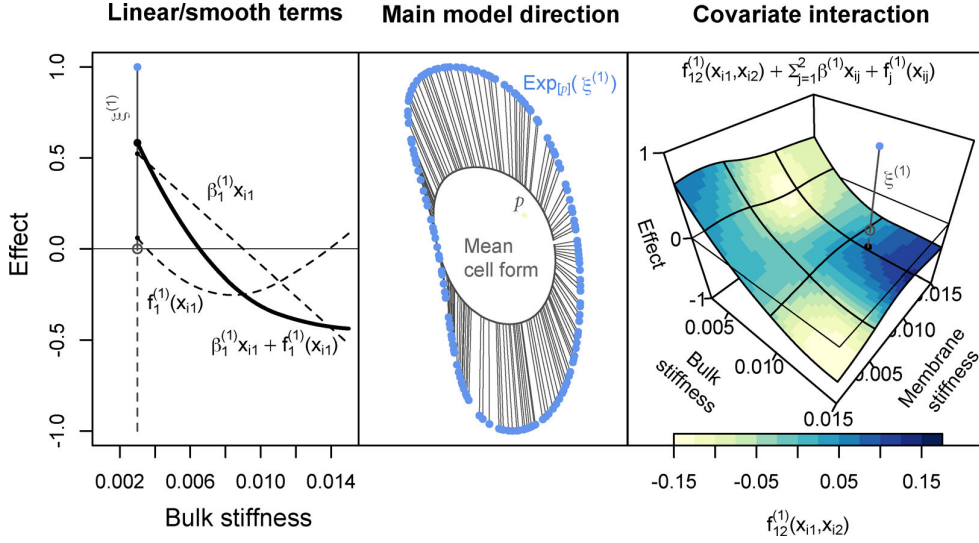


Figure 3. Center: the main direction $\xi^{(1)}$ of the model illustrated as vectors pointing from the overall mean cell form $[p]$ (gray curve) to the form $\text{Exp}_{[p]}(\xi^{(1)})$ (filled dots), which are both oriented as cells migrating rightwards. Left: Effects of the bulk stiffness x_{i1} into the direction $\xi^{(1)}$. A vertical line from 0, corresponding to $[p]$, to 1, corresponding to the full extent of $\xi^{(1)}$, underlines the connection between the plots and helps to visually assess the amount of change for a given value of x_{i1} . Right: The overall effect of x_{i1} and membrane stiffness x_{i2} , comprising linear, smooth and interaction effects, as a 3D surface plot. The heat map plotted on the surface shows only the interaction effect $f_{12}^{(1)}(x_{i1}, x_{i2})$ illustrating deviations from the marginal effects, which are of particular interest for CPM calibration.

A smooth effect is induced by the 2D viewing transformations resulting from tilting the planar outlines in a 3D coordinate system along their longitudinal axis by an angle of up to 60 degree toward the viewer ($z_1 = 60$) and away ($z_1 = -60$) (i.e., in a way not captured by 2D rotation invariance). Establishing ground truth models based on a fit to the bottle data, we simulate new responses $[y_1], \dots, [y_n]$ via residual resampling (Section S.5, supplementary materials) to preserve realistic autocorrelation. Subsequently, we randomly translate, rotate and scale $y_1, \dots, y_n \in \mathcal{Y}$ somewhat around the aligned shape/form representatives to obtain realistic samples.

The implied residual variance $\frac{1}{n} \sum_{i=1}^n \|\epsilon_i\|_i^2 = \frac{1}{n} \sum_{i=1}^n d_i^2$ ($[y_i], [\mu_i]$) on simulated datasets ranges around 105% of the predictor variance $\frac{1}{n} \sum_{i=1}^n \|h(\mathbf{x}_i)\|_i^2 = \frac{1}{n} \sum_{i=1}^n d_i^2(\mu_i, [p])$ in the form scenario and around 65% in the shape scenario. All simulations were repeated 100 times, fitting models with the model terms specified above and three additional nuisance effects: a linear effect β_{z_1} (orthogonal to $f_1(z_1)$), an effect f_2 of the same structure as f_1 but depending on an independently uniformly drawn variable z_2 , and a constant effect $h_0 \in T_{[p]}\mathcal{Y}_G^*$ to test centering around $[p]$. Base-learners are regularized to 4 degrees of freedom (step-length $\eta = 0.1$). Early-stopping is based on 10-fold cross-validation.

Form scenario: In the form scenario, the smooth covariate effect f_1 offers a particularly clear interpretation. TP factorization decomposes the true effect into its two relevant components, where the first (major) component corresponds to the bare projection of the tilted outline in 3D into the 2D image plane and the second to additional perspective transformations (Figure 4). For this effect, we observe a median relative mean squared error $\text{rMSE}(\hat{h}_j) = \sum_{i=1}^n \|\hat{h}_j(\mathbf{x}_i) - h_j(\mathbf{x}_i)\|_i^2 / \sum_{i=1}^n \|h(\mathbf{x}_i)\|_i^2$ of about 3.7% of the total predictor variance for small data settings with $n = 54$ and $k = 100$ (5.9% with $k = 40$), which reduces to 1.5% for $n = 162$ (for both

$k = 40$ and $k = 100$). It is typical for functional data that, from a certain point, adding more (highly correlated) evaluations per curve leads to distinctly less improvement in the model fit than adding further observations (compare, e.g., also Stöcker et al. 2021). In the extreme $k_i = 3$ scenario, we obtain an rMSE of around 15%, which is not surprisingly considerably higher than for the moderate settings above. Even in this extreme setting (Figure 4), the effect directions are captured well, while the size of the effect is underestimated. Rotation alignment based on only three points (which are randomly distributed along the curves) might considerably differ from the full curve alignment, and averaging over these sub-optimal alignments masks the full extend of the effect. Still, results are very good given the sparsity of information in this case. Having a simpler form, the binary effect β_κ is also estimated more accurately with an rMSE of around 1.5% for $n = 54$, $k = 100$ (1.9% for $k = 40$) and less than 0.8% for $n = 162$ (for both $k = 40$ and $k = 100$). The pole estimation accuracy varies on a similar scale.

Shape scenario: Qualitatively, the shape scenario shows a similar picture. For $k = 40$, we observe median rMSEs of 2.8% ($n = 54$) and 2.2% ($n = 162$) for $f_1(z_1)$, and 1.5% and 0.6% for the binary effect β_κ . For $k = 100$, accuracy is again slightly higher.

Nuisance effects and integration weights: Nuisance effects in the model where generally rarely selected and, if selected at all, only lead to a marginal loss in accuracy. The constant effect is only selected sometimes in the extreme triangle scenarios, when pole estimation is difficult. We refer to Brockhaus et al. (2017), who perform gradient boosting with functional responses and a large number of covariate effects with stability selection, for simulations with larger numbers of nuisance effects and further discussion in a related context, as variable selection is not our main focus here. Finally, simulations indicate that inner product weights implementing a trapezoidal rule for numerical

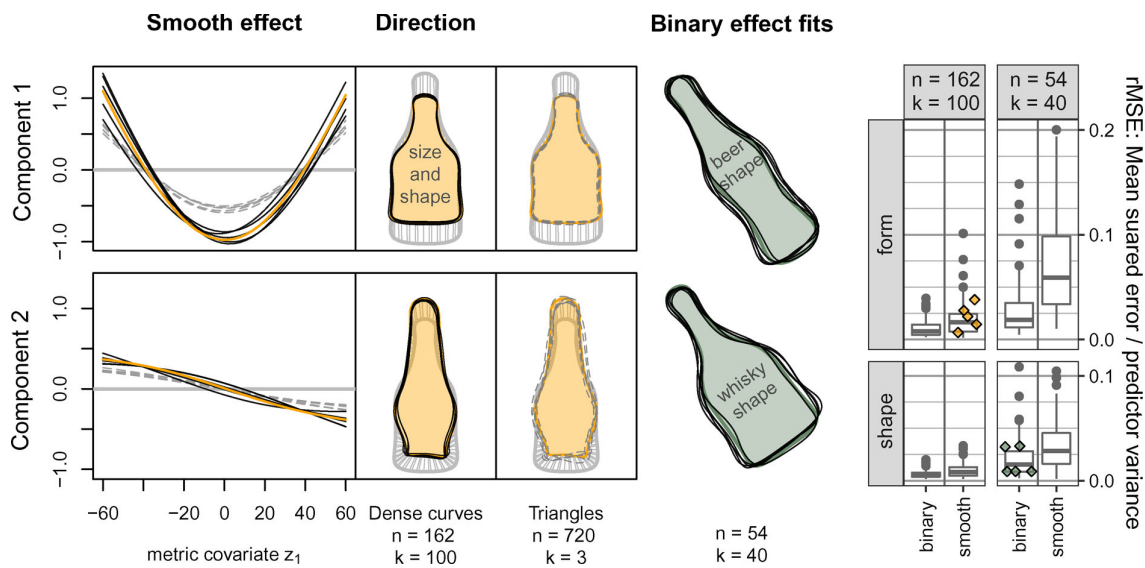


Figure 4. Left: First (row 1) and second (row 2) main components of the smooth effect $f_1(z_1)$ in the form scenario obtained from TP factorization. Normalized component directions are visualized as bottle outlines after transporting them to the true pole (gray solid outline). Underlying truth (solid lines / shaded areas) are plotted together with five example estimates for $n = 162$ and $k = 100$ (black solid lines) and the extremely sparse $k_i = 3$ setting (gray dashed lines). Center: Conditional means for both bottle types with fixed metric covariate $z_1 = 0$ in the shape scenario with $n = 54$ and $k = 40$. Five example estimates (black solid outlines) are plotted in front of the underlying truth (shaded areas). Right: rMSE of shown example estimates (jittered diamonds) contextualized with boxplots of rMSE distributions observed in respective simulation scenarios.

integration are slightly preferable for typical grid sizes ($k = 40, 100$), whereas weights of $1/k_i$ equal over all grid points within a curve gave slightly better results in the extreme $k_i = 3$ settings.

All in all, the simulations show that Riemannian L_2 -Boosting can adequately fit both shape and form models in a realistic scenario and captures effects reasonably well even for a comparably small number of sampled outlines or evaluations per outline.

6. Discussion and Outlook

Compared to existing (landmark) shape regression models, the presented approach extends linear predictors to more general additive predictors including also, for example, smooth nonlinear model terms and interactions, and yields the first regression approach for functional shape as well as form responses. Moreover, we propose novel visualizations based on TP factorization that, similar to FPC analysis, enable a systematic decomposition of the variability explained by an additive effect on tangent space level. Yielding meaningful coordinates for model effects, its potential for visualization will be useful also for FAMs in linear spaces and also beyond our model framework, such as we exemplarily illustrate for the nonparametric approach of Jeon and Park (2020) in Section S.8, supplementary materials.

Instead of operating on the original evaluations $y_i \in \mathbb{C}^{k_i}$ of response curves y_i as in all applications above, another frequently used approach expands $y_i, i = 1, \dots, n$, in a common basis first, before carrying out statistical analysis on coefficient vectors (compare Ramsay and Silverman (2005), Morris (2015), and Müller and Yao (2008) for smoothing spline, wavelet or FPC representations in FDA or Bonhomme et al. (2014) in shape analysis). Shape/form regression on the coefficients is, in fact, a special case of our approach, where the inner product is evaluated on the coefficients instead of evaluations (Section S.6, supplementary materials).

The proposed model is motivated by geodesic regression. However, in the multiple linear predictor, a linear effect of a single covariate does, in general, not describe a geodesic for fixed nonzero values of other covariate effects. Or put differently, $\text{Exp}_{[p]}(h_1 + h_2) \neq \text{Exp}_{\text{Exp}_{[p]}(h_1)}(h_2) \neq \text{Exp}_{\text{Exp}_{[p]}(h_2)}(h_1)$ in general. Thus, hierarchical geodesic effects of the form $\text{Exp}_{\text{Exp}_{[p]}(h_1)}(h_2)$, relevant, i.e., in mixed models for hierarchical/longitudinal study designs (Kim et al. 2017), present an interesting future extension of our model. Moreover, an “elastic” extension based on the square-root-velocity framework (Srivastava and Klassen 2016) presents a promising direction for future research, as do other manifold responses.

Supplementary Materials

Supplementary material with further details is provided in an online supplement.

Acknowledgments

We sincerely thank Nadja Pöllath for providing carefully recorded sheep astragalus data and important insights and comments, and Sophia Schaffer for running and discussing cell simulations and providing fully processed cell outlines.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

We gratefully acknowledge funding by grant GR 3793/3-1 from the German research foundation (DFG) and support by the Open Access Publication Fund of Humboldt-Universität zu Berlin.

References

- Adams, D., Rohlf, F., and Slice, D. (2013), “A Field Comes of Age: Geometric Morphometrics in the 21st Century,” *Hystrix, the Italian Journal of Mammalogy*, 24, 7–14. [1]
- Backenroth, D., Goldsmith, J., Harran, M. D., Cortes, J. C., Krakauer, J. W., and Kitago, T. (2018), “Modeling Motor Learning Using Heteroscedastic Functional Principal Components Analysis,” *Journal of the American Statistical Association*, 113, 1003–1015. [2]
- Baranyi, P., Yam, Y., and Várlaki, P. (2013), *Tensor Product Model Transformation in Polytopic Model-based Control*, Boca Raton, FL: CRC Press. [2]
- Bonhomme, V., Picq, S., Gaucherel, C., and Claude, J. (2014), “Momocs: Outline Analysis using R,” *Journal of Statistical Software*, 56, 1–24. [9,11]
- Brockhaus, S., Melcher, M., Leisch, F., and Greven, S. (2017), “Boosting Flexible Functional Regression Models with a High Number of Functional Historical Effects,” *Statistics and Computing*, 27, 913–926. [10]
- Brockhaus, S., Rügamer, D., and Greven, S. (2020), “Boosting Functional Regression Models with FDBOOST,” *Journal of Statistical Software*, 94, 1–50. [7]
- Brockhaus, S., Scheipl, F., and Greven, S. (2015), “The Functional Linear Array Model,” *Statistical Modelling*, 15, 279–300. [2,5,7,8]
- Bühlmann, P., and Yu, B. (2003), “Boosting with the L2 Loss: Regression and Classification,” *Journal of the American Statistical Association*, 98, 324–339. [2,6]
- Clutton-Brock, J., Dennis-Bryan, K., Armitage, P. L., and Jewell, P. A. (1990), “Osteology of the Soay Sheep,” *Bulletin of the British Museum (Natural History)*, 56, 1–56. [8]
- Cornea, E., Zhu, H., Kim, P., Ibrahim, J. G., and the Alzheimer’s Disease Neuroimaging Initiative. (2017), “Regression Models on Riemannian Symmetric Spaces,” *Journal of the Royal Statistical Society, Series B*, 79, 463–482. [1,4,5]
- Davis, B. C., Fletcher, P. T., Bullitt, E., and Joshi, S. (2010), “Population Shape Regression from Random Design Data,” *International Journal of Computer Vision*, 90, 255–266. [1]
- Dryden, I. L., and Mardia, K. V. (2016), *Statistical Shape Analysis: With Applications in R*, Chichester: Wiley. [1,2,4,7]
- Ferraty, F., Goia, A., Salinelli, E., and Vieu, P. (2011), “Recent Advances on Functional Additive Regression,” in *Recent Advances in Functional Data Analysis and Related Topics*, ed. F. Ferraty, pp. 97–102, Heidelberg: Springer. [2]
- Fletcher, P. T. (2013), “Geodesic Regression and the Theory of Least Squares on Riemannian Manifolds,” *International Journal of Computer Vision*, 105, 171–185. [1,5]
- Greven, S., and Scheipl, F. (2017), “A General Framework for Functional Regression Modelling,” (with discussion and rejoinder), *Statistical Modelling*, 17(1–2), 1–35 and 100–115. [1,2]
- Happ, C., and Greven, S. (2018), “Multivariate Functional Principal Component Analysis for Data Observed on Different (Dimensional) Domains,” *Journal of the American Statistical Association*, 113, 649–659. [2]
- Hinkle, J., Fletcher, P. T., and Joshi, S. (2014), “Intrinsic Polynomials for Regression on Riemannian Manifolds,” *Journal of Mathematical Imaging and Vision*, 50, 32–52. [1]
- Hofner, B., Hothorn, T., Kneib, T., and Schmid, M. (2011), “A Framework for Unbiased Model Selection Based on Boosting,” *Journal of Computational and Graphical Statistics*, 20, 956–971. [7]
- Hofner, B., Kneib, T., and Hothorn, T. (2016), “A Unified Framework of Constrained Regression,” *Statistics and Computing*, 26, 1–14. [5]
- Hong, Y., Singh, N., Kwitt, R., and Niethammer, M. (2014), “Time-Warped Geodesic Regression,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 105–112, Springer. [1]
- Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M., and Hofner, B. (2010), “Model-based Boosting 2.0,” *Journal of Machine Learning Research*, 11, 2109–2113. [5,6,7]
- Huckemann, S., Hotz, T., and Munk, A. (2010), “Intrinsic MANOVA for Riemannian Manifolds with an Application to Kendall’s Space of Planar Shapes,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32, 593–603. [1,4]
- Jeon, J. M., and Park, B. U. (2020), “Additive Regression with Hilbertian Responses,” *The Annals of Statistics*, 48, 2671–2697. [2,11]
- Jeon, J. M., Lee, Y. K., Mammen, E., and Park, B. U. (2022), “Locally Polynomial Hilbertian Additive Regression,” *Bernoulli*, 28, 2034–2066. [2]
- Jeon, J. M., Park, B. U., and Van Keilegom, I. (2021), “Additive Regression for Non-Euclidean Responses and Predictors,” *The Annals of Statistics*, 49, 2611–2641. [2]
- Jupp, P. E., and Kent, J. T. (1987), “Fitting Smooth Paths to Spherical Data,” *Journal of the Royal Statistical Society, Series C*, 36, 34–46. [1]
- Karcher, H. (1977), “Riemannian Center of Mass and Mollifier Smoothing,” *Communications on Pure and Applied Mathematics*, 30, 509–541. [6]
- Kendall, D. G., Barden, D., Carne, T. K., and Le, H. (1999), *Shape and Shape Theory* (Vol. 500), Chichester: Wiley. [2]
- Kent, J. T., Mardia, K. V., Morris, R. J., and Aykroyd, R. G. (2001), “Functional Models of Growth for Landmark Data,” in *Proceedings in Functional and Spatial Data Analysis*, 109115. [1]
- Kim, H. J., Adluru, N., Collins, M. D., Chung, M. K., Bendlin, B. B., Johnson, S. C., Davidson, R. J., and Singh, V. (2014), “Multivariate General Linear Models (mglim) on Riemannian Manifolds with Applications to Statistical Analysis of Diffusion Weighted Images,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2705–2712. [1]
- Kim, H. J., Adluru, N., Suri, H., Vemuri, B. C., Johnson, S. C., and Singh, V. (2017), “Riemannian Nonlinear Mixed Effects Models: Analyzing Longitudinal Deformations in Neuroimaging,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5777–5786. [11]
- Klingenberg, W. (1995), *Riemannian Geometry*, Berlin: de Gruyter. [2]
- Kneib, T., Hothorn, T., and Tutz, G. (2009), “Variable Selection and Model Choice in Geoadditive Regression Models,” *Biometrics*, 65, 626–634. [5]
- Kume, A., Dryden, I. L., and Le, H. (2007), “Shape-Space Smoothing Splines for Planar Landmark Data,” *Biometrika*, 94, 513–528. [1]
- Lay, D. M. (1967), “A Study of the Mammals of Iran: Resulting from the Street Expedition of 1962–63,” in *Fieldiana: Zoology* 54, Field Museum of Natural History. [8]
- Li, Y., and Ruppert, D. (2008), “On the Asymptotics of Penalized Splines,” *Biometrika*, 95, 415–436. [2,5]
- Lin, L., St. Thomas, B., Zhu, H., and Dunson, D. B. (2017), “Extrinsic Local Regression on Manifold-Valued Data,” *Journal of the American Statistical Association*, 112, 1261–1273. [1]
- Lin, Z., Müller, H.-G., and Park, B. U. (2020), “Additive Models for Symmetric Positive-Definite Matrices, Riemannian Manifolds and Lie Groups,” arXiv preprint arXiv:2009.08789. [2]
- Lutz, R. W., and Bühlmann, P. (2006), “Boosting for High-Multivariate Responses in High-Dimensional Linear Regression,” *Statistica Sinica*, 16, 471–494. [6]
- Mallasto, A., and Feragen, A. (2018), “Wrapped Gaussian Process Regression on Riemannian Manifolds,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5580–5588. [1]
- Mayr, A., Binder, H., Gefeller, O., and Schmid, M. (2014), “The Evolution of Boosting Algorithms,” *Methods of Information in Medicine*, 53, 419–427. [6]
- Meyer, M. J., Coull, B. A., Versace, F., Cinciripini, P., and Morris, J. S. (2015), “Bayesian Function-on-Function Regression for Multilevel Functional Data,” *Biometrics*, 71, 563–574. [2,5]
- Morris, J. S. (2015), “Functional Regression,” *Annual Review of Statistics and its Applications*, 2, 321–359. [2,11]
- Morris, J. S., and Carroll, R. J. (2006), “Wavelet-based Functional Mixed Models,” *Journal of the Royal Statistical Society, Series B*, 68, 179–199. [2]
- Müller, H.-G., and Yao, F. (2008), “Functional Additive Models,” *Journal of the American Statistical Association*, 103, 1534–1544. [2,5,11]
- Muralidharan, P., and Fletcher, P. T. (2012), “Sasaki Metrics for Analysis of Longitudinal Data on Manifolds,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1027–1034, IEEE. [1]
- Olsen, N. L., Markussen, B., and Raket, L. L. (2018), “Simultaneous Inference for Misaligned Multivariate Functional Data,” *Journal of the Royal Statistical Society, Series C*, 67, 1147–1176. [2]
- Pennec, X. (2006), “Intrinsic Statistics on Riemannian Manifolds: Basic Tools for Geometric Measurements,” *Journal of Mathematical Imaging and Vision*, 25, 127–154. [6]

- Petersen, A., and Müller, H.-G. (2019), “Fréchet Regression for Random Objects with Euclidean Predictors,” *The Annals of Statistics*, 47, 691–719. [1]
- Pigoli, D., Menafoglio, A., and Secchi, P. (2016), “Kriging Prediction for Manifold-Valued Random Fields,” *Journal of Multivariate Analysis*, 145, 117–131. [1]
- Pöllath, N., Schafberg, R., and Peters, J. (2019), “Astragalar Morphology: Approaching the Cultural Trajectories of Wild and Domestic Sheep Applying Geometric Morphometrics,” *Journal of Archaeological Science: Reports*, 23, 810–821. [7,8]
- R Core Team (2018), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. [7]
- Ramsay, J. O., and Silverman, B. W. (2005), *Functional Data Analysis*, New York: Springer. [1,11]
- Rosen, O., and Thompson, W. K. (2009), “A Bayesian Regression Model for Multivariate Functional Data,” *Computational Statistics & Data Analysis*, 53, 3773–3786. [2]
- Schafberg, R., and Wussow, J. (2010), “Julius Kühn. Das Lebenswerk eines agrarwissenschaftlichen Visionärs,” *Züchtungskunde*, 82, 468–484. [8]
- Schaffer, S. A. (2021), “Cytoskeletal Dynamics in Confined Cell Migration: Experiment and Modelling,” PhD thesis, LMU Munich. DOI:10.5282/edoc.28480. [9]
- Scheipl, F., Staicu, A.-M., and Greven, S. (2015), “Functional Additive Mixed Models,” *Journal of Computational and Graphical Statistics*, 24, 477–501. [2,5]
- Schiratti, J.-B., Allasonnière, S., Colliot, O., and Durrleman, S. (2017), “A Bayesian Mixed-Effects Model to Learn Trajectories of Changes from Repeated Manifold-Valued Observations,” *The Journal of Machine Learning Research*, 18, 4840–4872. [1]
- Shi, X., Styner, M., Lieberman, J., Ibrahim, J. G., Lin, W., and Zhu, H. (2009), “Intrinsic Regression Models for Manifold-Valued Data,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 192–199, Springer. [1]
- Srivastava, A., and Klassen, E. P. (2016), *Functional and Shape Data Analysis*, New York: Springer-Verlag. [2,11]
- Stöcker, A., Brockhaus, S., Schaffer, S. A., Bronk, B. v., Opitz, M., and Greven, S. (2021), “Boosting Functional Response Models for Location, Scale and Shape with an Application to Bacterial Competition,” *Statistical Modelling*, 21, 385–404. [2,6,10]
- Stöcker, A., Pfeuffer, M., Steyer, L., and Greven, S. (2022), “Elastic Full Procrustes Analysis of Plane Curves via Hermitian Covariance Smoothing,” <https://doi.org/10.48550/arXiv.2203.10522> [4]
- Thüroff, F., Goychuk, A., Reiter, M., and Frey, E. (2019), “Bridging the Gap between Single-Cell Migration and Collective Dynamics,” *eLife*, 8, e46842. [9]
- Volkman, A., Stöcker, A., Scheipl, F., and Greven, S. (2021), “Multivariate Functional Additive Mixed Models,” *Statistical Modelling*. <https://doi.org/10.1177/1471082X211056158> [2,9]
- Wood, S. N., Pya, N., and Säfken, B. (2016), “Smoothing Parameter and Model Selection for General Smooth Models,” *Journal of the American Statistical Association*, 111, 1548–1563. [5]
- Yao, F., Müller, H., and Wang, J. (2005), “Functional Data Analysis for Sparse Longitudinal Data,” *Journal of the American Statistical Association*, 100, 577–590. [1,4]
- Zeder, M. A. (2006), “Reconciling Rates of Long Bone Fusion and Tooth Eruption and Wear in Sheep (*Ovis*) and Goat (*Capra*),” *Recent Advances in Ageing and Sexing Animal Bones*, 9, 87–118. [8]
- Zhu, H., Chen, Y., Ibrahim, J. G., Li, Y., Hall, C., and Lin, W. (2009), “Intrinsic Regression Models for Positive-Definite Matrices with Applications to Diffusion Tensor Imaging,” *Journal of the American Statistical Association*, 104, 1203–1212. [1]
- Zhu, H., Li, R., and Kong, L. (2012), “Multivariate Varying Coefficient Model for Functional Responses,” *Annals of Statistics*, 40, 2634–2666. [2]
- Zhu, H., Morris, J. S., Wei, F., and Cox, D. D. (2017), “Multivariate Functional Response Regression, with Application to Fluorescence Spectroscopy in a Cervical Pre-cancer Study,” *Computational Statistics and Data Analysis*, 111, 88–101. [2]

6. Paper V: Elastic Shape Regression for Plane Curves

Paper V develops a quotient regression model introduced in Paper III for the quotient of the shape manifold introduced in Paper IV and the action of re-parametrization, which is by isometries. In this way, it combines regression for elastic curves with additive regression for functional planar shapes as a response given scalar covariates, i.e. it respects all invariances: rotation, translation, rescaling and re-parametrization. In addition, the necessary constraints needed to model symmetric shapes are provided, along with a demonstration of the interpretability of estimated nonlinear covariate effects in an analysis of bottle shapes.

Contributing article:

Stöcker, A., Steyer, L. and Greven, S. (2022). Elastic Shape Regression for Plane Curves. *Unpublished manuscript*

Declaration on personal contributions:

Here, the ideas of Paper I, Paper II and Paper IV are merged into a collaborative project, the actual execution of which is primarily carried out by Almond Stöcker. Sonja Greven and the author are involved in a supporting role, providing important advice, participating in discussions, and contributing to the proofs. This work is also part of Almond Stöcker's dissertation.

Elastic Shape Regression for Plane Curves

Almond Stöcker, Lisa Steyer, Sonja Greven

Humboldt-Universität zu Berlin

Abstract

For outline data such as arising for anatomical shapes in biomedical imaging, often only the shape of the outline rather than the used coordinate system or the parametrization of the outline curve are of interest. The square-root-velocity framework provides a basis for “elastic” statistical analysis of variability in the shapes of such curves, allowing to incorporate invariance with respect to the curve parameterization integrally into the data geometry, in addition to traditional shape invariances with respect to rotation, translation and scaling. However, little work has been done so far on elastic modeling of such data in dependence on covariates. We introduce an approach based on generalized additive regression that transfers the accustomed flexibility for scalar data to response shapes of plane curves, and provide necessary constraints required for modeling symmetric shapes. We illustrate interpretability of estimated non-linear covariate effects in an analysis of bottle shapes.

Keywords: Functional data, additive regression, square-root-velocity, geometric data, semi-parametric modeling

1 Introduction

Understanding shape variability of curves, for instance recorded in medical imaging, promises important insights in the areas of life sciences and beyond. In many data problems, say, when analyzing outlines of a particular brain area across different patients, the coordinate system applied for recording is likely arbitrary and size differences in patients are often not of interest. This has motivated statistical shape analysis (Dryden and Mardia, 2016) to define the shape of a plane curve as equivalence class modulo the shape invariances of translation, rotation and scale, equipped with a Riemannian manifold structure. Similarly, a curve is naturally described in parameterized form as a function, yet potentially only the image of the curve is of interest and analysis should then be invariant under re-parameterization (“warping”) – a problem closely related to the registration problem in functional data analysis (Marron et al., 2014). The square-root-velocity (SRV) framework (Srivastava and Klassen, 2016) provides a basis for statistical analysis of such shapes of curves modulo all mentioned invariances employing an “elastic” distance: Unlike for other approaches, re-parameterization proves isometric here, allowing to induce a quotient space distance on shapes of curves as infimum distance over its parameterizations. While first approaches to regression in this framework with shapes of curves as covariates are presented by Ahn et al. (2018) and Tucker et al. (2019), regression for such shapes as response variable are so far restricted to the work of Guo et al. (2020), who model tangent space principal component representations after warping alignment. However, this does not incorporate the elastic quotient space distance integrally into the model fit. Related regression models for (one-dimensional) functional data with warping-alignment (but no shape alignment) in the response were proposed by Matuk et al. (2021) and Hadjipantelis et al. (2014, 2015).

We introduce functional additive regression-type models (Greven and Scheipl, 2017; Morris, 2015) to flexibly model shapes of plane curves in dependence on covariates and base the entire model estimation on the elastic quotient space distance, which arises in the SRV framework and incorporates all considered invariances. The proposed approach extends earlier inelastic shape regression (Stöcker et al., 2022) combining gradient boosting for functional additive models (Brockhaus et al., 2015) with ideas of regression for manifold-valued responses (Cornea et al., 2017). Moreover, we consider the important special case of modeling curves with axial symmetry, and provide and implement corresponding required symmetry constraints. The approach is provided in the R package `manifoldboost` (github.com/Almond-S/manifoldboost/tree/elastic).

In Section 2, we provide a brief introduction into SRV-representation of plane curves (Section 2.1) and discuss generalized additive models from an object data perspective (Section 2.2) to motivate the proposed regression approach presented in Section 2.3. Having introduced the general model, we discuss constraints for modeling axial symmetric closed curves in Section 2.4 and present an elastic Riemannian L_2 -Boosting approach for model fitting in Section 2.5. In Section 3, we analyze bottle design based on outline shapes (Section 3.1) and use the analysis to motivate a simulation study investigating the impact of invariances on fitting performance (Section 3.2). Section 4 concludes with a discussion and outlook.

2 Elastic functional additive shape regression

2.1 Representation of shapes of plane curves in the SRV-framework

Identifying the real plane $\mathbb{R}^2 \cong \mathbb{C}$ with the complex numbers for convenience, we consider a *parameterized plane curve* an absolute continuous function $y : \mathcal{I} \rightarrow \mathbb{C}$ defined on an interval \mathcal{I} , where we assume y to be non-constant to avoid the degenerate case of a curve describing only a point and write $y \in \mathcal{AC}^*(\mathcal{I})$. For any such y , the component-wise derivative $\dot{y}(t) = dy(t)/dt$ exists for almost all $t \in \mathcal{I}$ and there exists a monotonously increasing *warping function* $\gamma : \mathcal{I} \rightarrow \mathcal{I}$ re-parameterizing the curve as $u = y \circ \gamma$ with constant speed, i.e. with $|\dot{u}(t)|$ constant for all t (e.g., Bruveris, 2016). Two parameterized curves $y_1, y_2 \in \mathcal{AC}^*(\mathcal{I})$ are called equivalent if they have the same constant speed parameterization $u_1 = u_2$, defining an *oriented curve* as their equivalence class. Although both are commonly referred to simply as “curves”, we explicitly write $[y]_w$ for the oriented curve described by a parameterized curve y for clarity. Mapping into an arbitrary coordinate system, the shape $[y]_s$ of y is defined as its equivalence class $[y]_s = \{\lambda \exp(\sqrt{-1}\omega)y + z \mid \lambda > 0, \omega \in \mathbb{R}, z \in \mathbb{C}\}$ over re-scaling by λ , rotation by ω radian, and translation by z . The definition directly carries over to the shape of $[y]_w$ as union over its representatives $[y] = \bigcup_{y \in [y]_w} [y]_s$ presenting our object of primary interest. The square-root-velocity (SRV) transform (Srivastava and Klassen, 2016), mapping $y \mapsto q$ with $q(t) = \dot{y}(t)/\sqrt{|\dot{y}(t)|}$ where defined and $q(t) = 0$ elsewhere, establishes a surjective map from $\mathcal{AC}^*(\mathcal{I})$ to $\mathbb{L}_{\mathbb{C}}^2(\mathcal{I})$, or briefly $\mathbb{L}_{\mathbb{C}}^2$, the Hilbert space of square-integrable complex-valued functions defined on \mathcal{I} (Bruveris, 2016). Loosing translation in the derivative, this yields a one-to-one identification of $[y]_s$ with $[q]_s = \{\lambda^2 \exp(\sqrt{-1}\omega)q \mid \lambda > 0, \omega \in \mathbb{R}\}$ on SRV-level. The quotient space of such $[q]_s$ with $q \neq 0$ corresponds to the complex projective space $\text{PL}_{\mathbb{C}}^2$ with a well-known symmetric Riemannian manifold structure (e.g., Klingenberg, 1995). This link is analogous to Kendall’s shape space (e.g., Dryden and Mardia, 2016) and a more detailed motivation of the geometry for modeling shapes of parameterized plane curves can be found in Stöcker et al. (2022). Inducing the geometry, however, via the SRV-representation of the curves allows to establish a suitable, elastic metric on $[\mathcal{AC}^*(\mathcal{I})]$, the space of oriented plane curve shapes $[y]$, as introduced by Srivastava et al. (2011) and defined below in Section 2.3. Modeling such $[y]$ as response objects in dependence on covariates is the target of this paper.

2.2 Generalized additive regression for modeling object data

Ever since Hastie and Tibshirani (1986) proposed generalized additive models as extension of generalized linear models (Nelder and Wedderburn, 1972) to non-linear covariate effects, a wealth of often inter-combinable extensions have been proposed (partly summarized in textbooks such as Fahrmeir et al., 2013; Wood, 2017; Stasinopoulos et al., 2017) leading to a versatile regression framework for statistical analysis in various data problems. While approaches so far have predominantly focused on scalar response variables Y , we take a geometric object data perspective on generalized additive models here to provide a roadmap for our model for shapes of plane curves. Their general model structure

$$g(\mu) = f(\mathbf{x}) = f_1(\mathbf{x}) + \cdots + f_J(\mathbf{x})$$

consists of three components: a target parameter μ of the distribution of Y depending on covariate values \mathbf{x} , an additive predictor $f(\mathbf{x}) = \sum_{j=1}^J f_j(\mathbf{x})$, and a link function g linking μ to the predictor.

Most commonly, μ presents a conditional mean of Y . The Fréchet mean (Fréchet, 1948; Ziezold, 1977) presents a general mean concept assuming Y a random element in a metric space (\mathcal{Y}, d) , i.e. a Borel-measurable map from some probability space into \mathcal{Y} . For simplicity, covariates \mathbf{X} are assumed a random vector of scalar covariates $\mathbf{x} \in \mathcal{X}$ in the following. A conditional Fréchet mean μ of Y , as modeled e.g. in the “Fréchet Regression” approach of Petersen and Müller (2019), is defined as a minimizer of the conditional expected squared distance

$$\mathbb{E}(d^2(\mu, Y) \mid \mathbf{X} = \mathbf{x}) = \sigma_{\mathbf{x}}^2 = \inf_{\mu' \in \mathcal{M}} \mathbb{E}(d^2(\mu', Y) \mid \mathbf{X} = \mathbf{x})$$

assuming finite variance(s) $\sigma_{\mathbf{x}}^2 < \infty$ and the model, which potentially restricts μ to some subspace $\mathcal{M} \subseteq \mathcal{Y}$. When d is the geodesic distance on a Riemannian manifold \mathcal{Y} , the Fréchet mean is typically referred to as intrinsic mean or Riemannian center of mass (Karcher, 1977; Afsari, 2011). In Euclidean spaces, it corresponds to the usual expected value.

While additive models have also been formulated on Lie groups (Lin et al., 2020), an approach extending and in the tradition of generalized linear models requires a linear structure for the space of the predictor, i.e. for the predictor $f : \mathcal{X} \rightarrow \mathcal{V}$ to map the covariates into a vector space \mathcal{V} . The predictor values can then be mapped into the space of the responses using a suitable response (inverse link) function g^{-1} . In practice, $f(\mathcal{X})$ typically restricts to a finite-dimensional subspace of \mathcal{V} with a basis $v_1, \dots, v_K \in \mathcal{V}$. This lets us follow an analogous approach to Brockhaus et al. (2015); Scheipl et al. (2016) for functional data, modeling covariate effect functions $f_j(\mathbf{x})$ as

$$f_j(\mathbf{x}) = \sum_{k=1}^K \sum_{h=1}^H \theta_{jhk} b_{jh}(\mathbf{x}) v_k$$

expanded in a finite tensor-product basis of the basis $\{v_k\}_k$ and some effect basis $b_{jh} : \mathcal{X} \rightarrow \mathbb{R}$, $h = 1, \dots, H$. Estimating $f_j(\mathbf{x})$ then reduces to estimating the $H \times K$ coefficient matrix $\Theta_j = \{\theta_{jhk}\}_{h,k}$. This approach effectively models each basis coefficient for the v_k as an additive function of the covariates. The tensor-product effect structure thus prepares the ground for directly building on covariate effects established for scalar additive models. Typical example effects of a metric covariate x_1 in \mathbf{x} include linear effects $f_j(\mathbf{x}) = \beta x_1$ (specifying $b_{j1}(\mathbf{x}) = x_1$, $H = 1$) and smooth spline effects with $\{b_{jh}\}_h$, say, a B-spline basis, where coefficient β like all $f_j(\mathbf{x})$ here is an element of \mathcal{V} . Effects of a categorical covariate $x_2 \in \{1, \dots, L\}$ are implemented by mapping the l th level to a contrast vector $\mathbf{b}_j(l)$ as in linear regression. Interactions and other types of effects are possible, and effect visualizations can be achieved by tensor-product factorization (Stöcker et al., 2022).

The link function g is commonly assumed invertible with the response function $g^{-1} : \mathcal{V} \rightarrow \mathcal{M}$ mapping the predictor to the desired model space \mathcal{M} for the response. Its choice is usually motivated by properties of the involved spaces, and aims at offering a natural and convenient interpretation. For the special case where \mathcal{Y} has a (symmetric) Riemannian manifold structure, the Riemannian exponential map $\text{Exp}_p : T_p\mathcal{Y} \rightarrow \mathcal{Y}$ takes a prominent role here, mapping a tangent vector $v \in \mathcal{V} = T_p\mathcal{Y}$ in the tangent space of \mathcal{Y} at $p \in \mathcal{Y}$ to a point in \mathcal{Y} . Although other options are possible (Cornea et al., 2017), the Exp map was established as a response function in generalized-linear-regression-type models for manifold-valued responses by Zhu et al. (2009); Shi et al. (2009); Kim et al. (2014); Cornea et al. (2017); Stöcker et al. (2022) generalizing geodesic regression (Fletcher, 2013) to multiple regression. Geodesic regression is the direct generalization of simple linear regression: a covariate value of $x_1 = 1$ of a single linear effect is mapped from the “intercept” p to $\mu = \text{Exp}_p(\beta x_1)$ at a distance $d(\mu, p) = \|\beta\|$ corresponding to the norm of the “slope” $\beta \in T_p\mathcal{Y}$. Conversely, this yields a Riemannian Log_p -link function given by the inverse of Exp_p , which can unrestrictively be assumed to exist almost surely for symmetric Riemannian manifolds (Pennec, 2006; Cornea et al., 2017). The Log_p -link maps $y \in \mathcal{Y}$ to the tangent space $T_p\mathcal{Y}$, which is equipped with a Hilbert space structure corresponding to the Riemannian metric on \mathcal{Y} .

For other cases than Riemannian manifolds, suitable choices of \mathcal{V} and of the response function are less straightforward. For elastic shape analysis, we propose in the following to build on the Riemannian manifold structure and choice of tangent space \mathcal{V} of the inelastic shape case, but to adjust the response function appropriately.

2.3 Functional additive regression for shapes of plane curves

Consider a sample of plane curves $y_1, \dots, y_n \in \mathcal{AC}^*(\mathcal{I})$ recorded together with vectors of scalar covariates $\mathbf{x}_1, \dots, \mathbf{x}_n$. We model the conditional Fréchet mean $[\mu_i]$ of their shapes $[y_i], i = 1, \dots, n$, considering the $([y_i], \mathbf{x}_i)$ independent realizations of response-covariate tuples with the response presenting a random element in the metric space $([\mathcal{AC}^*(\mathcal{I})], d)$. The elastic distance d on the shape space $[\mathcal{AC}^*(\mathcal{I})]$ proposed by Srivastava et al. (2011) is induced as

$$d([y_1], [y_2]) = \inf_{\gamma \in \Gamma} d_{\text{PL}_{\mathbb{C}}^2}([q_1]_s, [q_2 \circ \gamma \sqrt{\hat{\gamma}}]_s) = \inf_{\gamma \in \Gamma, \omega \in \mathbb{R}} d_{\mathbb{S}}(q_1, \exp(\omega \sqrt{-1}) q_2 \circ \gamma \sqrt{\hat{\gamma}})$$

by the geodesic distance $d_{\text{PL}_{\mathbb{C}}^2}$ on the complex projective space of the $[q_i]_s$, where q_i denotes the SRV-transform of $y_i, i = 1, 2$. The set Γ of warping functions γ contains the strictly increasing surjective differentiable functions $\gamma : \mathcal{I} \rightarrow \mathcal{I}$. When modeling closed curves on the interval $\mathcal{I} = [t_0, t_1]$, i.e. with $y_i(t_0) = y_i(t_1)$, Γ in addition contains all functions $\gamma : t \mapsto t + \tau - (t_1 - t_0) \mathbb{1}_{(t_1 - \tau, t_1)}(t)$ that shift the starting point by $\tau \in [0, t_1 - t_0]$, where $\mathbb{1}_{\mathcal{U}}(t) = 1$ if $t \in \mathcal{U}$ is contained in the set \mathcal{U} and 0 otherwise, as well as concatenations of functions in Γ . The metric on $\text{PL}_{\mathbb{C}}^2$ is in turn induced from the submanifold geometry of the Hilbert sphere $\mathbb{S} = \{q \in \mathbb{L}_{\mathbb{C}}^2 \mid \|q\| = 1\} \subset \mathbb{L}_{\mathbb{C}}^2$, where $\|q\| = (\int_{\mathcal{I}} |q(t)|^2 dt)^{1/2}$ denotes the standard norm on $\mathbb{L}_{\mathbb{C}}^2$. The geodesic distance $d_{\mathbb{S}}(q_1, q_2)$ on the sphere reflects the arc-length between unit-norm representatives q_1 and q_2 with $\|q_i\| = 1$. This corresponds to scaling curves $[y_i]_w$ to unit-length. Due to the SRV-representation, not only rotation by ω radian but also reparameterization by $\gamma \in \Gamma$ acts by isometries, i.e. for common actions $\exp(\omega \sqrt{-1}) y_i \circ \gamma, i = 1, 2$, the $\mathbb{L}_{\mathbb{C}}^2$ inner product $\langle q_1, q_2 \rangle = \langle \exp(\omega \sqrt{-1}) q_1 \circ \gamma \sqrt{\hat{\gamma}}, \exp(\omega \sqrt{-1}) q_2 \circ \gamma \sqrt{\hat{\gamma}} \rangle$ is left unchanged. This allows to define the quotient space distance d as infimum over distances in the original space. While we focus on d in the following, related alternative elastic distances on shapes of plane curves have been proposed, including the geodesic distance on the subspace of closed curves (Srivastava et al., 2011), a more general family of elastic distances (Kurtek and Needham, 2018), and the elastic full Procrustes distance (Stöcker et al., 2022).

We model the mean shape $[\mu_i]$ for the i th observation via the SRV-transform m_i of a unit-length curve mean representative $\mu_i \in \mathcal{AC}^*(\mathcal{I})$ using an additive model of the form

$$[\mu_i] = g_{[\psi]}^{-1}(f(\mathbf{x}_i)) = g_{[\psi]}^{-1}\left(\sum_{j=1}^J f_j(\mathbf{x}_i)\right)$$

induced by the Riemannian (inelastic) functional additive model

$$m_i = \text{Exp}_p(f(\mathbf{x}_i))$$

on SRV-level: we choose the Riemannian exponential $\text{Exp}_p(\beta) = \cos(\|\beta\|)p + \sin(\|\beta\|)\beta/\|\beta\|$ on \mathbb{S} as response function mapping the additive predictor $f(\mathbf{x})$ along great-arcs. Constraining tangent vectors $\beta \in T_p\mathbb{S}$ to the subspace horizontal to rotation, this also corresponds to the Riemannian exponential on $\text{PL}_{\mathbb{C}}^2$ and lets us identify $T_{[p]_s}\text{PL}_{\mathbb{C}}^2$ with the subspace $\mathcal{V}_p = \{q \in \mathbb{L}_{\mathbb{C}}^2 \mid \langle q, p \rangle = 0\}$ orthogonal to $p \in \mathbb{S}$ (compare, e.g., Stöcker et al., 2022; Dryden and Mardia, 2016; Klingenberg, 1995). Thus, common basis functions $\tilde{v}_k : \mathcal{I} \rightarrow \mathbb{R}$, $k = 1, \dots, K + 1$, used for functional additive models (Scheipl et al., 2015), such as polynomial splines, can be utilized for constructing tensor-product effects $f_j(\mathbf{x})$ after linear transformation to a constrained basis v_k , $k = 1, \dots, K$, spanning a K -dimensional subspace of \mathcal{V}_p (analogous to Stöcker et al., 2022). To obtain a transparent model space, we assume that the same basis $\{v_k\}_k$ is utilized for all f_1, \dots, f_J and also p , such that also m is in its span. Steyer et al. (2021) show identifiability of a representation of SRV-transforms in a B-spline basis of order one under warping, ensuring that for this choice, we can un-restrictively assume that $g_{[\psi]}([\mu]) = \text{Log}_p(m)$ yields a valid link function of the target mean shape $[\mu]$ modulo re-parameterization. The intercept p is typically specified as the SRV-transform of a representative $\psi \in \mathcal{AC}^*(\mathcal{I})$ of the unconditional Fréchet mean $[\psi]$ of the marginal distribution of $[y_1], \dots, [y_n]$. Correspondingly, effects $f_j(\mathbf{x})$ are typically constrained to be centered to zero mean $\sum_{i=1}^n f_j(\mathbf{x}_i) = 0$. Basing our implementation in the R package `manifoldboost` on the package `FDboost`, an overview over implemented covariate effects is provided by Brockhaus et al. (2020).

2.4 Modeling symmetric and closed shape means

In many data scenarios, such as the bottle design data presented in Section 3.1, it is desirable to model mean curves as *symmetric* by imposing respective constraints. For convenience, we consider curves defined on $\mathcal{I} = [-1, 1]$ in the following and call a function $f : [-1, 1] \rightarrow \mathbb{C}$ even if $f(t)^\dagger = f(-t)$ and odd if $f(t)^\dagger = -f(-t)$ for all $t \in [-1, 1]$, where $z^\dagger = \Re(z) - \sqrt{-1}\Im(z)$ denotes the complex conjugate of $z \in \mathbb{C}$. $[\mu]$ is called (axis)symmetric if there is an odd $\mu \in [\mu]$ (i.e. μ is symmetric about the imaginary axis) or, equivalently if there is an even $\mu \in [\mu]$ (i.e. μ is symmetric about the real axis). The back-transform given by $\tilde{\mu}(t) := \int_0^t m(s)|m(s)| ds$ (i.e. $\tilde{\mu} = \mu - \mu(0)$) is odd whenever its SRV-transform m is even (see Appendix A.1). Hence, we ensure symmetry of the mean shape $[\mu]$ by constraining the modeled m to be even. This can be implemented by utilizing even basis functions $v_k^{\Re} : [-1, 1] \rightarrow \mathbb{R}$ for its real part and odd basis functions $v_k^{\Im} : [-1, 1] \rightarrow \mathbb{R}$ for its imaginary part in the effect functions (with the same notion of odd/even in the real special case). Constraining a B-spline basis to even or odd splines presents linear constraints, which we implement via basis transforms for general use in the R package `mboost` (Hothorn et al., 2010).

In contrast to symmetry, closedness of curves – also often desired in practice – poses a more challenging, non-linear constraint. Under symmetry, however, we argue that good results can already be expected with only a simpler closedness constraint on SRV-level. The (shape of the) oriented curve $[\mu]_w$ is closed if any and hence all $\mu \in [\mu]_w$ are closed. If μ is closed and continuously

differentiable in the vicinity of $\mu(-1) = \mu(1)$, also its SRV-transform m is closed. The package `mboost` already offers a linear constraint for closed (cyclic) B-splines (Hofner et al., 2016), which we employ for m . However, closedness of m is not sufficient for closedness of $\tilde{\mu}$ but leaves a gap $\delta = \tilde{\mu}(1) - \tilde{\mu}(-1)$ between its end-points. The geometry of closed curves in the SRV-framework has been considered in the literature (Srivastava et al., 2011; Srivastava and Klassen, 2016) but involves the non-linear constraint $\delta = \int_{-1}^1 m(s)|m(s)| ds = 0$. Instead, we focus on implementation of the symmetry constraint here and naively close $\tilde{\mu}$ with a small line segment between the endpoints of both sides of the curve. While extending curves by a line segment to a closed curve is always possible, the symmetry constraint ensures that transitions are differentiable in typical cases (for details see Appendix A.1). This pragmatic solution will, thus, be satisfactory in many data problems of this type, avoiding further restrictions of the geometry and more expensive computations.

2.5 Model fitting using elastic Riemannian L_2 -Boosting

For model estimation, we adapt Riemannian L_2 -Boosting (Stöcker et al., 2022) to elastic fitting in the SRV-framework. Component-wise gradient boosting (Bühlmann and Hothorn, 2007) is a forward step-wise estimation procedure offering inherent variable selection and a high flexibility to fit with respect to various loss functions (Mayr et al., 2014a,b) by effectively fitting gradients of the target loss with separate “base-learners” with respect to penalized least-squares. The dual regularization imposed by the base-learner penalty and informed early stopping make boosting also well-suited for high-dimensional (functional) responses (Stöcker et al., 2018; Lutz and Bühlmann, 2006). In the case of the quadratic loss, gradient boosting reduces to L_2 -Boosting (Bühlmann and Yu, 2003) corresponding to iterative re-fitting of model residuals. Stöcker et al. (2022) generalize conventional Euclidean L_2 -Boosting to Riemannian L_2 -boosting fitting base-learners to *transported residuals* (Cornea et al., 2017) in an approach based on the functional data extension (Brockhaus et al., 2015) of the boosting framework of Hothorn et al. (2010). Computing transported residuals, however, involves concatenation of the Riemannian Log-map and parallel transport, which are, as such, not available in our case. Hence, we borrow the Log-map from PL_C^2 after preceding warping-alignment, which is along the lines of Srivastava and Klassen (2016). This analogous to the procedure for rotation and, after full alignment with respect to rotation and warping, the length of the residual reflects the distance $d([\hat{\mu}_i], [y_i])$ of a prediction $[\hat{\mu}_i]$ to the respective shape $[y_i]$. Using this generalization, we fit our additive model for shapes of plane curves in the SRV-framework with respect to the quadratic elastic loss $d^2([\hat{\mu}], [y])$, estimating the conditional Fréchet mean by successively reducing the empirical risk $\sum_{i=1}^n d^2([\hat{\mu}_i], [y_i])$ over observations $i = 1, \dots, n$ analogously to the Riemannian case. After initialization, the proposed boosting algorithm (Algorithm 1) repeatedly adds to the model predictor $\hat{f}(\mathbf{x})$ by iteratively A) computing warping-aligned transported residuals, B) fitting them with the base-learners corresponding to predictor components $f_j(\mathbf{x})$, and C) updating the best-performing base-learner, until a stopping criterion is met. The single steps are detailed in the following.

Algorithm 1: Elastic Riemannian L^2 -Boosting

Fix intercept p , specify step-length $\eta > 0$ and base-learner penalty, initialize $\hat{f}(\mathbf{x}) = 0$;

repeat

A) Computing residuals:

foreach $i = 1, \dots, n$ **do**

Predict mean shape representative $\hat{\mu}_i$ based on current predictor $\hat{f}(\mathbf{x}_i)$;

Warping-align $y_i \xrightarrow{\text{align to } \hat{\mu}_i} \tilde{y}_i$;

Map $\tilde{y}_i \xrightarrow{\text{SRV-trafo}} \tilde{q}_i \xrightarrow{\text{Log}} \tilde{\epsilon}_i \xrightarrow{\text{Transp}} \epsilon_i$ to transported residual $\epsilon_i \in T_{[p]_s} \text{PL}_{\mathbb{C}}^2$;

end

B) Fitting baselearners:

foreach $j = 1, \dots, J$ **do**

Fit j th base-learner to residuals $\epsilon_i, i = 1 \dots, n$, to obtain $\check{f}_j(\mathbf{x})$;

Determine insample performance ;

end

C) Updating the predictor:

Set $\hat{f}(\mathbf{x}) \leftarrow \hat{f}(\mathbf{x}) + \eta \check{f}_j(\mathbf{x})$ for the best performing base-learner j ;

until *stopping criterion is met*;

Initialization: The algorithm presupposes a fixed intercept p . However in practice, p is typically estimated as SRV-transform \hat{p} of a curve representative $\hat{\psi}$ of an estimate $[\hat{\psi}]$ of the overall Fréchet mean shape $[\psi]$ of the response. We obtain \hat{p} from an intercept model (i.e., with a single constant base-learner) fitted in a previous Riemannian L^2 -Boosting run. This fit is based on a preliminary intercept p_0 fitted for instance as $\mathbb{L}_{\mathbb{C}}^2$ -average on reasonably aligned curve data. Some alternatives to this choice are described in Section 3.2.

A) Computing residuals: In the Riemannian manifold of shapes of parameterized curves $[y_i]_s$ predicted as $[\hat{\mu}_i]_s$ via the SRV-transform \hat{m}_i of the predicted curve representative $\hat{\mu}_i$, transported residuals ϵ_i are defined as follows: first, a local residual $\tilde{\epsilon}_i \in T_{[\hat{m}_i]_s} \text{PL}_{\mathbb{C}}^2$ in the (linear) tangent space is obtained as $\tilde{\epsilon}_i = \text{Log}_{[\hat{m}_i]_s}([q_i]_s)$ from the SRV-transform q_i of y_i . Due to the geometry of $\text{PL}_{\mathbb{C}}^2$, this can effectively be computed using the Log-map on the sphere \mathbb{S} as $\tilde{\epsilon}_i = \text{Log}_{\hat{m}_i}(\tilde{q}_i)$ when $\tilde{q}_i \in [q_i]_s$ and \hat{m}_i are rotation-aligned (compare, e.g., Huckemann et al., 2010). The local residuals reflect the distance $\|\tilde{\epsilon}_i\| = d([\hat{m}]_s, [q_i]_s)$ and correspond to the negative gradient $\tilde{\epsilon}_i = -\nabla_{[\hat{m}]_s} d^2([\hat{m}]_s, [q_i]_s)$ pointing into the direction of loss-reduction (Pennec, 2006). However, for $i = 1, \dots, n$, they are elements of different spaces. Parallel-transport $\text{Transp}_{[\hat{m}_i]_s, [p]_s} : T_{[\hat{m}_i]_s} \text{PL}_{\mathbb{C}}^2 \rightarrow T_{[p]_s} \text{PL}_{\mathbb{C}}^2$ isometrically maps the local residuals to transported residuals $\epsilon_i = \text{Transp}_{[\hat{m}_i]_s, [p]_s}(\tilde{\epsilon}_i)$ in the space $\mathcal{V}_p \cong T_{[p]_s} \text{PL}_{\mathbb{C}}^2$ of the linear predictor. In Riemannian L^2 -Boosting (Stöcker et al., 2022), transported residual ϵ_i are repeatedly fit to reduce the loss. Details concerning the involved maps can be found, e.g., also in Cornea et al. (2017); Huckemann et al. (2010).

As rotation, warping presents an isometric action. To fit shapes of curves $[y_i]$ also involving warping-invariance, we proceed analogously to rotation, and warping align y_i to μ_i before computing transported residuals on the parameterized curve shapes $[\tilde{y}_i]_s$ of the aligned representatives $\tilde{y}_i \in [y_i]$ as described above. Due to alignment and concatenation of length-preserving maps, the quadratic loss on predictor-level $\|\text{Log}_p(\hat{m}_i) - \epsilon_i\|^2 = d_{\text{PL}_{\mathbb{C}}^2}^2([\hat{m}_i]_s, [\tilde{q}_i]_s) \approx d^2([\hat{\mu}_i], [y_i])$ approximates the target elastic loss. Hence, fitting warping-aligned transported residuals on predictor level, we may reduce the

loss on the level of curve shapes. Perfect equality in the second relation would require simultaneous rotation and warping alignment, but we approximate it by subsequent alignment for computational efficiency.

B) Fitting base-learners: Base-learners are associated with the additive model components $f_j(\mathbf{x})$, $j = 1, \dots, J$, by considering them as individual predictors fitted to a sample of pseudo-responses $\epsilon_i \in \mathcal{V}_p$ in the Hilbert space \mathcal{V}_p at covariate values $\mathbf{x}_i \in \mathcal{X}$, $i = 1, \dots, n$. As elements of \mathcal{V}_p , they are fitted with respect to the penalized least-squares criterion to obtain $\check{f}_j = \arg \min_{f_j} \sum_{i=1}^n \|f_j(\mathbf{x}_i) - \epsilon_i\|^2 + \text{pen}_j(f_j)$. Using tensor-product effects $f_j(\mathbf{x}) = \sum_{k=1}^K \sum_{h=1}^H \theta_{jhk} b_{jh}(\mathbf{x}) v_k$ and a non-negative definite quadratic penalty term $\text{pen}_j(f_j)$, \check{f}_j is given by the well-known linear estimator for the vector of coefficients θ_{jhk} . Typically, $\text{pen}_j(f_j)$ is induced by suitable penalties for the basis $\{v_k\}_k$ in \mathcal{V} and the scalar effect basis $\{b_{jh}\}_h$. For B-splines, ridge or higher-order difference penalties on the coefficients θ_{jhk} present convenient choices (for details see, e.g., Brockhaus et al., 2015; Stöcker et al., 2022). For comparability across base-learners, the penalties are typically specified to achieve the same effective degrees of freedom (Hofner et al., 2011) for $j = 1, \dots, J$. The in-sample performance of the j th base-learner is then measured in terms of its residual sum of squares $\text{RSS}_j = \sum_{i=1}^n \|\check{f}_j(\mathbf{x}_i) - \epsilon_i\|^2$.

C) Updating the predictor: In each boosting iteration, only the base-learner with lowest RSS_j is added to the current predictor, weighted with a step-length of typically $\eta = 0.1$. If a base-learner is never selected, the corresponding covariate effect drops out of the model. If it has been selected already, the addition results in a coefficient update.

Stopping the algorithm early provides important means of regularization in high-dimensional data scenarios (Mayr et al., 2012). We select the stopping iteration via curve-wise cross-validation. For functional responses, this has proven a valuable tool to avoid over-fitting also in scenarios with high auto-correlation without explicit modeling of the covariance structure (Stöcker et al., 2018).

In practice, curves y_i , $i = 1, \dots, n$, are recorded at discrete sampling points and computations involving $\mathbb{L}_{\mathbb{C}}^2$ inner products are approximated by numerical integration as described by Stöcker et al. (2022). For warping-alignment based on discretely recorded curves, we rely on the approach of Steyer et al. (2021) and its implementation in the R package `elasdics` (Steyer, 2021).

3 Analysis of bottle design

3.1 Modeling bottle outline shapes

Shapes of everyday objects yield an ideal platform for illustration and evaluation of shape analysis, providing intuitive visual access to assess even small changes in shape. Bonhomme et al. (2014) provide a dataset of whisky and beer bottle outlines of 20 different brands, each with their characteristic designs. Based on the $n = 40$ recorded curves y_i , $i = 1, \dots, n$, we model their conditional mean shape $[\mu_i]$ with representatives $\mu_i \in \mathcal{AC}^*(\mathcal{I})$ in dependence on their bottle `type` (whisky/beer) and `size` in centiliter (covariates $\mathbf{x}_i = (x_{\text{type},i}, x_{\text{size},i})^\top$) as

$$[\mu_i] = g_{[p]}^{-1}(\alpha_{\text{type},i} + \beta x_{\text{size},i} + \beta_{\text{type}} x_{\text{size},i} + f(x_{\text{size},i}) + f_{\text{type}}(x_{\text{size},i}))$$

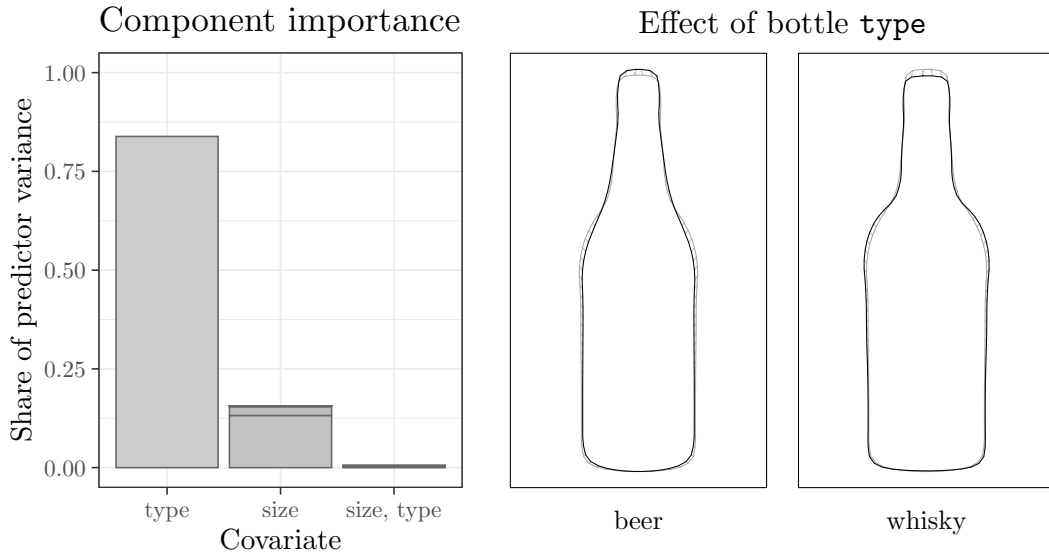


Figure 1: *Left:* Shares $\sum_{i=1}^n (\hat{f}_j^{[k]}(\mathbf{x}_i))^2 / \sum_{i=1}^n \sum_{j=1}^J \|\hat{f}_j(\mathbf{x}_i)\|^2$ of the variance of each (centered) factorized effect component $\hat{f}_j^{[k]}(\mathbf{x})$ selected into the model in overall predictor variance. Bars for its factorization components are stacked for each base-learner. *Right:* Estimated elastic mean shape of beer and whisky bottles setting **size**-effects $\hat{f}(x_{\text{size}}) + \hat{f}_{\text{type}}(x_{\text{size}}) = 0$. Bottle outlines are plotted aligned to the estimated overall mean shape (*grey line*) and corresponding time-points are connected by line segments.

via the unit-norm SRV-transform

$$m_i = \text{Exp}_p(\alpha_{\text{type},i} + \beta x_{\text{size},i} + \beta_{\text{type}} x_{\text{size},i} + f(x_{\text{size},i}) + f_{\text{type}}(x_{\text{size},i}))$$

of μ_i with an effect-coded binary effect $x_{\text{type}} \mapsto \alpha_{\text{type}} \in \mathcal{V}_p$ and, for **size**, a linear effect with coefficient β and a smooth effect $f(x_{\text{size}})$ centered around the linear effect, as well as their interactions with **type**. The effect functions f and f_{type} are modeled as cubic B-splines and m and p with piece-wise linear B-splines with symmetry and closedness constraints (adjusting penalty matrices correspondingly). In covariate direction, a second order difference penalty on coefficients implements equal effective degrees of freedom for all base-learners. For model fitting, the densely observed response curves are regularly evaluated at 100 points following a consistent parameterization scheme (constant-speed between landmarks). Although irregular sampling is possible, the regular design allows use of the functional linear array model (Brockhaus et al., 2015) for efficient computations (ca. 70 seconds for a single fit followed by 7.6 minutes of cross-validation on a regular computer without parallelization). After 10-fold curve-wise cross-validation, the algorithm with step-length $\eta = 0.1$ is stopped after 30 iterations resulting in an estimated predictor $\hat{f}(x_{\text{type}}, x_{\text{size}}) = \hat{\alpha}_{\text{type}} + \hat{f}(x_{\text{size}}) + \hat{f}_{\text{type}}(x_{\text{size}})$ omitting linear terms for **size**. The effect of **type** is illustrated in Fig. 1 presenting the largest effect in the model. As typical for shape variation, differences are comparably small after registration. Yet, they reflect characteristic design patterns, with whisky bottles exhibiting more pronounced “shoulders” and more tendency towards vaulted bottle necks.

For visualization of the **size** effect in Fig. 2, we employ tensor-product factorization (Stöcker et al., 2022) to decompose $\hat{f}(x_{\text{size}}) = \sum_{k=1}^{K'} \hat{v}^{[k]} \hat{f}^{[k]}(x_{\text{size}})$, with K' the minimum of marginal basis

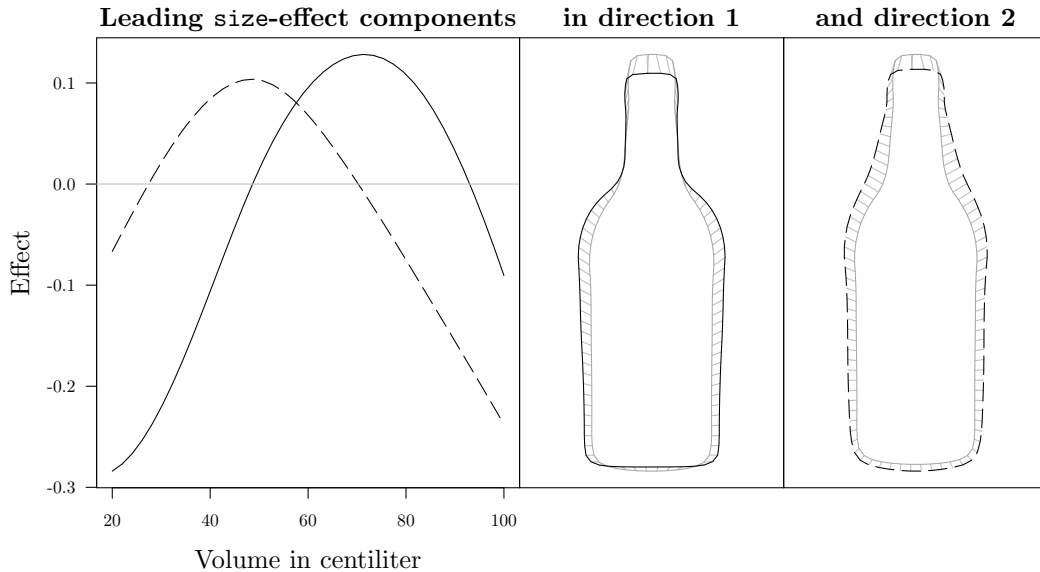


Figure 2: The two leading size-effect components $\hat{f}^{[1]}(x_{\text{size}})$ (*black solid lines*) and $\hat{f}^{[2]}(x_{\text{size}})$ (*black dashed lines*), explaining around 13.2% and 2.2% of the predictor variance respectively, depicted together with their respective directions $\hat{v}^{[1]}$ and $\hat{v}^{[2]}$. Directions are illustrated by showing bottle outlines represented by $\zeta = \text{Exp}_p(\hat{v}^{[k]})$ aligned to the overall mean shape (*gray*). Accordingly, the shown changes in the bottle outlines reflect an effect of $\hat{f}^{[k]}(x_{\text{size}}) = 1$, $k = 1, 2$.

dimensions, into independent effect components $\hat{f}^{[k]} : [0, 100] \rightarrow \mathbb{R}$ presenting scalar effects into orthogonal effect directions $\hat{v}^{[k]} \in T_{[p]}\text{PLC}^2$ sorted with decreasing effect variance $\frac{1}{n} \sum_{i=1}^n (\hat{f}^{[k]}(x_{\text{size}}))^2$ over the data. The decomposition lets us plot the effect despite its non-linearity and allows to depict also visually small effects on a suitable scale. Effects into the main direction $\hat{v}^{[1]}$ and the second direction $\hat{v}^{[2]}$ effectively explain all predictor variance of the **size**-effect (Fig. 1). The first reflects a broadening or tightening of the bottle shoulders for $\hat{f}^{[1]}(x_{\text{size}}) > 0$ or < 0 , respectively. A positive or negative second component $\hat{f}^{[2]}(x_{\text{size}})$ leads to a more wedge-shaped or more champagne-bottle-shaped neck of the bottle. The estimated interaction effect of **size** and **type** is vanishingly small in size and, thus, not shown. Even though the **size** distribution of beer ($x_{\text{size}} \in [25, 75]$, average $\bar{x}_{\text{size}} \approx 42$ centiliter) and whisky bottles ($x_{\text{size}} \in [70, 100]$, $\bar{x}_{\text{size}} \approx 73$) in the data overlap, their ranges clearly differ and the **size**-effect is highly correlated with **type**. Moreover, beverage brands are not selected representatively. Hence, we avoid a deeper interpretation, remaining with the illustration of the proposed model that captures familiar directions of shape variability in the data.

3.2 Empirical evaluation of elastic Riemannian L_2 -Boosting

Performance of model-based boosting was investigated and justified in simulation studies in various advanced modeling scenarios (e.g., Thomas et al., 2018) and also in (inelastic) modeling of functional and shape responses (Brockhaus et al., 2015; Stöcker et al., 2022). Boosting is generally known for its slow over-fitting behavior (Bühlmann and Hothorn, 2007). Nevertheless, early stopping is important for variable selection (investigated, e.g., by Hofner et al., 2011; Brockhaus et al., 2018) as well as for comparably small sample sizes of highly auto-correlated response curves in functional

models (Stöcker et al., 2018). The SRV framework is well-established for modeling shapes of curves (Srivastava and Klassen, 2016), good performance of the utilized warping-alignment procedure has been shown by Steyer et al. (2021), and good fitting behavior of Riemannian L_2 -Boosting in a related shape geometry has been validated by Stöcker et al. (2022). Here, we thus focus on warping invariance in the fitting behavior of our elastic regression approach and compare this also to the role of shape invariances. Although the model is widely invariant under warping and shape preserving transformations, the estimate \hat{p} of the SRV representative p of the intercept $[\psi]$ serves as starting point and typically depends in turn on a starting value \hat{p}_0 depending on the starting parameterization and positioning of the recorded curve representatives y_1, \dots, y_n . Initially aligning all curves to \hat{p} , the model fit then indirectly also depends on “reasonable” starting representatives. While indicating a good performance overall, the simulations will hence also show that a good model fit relies on a good fit of the intercept.

To provide a realistic scenario and control the sources of variability, we simulate datasets by sampling from the bottle outline dataset of Section 3.1, applying random warping and/or random positioning (i.e., random translation, rotation, and scaling) to the original curves. For random warping, original curves are interpolated at a total of 100 points along the bottle outlines (of 123 to 193 original sample points), which are then considered as the observations sampled on a fixed regular grid. All random transformations are applied with a moderate variability around the original curves, which already exceeds the warping variability observed in usual data settings where curve data is commonly more or less registered with similar parameterizations (for simulation details see Appendix A.2). We sample response-covariate tuples without replacement, such that variability in scenarios with all $n = 40$ observations is exclusively due to the random transformations. Scenarios with a sample size of $n = 30$ also reflect generalization error, subsampling 75% of the data stratified with respect to bottle **type**. In addition to these main scenarios, we also consider one $n = 80$ scenario with all observations twice in the data but with different random transformations. For each scenario, 100 simulated datasets are fit with the bottle model of Section 3.1, considering the original fit as ground truth and fixing the number of boosting iterations to 30 to speed up computations.

Given the relatively small effects and sample size and the high correlation between **type** and **size** effects, covariate effects are captured well (Fig. 3): In the $n = 30$ scenario with the original starting parameterizations and positioning of the curves, effects are mostly estimated comparably accurately with mean squared errors (MSE) below 5% of the original additive predictor variance in the data (corresponding to about up to 8% of the variance of the original **type**-effect). Outliers are likely due to uncertainty in the choice of linear or non-linear (“smooth”) effects. Under random warping and positioning of the curves, errors of the **type**-effect increase to a median MSE of 10% of the total original additive predictor variance. Although with distinctly smaller MSE, it is evident that random transformations affect the estimation of the effects to some extent also in scenarios based on the entire original data ($n = 40$ and $n = 80$). Here, the more complex transformation given by random warping shows a larger impact than random positioning, leading to larger MSE.

Tracing error resulting from the random transformations to its root, leads to the estimation of the intercept $[\psi]$ as overall shape mean of the curves as its cause. Our applied default estimator $[\hat{\psi}]$ shows a good performance in terms of $d^2([\hat{\psi}], [\psi]) \ll \sigma_0^2$ ranging mostly below 1% of the total variance $\sigma_0^2 = \frac{1}{n} \sum_{i=1}^n d^2([\hat{y}_i], [\psi])$ obtained from the original model fit. Yet, the starting parameterization still shows a strong effect, in the sense that without random warping the error decreases to nearly zero. Visual inspection shows that, while bottle proportions (and also the direction of the **type**-effect) are captured well, edges perceived as characteristic landmarks are slightly over-smoothed. As model effects take their origin at $[\hat{\psi}]$, this lack of detail is carried forward to model prediction and visualization. The over-smoothing behavior can be explained by the fact that $[\hat{\psi}]$ is based in turn on

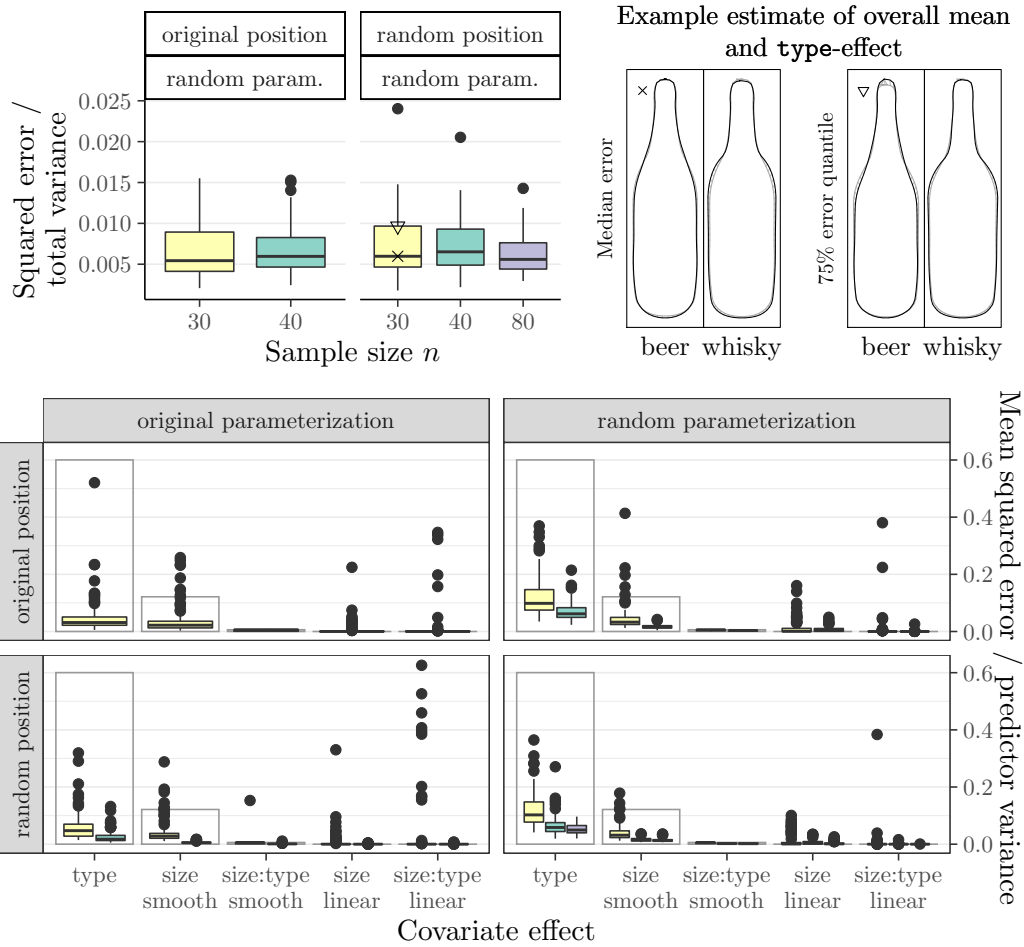


Figure 3: *Top left:* Distributions of intercept (overall mean) estimation accuracy (squared errors $d^2([\hat{\psi}], [\psi]) / \sigma_0^2$ relative to total variance) for simulation scenarios with random warping of bottle outline representatives y_1, \dots, y_n . For concise display, relative errors in scenarios without random warping are not shown, being very small (below $2 \cdot 10^{-4}$ for $n = 30$ and below $4 \cdot 10^{-5}$ for $n = 40$). *Top right:* Two example estimates of the bottle **type** effect (*black*) in front of the overall mean estimate (*gray*), corresponding to the depiction of the original effect in Fig. 1, for simulation runs marked with \times and Δ in the plot on the left. *Bottom:* MSE distributions for covariate effects in the model relative to the overall variance $\frac{1}{n} \sum_{i=1}^n \|f(\mathbf{x}_i)\|^2$ of the (centered) additive predictor, for simulation scenarios with and without random positioning and warping of recorded curves. MSEs are relative to the fit from Section 3.1 taken as true values. Bars reflecting the single effect variances $\frac{1}{n} \sum_{i=1}^n \|f_j(\mathbf{x}_i)\|^2$, $j = 1, \dots, n$, are added for individual comparison. For neither of the random transformations, only the $n = 30$ setting is depicted, reflecting the generalization error of the model with naively aligned curve data as underlying the original model fit. Other settings have zero error here by design.

elastic Riemannian L_2 -Boosting, where the preliminary intercept $[\hat{\psi}_0]$ does depend on the specific representatives, since its representative p_0 is estimated without warping/rotation alignment as L_C^2 -average of the SRV-transforms q_1, \dots, q_n of recorded curves. Mismatch in warping and rotation masks distinct curve features by averaging over miss-aligned representatives (compare also Stöcker et al., 2022,?). Hence, although in principle the original $[\psi]$ could be retrieved from a different starting point p_0 , lacking features in p_0 to align to can render it difficult to fully estimate these features in $[\hat{\psi}]$.

Various possibilities exist to avoid this problem by choosing a better starting point that exhibits the desired features: a) in our experience also from Steyer et al. (2021), using similar initial parameterizations (such as constant-speed parameterization) in curves y_1, \dots, y_n already yields a well-working default starting point $[\hat{\psi}_0]$ for the estimator $[\hat{\psi}]$ utilized in this paper, as illustrated by the natural bottle appearance in the original model fit in 3.1. b) in particular for sparsely recorded curves, the estimator $[\hat{\psi}_{\text{eFP}}]$ of the elastic full Procrustes shape mean $[\psi_{\text{eFP}}]$ proposed by Stöcker et al. (2022) and implemented in the R package `elastes` (github.com/mpff/elastes) presents an attractive choice for $[\psi_0]$ due to its fit based on Hermitian covariance smoothing. c) if a good template curve is available, it can be directly used to represent $[\psi_0]$. Such a curve might be simply selected from the dataset. d) as an alternative to our overall elastic shape mean estimation approach, $[\hat{\psi}]$ might be obtained from the implementation in R package `fdasrvf` (Tucker, 2017). An approach to landmark-constrained elastic shape mean estimation was proposed by Strait et al. (2017).

Nonetheless, we keep the straightforward estimator here to illustrate the role of the intercept: as it presents the starting point of the model fit, prediction and visualization, it has a strong impact on the model results. Inaccuracy in details of the fit of the intercept are likely carried forward. In general, this is not problematic, since the intercept can be estimated very accurately as overall shape mean. However, to capture also shape details well, it is recommended to ensure that the fit of the overall shape mean is fully satisfying, which requires a starting point that contains all important features of the shape.

4 Discussion

Depending on the data problem, different modifications of the presented elastic regression approach for shapes of plane curves might be of interest: further development will be needed to model (non-symmetric) closed curves with closedness explicitly integrated into the model, while regression for open curves is already covered in our framework. Instead of modeling the shape of the curves, it might also desirable to model the “form” (or size-and-shape) of curves without scale invariance (analogously to Stöcker et al., 2022), or to model curves with a fixed coordinate system without shape invariances. Integrating different intercept options mentioned in Section 3.2 into our software package will improve flexible usability. The architecture of our R package `manifoldboost` is designed to simplify modular extension to such variations in the response geometry and model fit, adding to the modular covariate effect specification borrowed from scalar additive models. Finally, applying our approach to further data sets will illustrate flexibility and usefulness of the proposed model framework for analyzing data problems of scientific interests.

Appendix

A.1 Closing symmetric curves

To avoid non-linear constraints guaranteeing closedness of a curve $\mu \in \mathcal{AC}^*([-1, 1])$ via its SRV-transform m , we argue that unconstrained estimation already promises satisfactory results when modeling symmetric curves, since in this case, μ can be differentially extended by a line segment to obtain a closed curve under mild assumptions. For a symmetric shape $[\mu]$ of μ , we assume without loss of generality that its SRV-transform m is even (in general it could be rotated or based on a different parameterization). Modeling μ continuously differentiable, m is also assumed closed and continuous in the following. In this case, also $\dot{\mu}$ is even and closed, and the back-transform $\tilde{\mu} = \int_0^t m(s) ds$ is odd. For simplicity and without loss of generality, we assume $\mu = \tilde{\mu}$. Our aim is to close the gap $\delta = \mu(-1) - \mu(1)$ by a line segment such that the resulting curve μ^* is differentiable. Lemma 1 below yields that under the given assumptions $\delta \in \mathbb{R}$, $\dot{\mu}(0) \in \mathbb{R}$ and $\dot{\mu}(1) = \dot{\mu}(-1) \in \mathbb{R}$. Hence, when considering the two symmetric sides of the curve described by $\mu|_{[0,1]}$ and $\mu|_{[-1,0]}$ restricting μ to the respective interval, directions at the endpoints of the sides of μ are all orthogonal to the imaginary axis presenting the symmetry axis. Hence, differentiable closing will be possible if $\dot{\mu}(1)$ and $\dot{\mu}(0)$ have the right combination of signs, for which three cases have to be distinguished (assuming a parameterization with $\dot{\mu}(1) \neq 0$ and $\dot{\mu}(0) \neq 0$ and a relevant gap $\delta \neq 0$):

If $\delta \dot{\mu}(1) > 0$, μ can be directly extended to a differentiable closed curve $\mu^* : [-1 - \frac{\delta}{2\dot{\mu}(1)}, 1 + \frac{\delta}{2\dot{\mu}(1)}] \rightarrow \mathbb{C}$ with

$$\mu^*(t) = \begin{cases} \mu(t) & \text{for } t \in [-1, 1] \\ \dot{\mu}(1)(t-1) + \mu(1) & \text{for } t > 1 \\ \dot{\mu}(1)(t+1) + \mu(-1) & \text{for } t < -1 \end{cases}$$

If $\delta \dot{\mu}(0) < 0$, the two sides $\mu|_{[0,1]}$ and $\mu|_{[-1,0]}$ of the symmetric curve can be shifted to close the curve at $-1/1$ while opening it at 0. Then, we may differentially extend them at 0 to obtain a closed curve $\mu^* : [-1 + \frac{\delta}{2\dot{\mu}(0)}, 1 - \frac{\delta}{2\dot{\mu}(0)}] \rightarrow \mathbb{C}$ as

$$\mu^*(t) = \begin{cases} \mu(t - \frac{\delta}{2\dot{\mu}(0)}) - \frac{\delta}{2} & \text{for } t \in [-1 + \frac{\delta}{2\dot{\mu}(0)}, \frac{\delta}{2\dot{\mu}(0)}] \\ \mu(t + \frac{\delta}{2\dot{\mu}(0)}) + \frac{\delta}{2} & \text{for } t \in [-\frac{\delta}{2\dot{\mu}(0)}, 1 - \frac{\delta}{2\dot{\mu}(0)}] \\ \dot{\mu}(0)t & \text{otherwise.} \end{cases}$$

Although involving the shift, the second option in fact corresponds to the first after simple reparameterization as $\mu'(t) = \mu(t-1)$ for $t \in [0, 1]$ and $\mu'(t) = \mu(t+1)$ for $t \in [-1, 1)$, switching $t = 0$ with $t = \pm 1$.

If $\delta \dot{\mu}(1) < 0$ and $\delta \dot{\mu}(0) > 0$, μ cannot be differentially closed by a line segment, since $\dot{\mu}(1)$ points in the same direction as $\dot{\mu}(0)$ and away from 0. We do not implement a constraint to avoid this case, since we would hardly expect to encounter in practice: being bound to values in \mathbb{R} by the symmetry constraint, $\dot{\mu}(1)$ and $\dot{\mu}(0)$ can only point into the right direction for closing or precisely into the opposite direction. This makes it unlikely that, when all curves y_1, \dots, y_n in the data are closed and, hence, in line with the constraint, $\dot{\mu}(1)$ and $\dot{\mu}(0)$ still point into the wrong direction for closing.

Lemma 1. *For an even SRV-transform $m : [-1, 1] \rightarrow \mathbb{C}$ of a plane curve $\mu \in \mathcal{AC}^*([-1, 1])$,*

- i) the back-transform $\tilde{\mu}(t) = \int_0^t m(s)|m(s)| ds$ is odd.*

ii) the gap between the endpoints of μ is a real number $\delta = \mu(-1) - \mu(1) \in \mathbb{R}$.

iii) if m is closed, we have $m(1) \in \mathbb{R}$ and, hence, also $\dot{\mu}(1) \in \mathbb{R}$.

Proof. i) follows by plugging $m(t)^\dagger = m(-t)$ into the definition of $\tilde{\mu}$:

$$\begin{aligned}\tilde{\mu}(t)^\dagger &= \int_0^t m(s)^\dagger |m(s)^\dagger| ds = \int_0^t m(-s) |m(-s)| ds \\ &= - \int_0^{-t} m(s) |m(s)| ds = -\tilde{\mu}(-t).\end{aligned}$$

To see ii), first note that $\mu(t) = \tilde{\mu}(t) + z$ for some $z \in \mathbb{C}$ and, thus, $\delta = \tilde{\mu}(-1) - \tilde{\mu}(1)$. Hence,

$$\begin{aligned}2 \Im(\delta) &= \delta - \delta^\dagger = \tilde{\mu}(-1) - \tilde{\mu}(1) - (\tilde{\mu}(-1)^\dagger - \tilde{\mu}(1)^\dagger) \\ &= -\tilde{\mu}(1)^\dagger - \tilde{\mu}(1) + \tilde{\mu}(1) + \tilde{\mu}(1)^\dagger = 0\end{aligned}$$

by repeatedly applying i). iii) immediately follows from $m(1)^\dagger \stackrel{\text{even}}{=} m(-1) \stackrel{\text{closed}}{=} m(1)$. \square

A.2 Simulating curves with random warping and positioning

To control variability of random transformations applied in the simulation study to a moderate amount (exceeding what we expect to find in typical data but not completely arbitrary), we draw sampling points of a randomly transformed version \tilde{y}_i of an original curve $y_i : [0, 1] \rightarrow \mathbb{C}$, given by the sample polygon of the i th curve in our original data from 3.1 with the corresponding initial parameterization on $[0, 1]$, as

$$\tilde{y}_i(t_l) = \lambda_i \exp(\omega_i) y_i(\gamma_i(t_l)) + z_i \quad (l = 1, \dots, 100)$$

where $\lambda_i > 0$, $\omega_i \in \mathbb{R}$, $z_i = z_i^{\Re} + z_i^{\Im} \sqrt{-1} \in \mathbb{C}$, and $0 = \gamma_i(t_1) < \dots < \gamma_i(t_{100}) = t_{100}$ are randomly drawn independently for the i th curve in the simulated data corresponding to the i th curve in the original dataset. In scenarios with random positioning, we draw

$$\begin{aligned}\lambda_i &\sim \text{Gamma}(100, 100) \quad (\text{given with shape and rate parameter}) \\ \omega_i &\sim \text{N}\left(0, \frac{\pi^2}{400}\right), \quad z_i^{\Re} \sim \text{N}(0, \sigma_{\Re}^2), \quad z_i^{\Im} \sim \text{N}(0, \sigma_{\Im}^2)\end{aligned}$$

where $\mathbb{E}(\lambda_i) = 1$ with standard deviation $\text{sd}(\lambda_i) = 0.1$, the standard deviation of ω_i corresponds to a rotation about ca. 9 degrees, and σ_{\Re}^2 and σ_{\Im}^2 are selected to reflect the standard deviation of the evaluations of the original curve along the real and imaginary axis, respectively. In scenarios with random warping, we draw $\gamma(t_l) = \frac{\sum_{l'=2}^l \Delta_{l'}}{\sum_{l'=2}^{100} \Delta_{l'}} t_{100}$ with

$$\Delta_l \sim \text{Gamma}(3, 3)$$

such that $\mathbb{E}(\Delta_l) = 1$ and $\text{sd}(\Delta_l) = \frac{1}{3}$. Figure 4 illustrates the resulting variability with random positioning and warping in different samples of one example bottle outline.

References

Afsari, B. (2011). Riemannian L^p center of mass: existence, uniqueness, and convexity. *Proceedings of the American Mathematical Society* 139(2), 655–673.

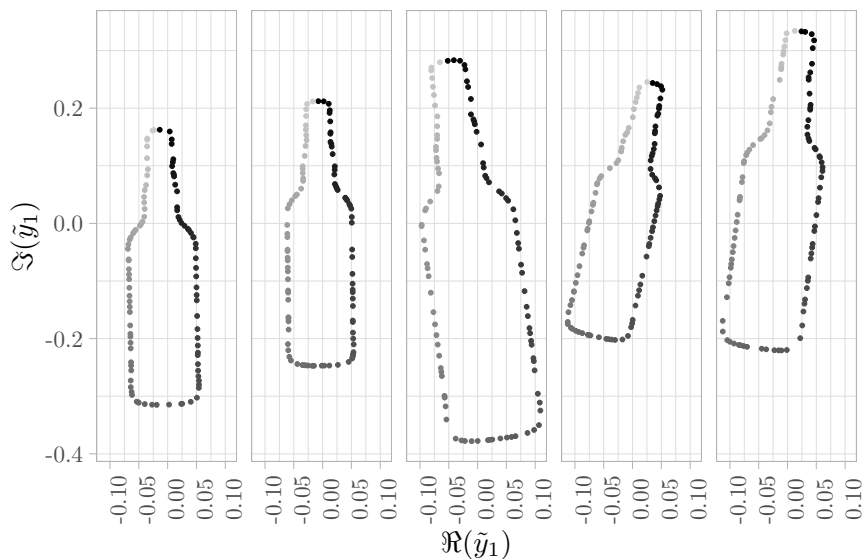


Figure 4: Five example draws of the first curve in the original dataset applying random warping and positioning. The parameterization of the 100 sample points per curve is always initialized with a regular grid starting and ending at the top of the bottle.

- Ahn, K., J. Derek Tucker, W. Wu, and A. Srivastava (2018). Elastic handling of predictor phase in functional regression models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 324–331.
- Bonhomme, V., S. Picq, C. Gaucherel, and J. Claude (2014). Momocs: Outline analysis using R. *Journal of Statistical Software* 56(13), 1–24.
- Brockhaus, S., A. Fuest, A. Mayr, and S. Greven (2018). Signal regression models for location, scale and shape with an application to stock returns. *Journal of the Royal Statistical Society: Series C* 67(3), 665–686.
- Brockhaus, S., D. Rügamer, and S. Greven (2020). Boosting functional regression models with fdboost. *Journal of Statistical Software* 94, 1–50.
- Brockhaus, S., F. Scheipl, and S. Greven (2015). The Functional Linear Array Model. *Statistical Modelling* 15(3), 279–300.
- Bruveris, M. (2016). Optimal reparametrizations in the square root velocity framework. *SIAM Journal on Mathematical Analysis* 48(6), 4335–4354.
- Bühlmann, P. and T. Hothorn (2007). Boosting algorithms: Regularization, prediction and model fitting (with discussion). *Statistical Science* 22(4), 477–505.
- Bühlmann, P. and B. Yu (2003). Boosting with the L2 loss: regression and classification. *Journal of the American Statistical Association* 98(462), 324–339.

- Cornea, E., H. Zhu, P. Kim, J. G. Ibrahim, and the Alzheimer’s Disease Neuroimaging Initiative (2017). Regression models on Riemannian symmetric spaces. *Journal of the Royal Statistical Society: Series B* 79(2), 463–482.
- Dryden, I. L. and K. V. Mardia (2016). *Statistical Shape Analysis: With Applications in R*. John Wiley & Sons.
- Fahrmeir, L., T. Kneib, S. Lang, and B. Marx (2013). Regression models. In *Regression*, pp. 21–72. Springer.
- Fletcher, P. T. (2013). Geodesic regression and the theory of least squares on Riemannian manifolds. *International Journal of Computer Vision* 105(2), 171–185.
- Fréchet, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. In *Annales de l’institut Henri Poincaré*, Volume 10, pp. 215–310.
- Greven, S. and F. Scheipl (2017). A general framework for functional regression modelling (with discussion and rejoinder). *Statistical Modelling* 17(1-2), 1–35 and 100–115.
- Guo, M., J. Su, L. Sun, and G. Cao (2020). Statistical regression analysis of functional and shape data. *Journal of Applied Statistics* 47(1), 28–44.
- Hadjipantelis, P. Z., J. A. D. Aston, H. G. Müller, and J. P. Evans (2015). Unifying amplitude and phase analysis: A compositional data approach to functional multivariate mixed-effects modeling of mandarin chinese. *Journal of the American Statistical Association* 110(510), 545–559.
- Hadjipantelis, P. Z., J. A. D. Aston, H. G. Müller, and J. Moriarty (2014). Analysis of spike train data: A multivariate mixed effects model for phase and amplitude. *Electronic Journal of Statistics* 8, 1797–1807.
- Hastie, T. and R. Tibshirani (1986). Generalized additive models. *Statistical Science*, 297–310.
- Hofner, B., T. Hothorn, T. Kneib, and M. Schmid (2011). A framework for unbiased model selection based on boosting. *Journal of Computational and Graphical Statistics* 20(4), 956–971.
- Hofner, B., T. Kneib, and T. Hothorn (2016). A unified framework of constrained regression. *Statistics and Computing* 26(1), 1–14.
- Hothorn, T., P. Bühlmann, T. Kneib, M. Schmid, and B. Hofner (2010). Model-based boosting 2.0. *Journal of Machine Learning Research* 11, 2109–2113.
- Huckemann, S., T. Hotz, and A. Munk (2010). Intrinsic MANOVA for Riemannian manifolds with an application to Kendall’s space of planar shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32(4), 593–603.
- Karcher, H. (1977). Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics* 30(5), 509–541.
- Kim, H. J., N. Adluru, M. D. Collins, M. K. Chung, B. B. Bendlin, S. C. Johnson, R. J. Davidson, and V. Singh (2014). Multivariate general linear models (mgglm) on Riemannian manifolds with applications to statistical analysis of diffusion weighted images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2705–2712.
- Klingenberg, W. (1995). *Riemannian geometry*. de Gruyter.

- Kurtek, S. and T. Needham (2018). Simplifying transforms for general elastic metrics on the space of plane curves. *arXiv preprint arXiv:1803.10894*.
- Lin, Z., H.-G. Müller, and B. U. Park (2020). Additive models for symmetric positive-definite matrices, riemannian manifolds and lie groups. *arXiv preprint arXiv:2009.08789*.
- Lutz, R. W. and P. Bühlmann (2006). Boosting for high-multivariate responses in high-dimensional linear regression. *Statistica Sinica*, 471–494.
- Marron, J., J. O. Ramsay, L. M. Sangalli, and A. Srivastava (Eds.) (2014). Statistics of time warpings and phase variations [special section]. *Electronic Journal of Statistics* 8(2), 1697–1939.
- Matuk, J., K. Bharath, O. Chkrebti, and S. Kurtek (2021). Bayesian framework for simultaneous registration and estimation of noisy, sparse, and fragmented functional data. *Journal of the American Statistical Association*, 1–17.
- Mayr, A., H. Binder, O. Gefeller, and M. Schmid (2014a). The evolution of boosting: From machine learning to statistical modelling. *Methods of information in medicine* 53, 419–27.
- Mayr, A., H. Binder, O. Gefeller, and M. Schmid (2014b). Extending statistical boosting: An overview of recent methodological developments. *Methods Inf Med* 53, 428–35.
- Mayr, A., B. Hofner, and M. Schmid (2012). The importance of knowing when to stop. *Methods of Information in Medicine* 51(02), 178–186.
- Morris, J. S. (2015). Functional Regression. *Annual Review of Statistics and its Applications* 2, 321–359.
- Nelder, J. A. and R. W. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)* 135(3), 370–384.
- Pennec, X. (2006). Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision* 25(1), 127–154.
- Petersen, A. and H.-G. Müller (2019). Fréchet regression for random objects with euclidean predictors. *The Annals of Statistics* 47(2), 691–719.
- Scheipl, F., J. Gertheiss, and S. Greven (2016). Generalized functional additive mixed models. *Electronic Journal of Statistics* 10(1), 1455–1492.
- Scheipl, F., A.-M. Staicu, and S. Greven (2015). Functional additive mixed models. *Journal of Computational and Graphical Statistics* 24(2), 477–501.
- Shi, X., M. Styner, J. Lieberman, J. G. Ibrahim, W. Lin, and H. Zhu (2009). Intrinsic regression models for manifold-valued data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 192–199. Springer.
- Srivastava, A., E. Klassen, S. H. Joshi, and I. H. Jermyn (2011). Shape analysis of elastic curves in Euclidean spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(7), 1415–1428.
- Srivastava, A. and E. P. Klassen (2016). *Functional and Shape Data Analysis*. Springer-Verlag.

- Stasinopoulos, M. D., R. A. Rigby, G. Z. Heller, V. Voudouris, and F. De Bastiani (2017). *Flexible regression and smoothing: using GAMLSS in R*. CRC Press.
- Steyer, L. (2021). *elasdics: Elastic Analysis of Sparse, Dense and Irregular Curves*. R package version 0.1.3.
- Steyer, L., A. Stöcker, and S. Greven (2021). Elastic analysis of irregularly or sparsely sampled curves. *arXiv preprint arXiv:2104.11039*.
- Stöcker, A., S. Brockhaus, S. Schaffer, B. von Bronk, M. Opitz, and S. Greven (2018). Boosting functional response models for location, scale and shape with an application to bacterial competition. *arXiv preprint arXiv:1809.09881*, <https://arxiv.org/abs/1809.09881>.
- Stöcker, A., M. Pfeuffer, L. Steyer, and S. Greven (2022). Elastic full procrustes analysis of plane curves via hermitian covariance smoothing. *arXiv preprint arXiv:2203.10522*.
- Stöcker, A., L. Steyer, and S. Greven (2022). Functional additive regression on shape and form manifolds of planar curves. *arXiv preprint arXiv:2109.02624*.
- Strait, J., S. Kurtek, E. Bartha, and S. N. MacEachern (2017). Landmark-constrained elastic shape analysis of planar curves. *Journal of the American Statistical Association* 112(518), 521–533.
- Thomas, J., A. Mayr, B. Bischl, M. Schmid, A. Smith, and B. Hofner (2018). Gradient boosting for distributional regression: faster tuning and improved variable selection via noncyclical updates. *Statistics and Computing* 28(3), 673–687.
- Tucker, J. D. (2017). *fdasrvf: Elastic Functional Data Analysis*. R package version 1.8.3.
- Tucker, J. D., J. R. Lewis, and A. Srivastava (2019). Elastic functional principal component regression. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 12(2), 101–115.
- Wood, S. (2017). *Generalized Additive Models: An Introduction with R* (2 ed.). Chapman and Hall/CRC.
- Zhu, H., Y. Chen, J. G. Ibrahim, Y. Li, C. Hall, and W. Lin (2009). Intrinsic regression models for positive-definite matrices with applications to diffusion tensor imaging. *Journal of the American Statistical Association* 104(487), 1203–1212.
- Ziezold, H. (1977). 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*, pp. 591–602. Springer.

7. Paper VI: Principal Component Analysis in Bayes Spaces for Sparsely Sampled Density Functions

In contrast to papers I to V, paper VI does not deal with any curve or shape space, but focuses on the analysis of one-dimensional probability density functions, considering them as equivalence classes with respect to rescaling (see Subsection 1.3.3). It proposes a novel approach to functional principal component analysis (FPCA, see Subsection 1.1.3) in Bayes spaces based on discrete samples drawn from each density. For modeling, the isometric isomorphism between the Bayes space and \mathbb{L}_0^2 , the space of square integrable functions integrating to zero is used, and the underlying functional densities are treated as latent variables in a maximum likelihood framework. Estimation is performed using a Monte Carlo expectation maximization (MCEM) algorithm. The paper demonstrates the applicability of the method for analyzing the distribution of maximum daily temperatures in Berlin over the last 70 years and the distribution of rental prices in the districts of Munich.

Contributing article:

Steyer, L. and Greven, S. (2023). Principal component analysis in Bayes spaces for sparsely sampled density functions. *arXiv pre-print*, arXiv:2309.11352

Declaration on personal contributions:

The author of this thesis has carried out major parts of the project independently, with important and detailed advice and discussions from Sonja Greven in a supporting role.

PRINCIPAL COMPONENT ANALYSIS IN BAYES SPACES FOR SPARSELY SAMPLED DENSITY FUNCTIONS

A PREPRINT

Lisa Steyer & Sonja Greven

21st September 2023

ABSTRACT

This paper presents a novel approach to functional principal component analysis (FPCA) in Bayes spaces in the setting where densities are the object of analysis, but only few individual samples from each density are observed. We use the observed data directly to account for all sources of uncertainty, instead of relying on prior estimation of the underlying densities in a two-step approach, which can be inaccurate if small or heterogeneous numbers of samples per density are available. To account for the constrained nature of densities, we base our approach on Bayes spaces, which extend the Aitchison geometry for compositional data to density functions. For modeling, we exploit the isometric isomorphism between the Bayes space and the L^2 subspace L_0^2 with integration-to-zero constraint through the centered log-ratio transformation. As only discrete draws from each density are observed, we treat the underlying functional densities as latent variables within a maximum likelihood framework and employ a Monte Carlo Expectation Maximization (MCEM) algorithm for model estimation. Resulting estimates are useful for exploratory analyses of density data, for dimension reduction in subsequent analyses, as well as for improved preprocessing of sparsely sampled density data compared to existing methods. The proposed method is applied to analyze the distribution of maximum daily temperatures in Berlin during the summer months for the last 70 years, as well as the distribution of rental prices in the districts of Munich.

1 Introduction

A classic task in statistics is to estimate the underlying density from sample data, since density functions can be used to describe the distribution of real-valued random variables. However, when not all observations are identically distributed, but constitute repeated draws from a set of density functions f_1, \dots, f_n , e.g. for n different individuals, distributional properties of these density functions themselves may be the actual target of a statistical analysis. Examples in which densities or distributions are considered the observational units exist in many different fields. These include the size distributions of different zooplankton in oceanology (Nerini and Ghattas, 2007), the distributions of firm size in econometrics (Huynh and Jacho-Chávez, 2010), mortality densities at different locations in epidemiology (Scimone et al., 2021), and the densities of glucose levels among several diabetes patients in medical research (Matabuena et al., 2021).

Probability density functions can be seen as a special case of functional data, but considering them as the unit of observation poses two major challenges. First, any density function must be non-negative and integrate to one to be valid. Second, in practice, density functions are often unobserved and accessible only through discrete samples. That is, for each density function f_i , $i = 1, \dots, n$ there is usually only an independent and identically distributed sample $x_{ij} \sim f_i$, $j = 1, \dots, m_i$ available. Our goal is to develop a Principal Component Analysis (PCA) for densities that can handle both of these challenges. A PCA for density data is of interest for several reasons. First, resulting estimates are useful for exploratory analyses to better understand the main modes of variation in density data. Second, the resulting dimension reduction allows to succinctly describe differences and trends in densities and the corresponding principal components (PCs) can be used as a parsimonious data-driven basis in subsequent analyses, as common in functional data analysis (Yao et al., 2005; Chiou and Li, 2007; Scheipl et al., 2015). Third, the reconstructed densities resulting from the PCA can also be used as improved preprocessing of sparsely sampled density data compared to existing methods, if subsequent analysis methods need observed or reconstructed density data as input (Scimone et al., 2021; Maier et al., 2022).

Existing research has primarily addressed one of the two challenges associated with studying densities as functions, while overlooking the other. Initially, Kneip and Utikal (2001) used functional principal component analysis (FPCA) in the unbounded space \mathbb{L}^2 of quadratic integrable functions without considering the density constraints. They did, however, account for discretely observed data by estimating the covariance surface based on the combined observations from all densities. In recent years, researchers have directed their attention towards incorporating the intrinsic geometric constraints in the space of density functions. The following discussion provides an overview of their work. However, when dealing with discretely observed data, their primary approach has been to estimate the observed densities through preprocessing steps, such as aggregating the data using histograms or kernel density estimates, and ignoring the reconstruction uncertainty in the further analysis.

To address the density constraints, several metrics have been considered for the space of probability density functions. In particular, the Wasserstein distance (Panaretos and Zemel, 2019) for probability measures is widely used, while the Fisher-Rao metric (Srivastava et al., 2007) imposes a manifold structure on the space of density functions. Although statistical analysis can be performed directly on manifolds (e.g. geodesic PCA in Wasserstein space (Bigot et al., 2013)), it is often more convenient to map the densities to a space with a simpler structure, perform statistical analysis there, and then back-transform the results to the original density space. Various transformations have been considered, such as the log-hazard and log-quantile density transformations (Petersen and Müller, 2016) for mapping the density functions to a Hilbert space. In particular, Hron et al. (2016) used the centered log-ratio (clr) transformation to obtain FPCA for densities. The clr transformation is particularly useful here because it defines a one-to-one mapping between the squared-log integrable (proper and improper) density functions and the separable Hilbert space \mathbb{L}_0^2 , which represents the space of square-integrable functions that integrate to zero. This means that the clr transformation also induces a Hilbert space structure on the space of squared-log integrable (proper and improper) density functions, which is called Bayes Hilbert space (Egozcue et al., 2006; van den Boogaart et al., 2014).

This paper develops Functional Principal Component Analysis (FPCA) in the Bayes spaces for the setting where only samples from the set of densities of interest are available. Our approach utilizes the observed data directly for calculations instead of estimating the underlying densities beforehand in a two-step approach. Estimating density f_i becomes particularly challenging when dealing with small sample sizes m_i in each observation unit. To address this, a two-step approach is proposed by Qiu et al. (2022) for cases with heterogeneous sample sizes. Firstly, the underlying process is estimated based on a subsample where each density is densely observed. Then, the remaining densities are estimated using the process estimated in the first step as a prior. However, this approach is only feasible when there exists a representative subset of densities that have been densely observed.

The advantage of our approach is its applicability even when all densities are sparsely observed. To achieve this, we incorporate the observed data of all densities in a maximum likelihood framework, treating the underlying densities as latent variables. To estimate the model, we utilize the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). As the expectation step in our framework is not analytically solvable, requiring a Monte Carlo approximation, we base the estimation of our model on the so-called Monte Carlo Expectation-Maximization (MCEM) adaptation of the EM algorithm (Wei and Tanner, 1990).

We organize our contributions as follows. First, in Section 2, we introduce the latent density model and develop an MCEM algorithm suitable for this scenario. Then, in Section 3, we apply our methodology in two different applications. First, in the context of climate change and changing (extreme) temperatures in particular, we study how the distribution of maximum temperatures during the summer months in Berlin has evolved over the last 70 years. Second, we look at the distribution of rental prices in the different districts of Munich. In Section 4, we use a simulation to demonstrate that our method is particularly effective for analyzing densities when few observations drawn from them are available. Finally, we conclude the paper with a discussion in Section 5.

2 Principal component analysis (PCA) for densities based on individual samples

We consider densities as elements of the Bayes Hilbert space, which has proven to be a valuable framework for modeling densities. In order to perform Principal Component Analysis (PCA) in the Bayes Hilbert space, which is a separable Hilbert space, we employ the Karhunen-Loève decomposition. While PCA was originally developed as a dimension reduction tool for finite-dimensional data, the Karhunen-Loève decomposition extends this concept to infinite-dimensional Hilbert spaces (Hsing and Eubank, 2015).

2.1 PCA in Bayes Hilbert spaces

We start with reviewing the structure of the Bayes Hilbert space. For simplicity and as it seems natural in most applications, we restrict ourselves in this work mainly to densities with respect to the Lebesgue measure λ defined on a compact interval $I \subset \mathbb{R}$ although the construction can be done for general measures (van den Boogaart et al., 2014). In Subsection 2.4, we briefly discuss how the case of compositional data can be treated in a similar way.

Theorem 2.1 (Bayes Hilbert space (Egozcue et al., 2006)). *Let $B = \{f = \exp(g) | g \in \mathbb{L}^2(I)\}$ and consider the equivalence relation $f_1 \sim f_2 \Leftrightarrow \exists \alpha > 0 : f_1 = \alpha f_2$ for $f_1, f_2 \in B$. Denote by $\mathcal{B} = B/\sim$ the set of equivalence classes $[f]$ with $f \in B$. Then \mathcal{B} equipped with the operations*

$$\oplus \text{ (addition) given by the perturbation operator } [f_1] \oplus [f_2] = [f_1 \cdot f_2] \text{ for all } [f_1], [f_2] \in \mathcal{B},$$

$$\odot \text{ (scalar multiplication) given by the powering operation } \alpha \odot [f] = [f^\alpha] \text{ for all } [f] \in \mathcal{B}, \alpha \in \mathbb{R} \text{ and}$$

$$\langle \cdot, \cdot \rangle_{\mathcal{B}} \text{ the scalar product given via } \langle [f_1], [f_2] \rangle_{\mathcal{B}} = \frac{1}{2|I|} \int_I \int_I \log \left(\frac{f_1(x)}{f_1(y)} \right) \log \left(\frac{f_2(x)}{f_2(y)} \right) dx dy \text{ for all } [f_1], [f_2] \in \mathcal{B}$$

is a separable Hilbert space.

As a separable Hilbert space the Bayes Hilbert space is isometrically isomorphic to any other infinite dimensional separable Hilbert space, in particular \mathcal{B} is isometrically isomorphic to \mathbb{L}_0^2 , the space of square-integrable functions integrating to zero, via the centered log-ratio transformation.

Lemma 2.2 (Centered log-ratio transformation). *The centered log-ratio (clr) transformation*

$$\text{clr} : \mathcal{B} \rightarrow \mathbb{L}_0^2, \quad [f] \mapsto \log(f) - \frac{1}{|I|} \int_I \log(f(x)) dx \quad (1)$$

is a bijective isometry with the inverse given via $\text{clr}^{-1}(g) = [\exp(g)]$ for all $g \in \mathbb{L}_0^2$.

In particular, the clr transformation is well defined as $\text{clr}([\alpha f]) = \alpha \text{clr}([f])$ for all $f \in \mathcal{B}$ and $\alpha \in \mathbb{R}$. A detailed proof for Lemma 2.2 can be found in Appendix A.1 (compare with van den Boogaart et al. (2014) for densities on general measure spaces). This identification of the Bayes space \mathcal{B} with \mathbb{L}_0^2 allows statistical modeling to be performed in \mathbb{L}_0^2 instead of directly in \mathcal{B} , allowing the application of techniques developed for functional data. In particular, Hron et al. (2016) used this correspondence to introduce PCA in Bayes space, while Scimone et al. (2021) and Maier et al. (2022) exploit its use for regression purposes.

Similarly, we will achieve a principal component decomposition of observed densities f_1, \dots, f_n , $n \in \mathbb{N}$ via considering them being the back-transforms of realizations g_1, \dots, g_n of a stochastic process $\mathcal{G} = \{G(x)\}_{x \in I} \subset \mathbb{L}_0^2$ characterized by its mean function $\mu(x) = \mathbb{E}(G(x))$ and covariance kernel $K(x_1, x_2) = \text{Cov}(G(x_1), G(x_2))$ for all $x, x_1, x_2 \in I$. The Karhunen-Loève decomposition (Karhunen, 1946; Loève, 1946) then yields the functional principal component representation

$$G(x) = \mu(x) + \sum_{k=1}^{\infty} Z_k \varphi_k(x) \quad (2)$$

where φ_k , $k \in \mathbb{N}$ are the orthonormal eigenfunctions of the covariance operator $\mathbb{L}_0^2 \rightarrow \mathbb{L}_0^2$, $g \mapsto \int_I K(x_1, \cdot) g(x_1) dx_1$ associated with the covariance kernel K and uncorrelated principal component scores Z_k of decreasing importance, with $\mathbb{E}(Z_k) = 0$ and $\text{Var}(Z_k) = \sigma_k^2$ the corresponding eigenvalues, $\sigma_1^2 \geq \sigma_2^2 \geq \dots \geq 0$. For more details on this decomposition for second-order stochastic processes refer to Hsing and Eubank (2015).

For a given sample of (fully observed) functions g_1, \dots, g_n , $n \in \mathbb{N}$ the unknown parameters φ_k, Z_k of this process could then be estimated via computing the eigendecomposition of the sample covariance operator associated with the sample covariance kernel $\hat{K}_n(x_1, x_2) = \frac{1}{n} \sum_{i=1}^n (g_i(x_1) - \hat{\mu}(x_1))(g_i(x_2) - \hat{\mu}(x_2))$ with $\hat{\mu}(x) = \frac{1}{n} \sum_{i=1}^n g_i(x)$ for all $x, x_1, x_2 \in I$. Thus, the eigenfunctions φ_k could be estimated as the eigenfunctions $\hat{\varphi}_k$ of the sample covariance and the distribution of Z_k as the empirical distribution of the factor loadings $z_{ik} = \int_I (g_i(x) - \hat{\mu}(x)) \hat{\varphi}_k dx$, $i = 1, \dots, n$ for all $k = 1, \dots, N$ where N is the number of non-zero eigenvalues of the sample covariance.

The correspondence of \mathbb{L}_0^2 and the Bayes Hilbert space via the clr transformation (Lemma 2.2) gives an analog principal component decomposition to Equation (2) for densities. The process $\{\text{clr}^{-1}(G)(x)\}_{x \in I}$ is given as

$$\text{clr}^{-1}(G)(x) = \text{clr}^{-1}(\mu)(x) \oplus \bigoplus_{k=1}^{\infty} \exp(Z_k) \odot \text{clr}^{-1}(\varphi_k)(x) \quad (3)$$

inheriting the properties of the decomposition of \mathcal{G} in \mathbb{L}_0^2 via the clr transformation (Equation (1)). Namely, we obtain orthonormal eigenfunctions since $\langle \text{clr}^{-1}(\varphi_k), \text{clr}^{-1}(\varphi_l) \rangle_{\mathcal{B}} = \langle \varphi_k, \varphi_l \rangle_{\mathbb{L}_2} = 0$ for all $k \neq l$ and $\|\varphi\|_{\mathcal{B}} = \|\varphi\|_{\mathbb{L}_2} = 1$ for all $k \in \mathbb{N}$ and principal components scores $\exp(Z_k)$, $k \in \mathbb{N}$ with $Z_k \perp Z_l$ and $\mathbb{E}(Z_k) = 0$ for all $k, l \in \mathbb{N}$. If we additionally assume that G is Gaussian, this also implies that the principal component scores $\exp(Z_k)$, $k \in \mathbb{N}$ are independent (and therefore also uncorrelated) and we can compute their expectation as the evaluation of the moment generating function, i.e. $\mathbb{E}(\exp(Z_k)) = \exp(0.5\sigma_k^2)$, where $\sigma_k^2 = \text{Var}(Z_k)$.

Note that elements in \mathcal{B} are equivalence classes with respect to scalar multiplication. As a result, they either possess a unique representative which is a proper density function, or solely consist of improper densities. In practical applications, the focus is often on modeling proper density functions. Consequently, it becomes necessary to impose restrictions on \mathbb{L}_0^2 accordingly. Specifically, only those functions $g \in \mathbb{L}_0^2$ that satisfy $\int_I \exp(g(x))dx < \infty$ are suitable for representing proper densities. This requirement can be effectively met, for example, by using spline representations for elements of \mathbb{L}_0^2 (Machalová et al., 2021; Maier et al., 2022) since the exponential transformation of a spline is bounded, guaranteeing integrability over I . Furthermore, if we consider a finite set of bounded density functions f_1, \dots, f_n , such as those obtained from kernel density estimates or histograms, and apply principal component decomposition in the Bayes space (3), we are modeling only functions in the span of the data, which are also bounded. Consequently, in such cases, the eigenfunctions (and any linear combinations of them) are inherently proper densities.

Each equivalence class $[f]$ in the subset of \mathcal{B} constructed such that it contains only proper densities can be uniquely identified with the element $\tilde{f} \in [f]$ that satisfies $\int_I \tilde{f}(x)dx = 1$. To simplify the notation, albeit with a slight misuse, we will refer to the equivalence class by this particular element, $\tilde{f} = [f]$ in the following discussion. Hence, if $[\exp(g)]$ contains proper density functions, i.e. if $\int_I \exp(g(x))dx$ is finite, we denote by

$$\text{clr}^{-1}(g) = \frac{\exp(g)}{\int_I \exp(g(x))dx} \quad (4)$$

the back-transformed element in \mathcal{B} under the inverse clr transformation.

2.2 Likelihood formulation assuming latent densities

We have seen in the previous section that the correspondence of \mathbb{L}_0^2 and the Bayes Hilbert space via the clr transformation provides a convenient approach for conducting PCA on fully observed densities, since it allows estimation via first transforming the density functions to the Hilbert space \mathbb{L}^2 using the clr transformation, and then performing the principal component decomposition in this well-known function space. This procedure for fully observed density functions was previously suggested by Hron et al. (2016). However, their approach motivates the so-called simplicial principal component analysis solely as a maximization problem. In other words, they seek to find the projections of the observed densities that maximize the variance along their directions. While this leads to the same decomposition for fully observed density functions when the covariance kernel is estimated using the sample covariance kernel, the stochastic process perspective becomes especially useful in the more common scenario we have in mind.

We focus on analyzing densities that are neither directly observable nor can be satisfactorily estimated from observed data as a preprocessing step. Instead, we assume that we have access to samples x_{i1}, \dots, x_{im_i} , where $m_i \in \mathbb{N}$, drawn from each probability distribution with density f_i , where $i = 1, \dots, n$. Our objective is then to conduct Maximum-Likelihood estimation for the parameters μ and K of the underlying process \mathcal{G} in \mathbb{L}_0^2 based on these samples. By estimating these parameters, we can compute the eigenvalues and eigenfunctions of the estimated covariance operator, which allows us to obtain the principal component decomposition which directly yields the principal component decomposition in the Bayes space via the inverse clr transformation. To accomplish this, we need to make a distributional assumption for G . For simplicity, we assume that G follows a Gaussian process with a finite Karhunen-Loève decomposition. More precisely, we assume the following model.

Definition 2.3 (Latent density model). *Let $GP(\mu, K)$ be a Gaussian process with mean function μ and covariance kernel K taking values in a finite dimensional subspace $\mathcal{H} \subset \mathbb{L}_0^2$ consisting of bounded functions. Then we assume the following data generating process*

$$X_{ij} \stackrel{i.i.d.}{\sim} \text{clr}^{-1}(G_i) = \frac{\exp(G_i)}{\int_I \exp(G_i(x))dx}$$

with G_i being independent replicates of $GP(\mu, K)$ for all $i = 1, \dots, n$, $n \in \mathbb{N}$, $j = 1, \dots, m_i$.

Assuming a finite dimensional subspace \mathcal{H} is not restrictive in practice, where observed data always lies in a finite dimensional subspace. This assumption allows us to apply maximum likelihood theory, since in this case the parameters μ and K of the underlying Gaussian process can be described by a finite set of real-valued parameters and the likelihood has a similar form as the marginal likelihood of a mixed model. More precisely, we can formulate the following corollary (proof in Appendix A.2).

Corollary 2.4. *For any orthonormal basis $e_1, \dots, e_N \subset \mathbb{L}^2$ with $\mathcal{H} \subseteq \text{span}\{e_1, \dots, e_N\}$ we have that $G \stackrel{i.i.d.}{\sim} GP(\mu, K)$ is equivalent to $G = \sum_{k=1}^N \theta_k e_k$ with $\boldsymbol{\theta} = (\theta_1, \dots, \theta_N)^T \stackrel{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{\nu}, \boldsymbol{\Sigma})$ for $\mu = \sum_{k=1}^N \nu_k e_k$ and $K(x_1, x_2) = \sum_{k=1}^N \sum_{l=1}^N e_k(x_1) e_l(x_2) \Sigma_{kl}$, where $\boldsymbol{\Sigma} = (\Sigma_{kl})_{k,l=1, \dots, N}$ and $\boldsymbol{\nu} = (\nu_1, \dots, \nu_N)^T$.*

If $\mathbf{v}_1, \dots, \mathbf{v}_N$ are the eigenvectors of $\boldsymbol{\Sigma}$ with corresponding eigenvalues $\sigma_1^2, \dots, \sigma_N^2$ then $\varphi_l = \sum_{k=1}^N \nu_k e_k$, $l = 1, \dots, N$ are the eigenfunctions of the covariance operator given by the covariance function K with the same eigenvalues $\sigma_1^2, \dots, \sigma_N^2$, where $\mathbf{v}_l = (v_{l1}, \dots, v_{lN})$ for all $l = 1, \dots, N$.

With this equivalence, the latent density model 2.3 can also be written as

$$X_{ij} \stackrel{i.i.d.}{\sim} \text{chr}^{-1}(G_i) = \frac{\exp(G_i)}{\int_I \exp(G_i(x)) dx} \quad \text{with} \quad G_i = \sum_{k=1}^N \theta_{ik} e_k \quad \text{and} \quad \boldsymbol{\theta}_i = (\theta_{i1}, \dots, \theta_{iN}) \stackrel{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{\nu}, \boldsymbol{\Sigma}) \quad (5)$$

and estimation of the parameters μ and K is equivalent to estimation of $\boldsymbol{\nu}$ and $\boldsymbol{\Sigma}$. Note that we do not assume $\text{span}\{e_1, \dots, e_N\} = \mathcal{H}$, we only need to cover \mathcal{H} , since for any finite basis $\{e_1, \dots, e_N\}$ in \mathbb{L}^2 the sum to zero constrain carries over to the parameters $\boldsymbol{\nu}$ and $\boldsymbol{\Sigma}$. That means if $\int_I e_k(x) dx = \int_I e_l(x) dx$ for all $k, l = 1, \dots, N$, one just needs to ensure that the entries in $\boldsymbol{\nu}$ as well as the entries of all eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_N$ of $\boldsymbol{\Sigma}$ sum to zero.

Corollary 2.4 also implies that the maximum likelihood estimators will be asymptotically consistent, if model 2.3 is correctly specified, that is the image \mathcal{H} of the the process G is actually in $\text{span}\{e_1, \dots, e_N\}$. An interesting question for future research is whether in the case of misspecification, in particular when the image space \mathcal{H} is not finite, a sequence of finite-dimensional subspaces of \mathbb{L}^2 can be constructed such that the corresponding estimators are asymptotically consistent in this case as well.

To estimate the finite dimensional parameters $\boldsymbol{\nu}$ and $\boldsymbol{\Sigma}$ of the latent density model via maximum likelihood, we need to maximise the likelihood function given the realizations $\mathbf{x}_i = (x_{i1}, \dots, x_{im_i})^T$ from the random sample $\mathbf{X}_i = (X_{i1}, \dots, X_{im_i})^T$, $i = 1, \dots, n$. Let $G_i = \sum_{k=1}^N \theta_{ik} e_k$ and $\boldsymbol{\theta}_i = (\theta_{i1}, \dots, \theta_{iN})^T$ for all $i = 1, \dots, n$. Then the marginal likelihood for the parameters $\boldsymbol{\nu}$ and $\boldsymbol{\Sigma}$ is given as

$$L(\boldsymbol{\nu}, \boldsymbol{\Sigma} | \mathbf{x}_1, \dots, \mathbf{x}_n) = \prod_{i=1}^n \int_{\mathbb{R}^N} \frac{\exp(\sum_{j=1}^{m_i} \sum_{k=1}^N \theta_{ik} e_k(x_{ij})) p(\boldsymbol{\theta}_i | \boldsymbol{\nu}, \boldsymbol{\Sigma})}{\left(\int_I \exp(\sum_{k=1}^N \theta_{ik} e_k(x)) dx \right)^{m_i}} d\boldsymbol{\theta}_i. \quad (6)$$

For a detailed derivation, please refer to Appendix A.3. Maximizing this marginal likelihood can be seen as an empirical Bayes approach, where the prior for $\boldsymbol{\theta}_i$ is a multivariate normal distribution with mean $\boldsymbol{\nu}$ and covariance $\boldsymbol{\Sigma}$. Note that by p we denote a general density function, for example here $p(\boldsymbol{\theta}_i | \boldsymbol{\nu}, \boldsymbol{\Sigma})$ denotes the density of a multivariate normal distribution with parameters $\boldsymbol{\nu}$ and $\boldsymbol{\Sigma}$.

Due to the complicated nature of the likelihood function and the potential abundance of parameters in $\boldsymbol{\nu}$ and $\boldsymbol{\Sigma}$, numerical optimization of (6) is challenging. Therefore, in the following section, we use the Monte Carlo Expectation Maximization (MCEM) algorithm as a numerical method to effectively tackle this maximization problem.

2.3 Model estimation using an MCEM algorithm

The EM algorithm, developed by Dempster et al. (1977), addresses the challenge of maximum likelihood estimation in the presence of incomplete or missing data. This algorithm provides a framework for estimating the parameters of statistical models that involve unobserved or latent variables. It iteratively updates parameter estimates by incorporat-

ing both observed data and estimates of the missing data. In our specific context, we want to use the EM algorithm to handle latent, unobserved densities f_1, \dots, f_n , along with observed values $\mathbf{x}_1, \dots, \mathbf{x}_n$ sampled from these densities. Central to the EM algorithm is the notion of the expected complete-data log-likelihood, which in our case becomes

$$\begin{aligned} Q(\boldsymbol{\nu}, \boldsymbol{\Sigma} | \boldsymbol{\nu}^{(h)}, \boldsymbol{\Sigma}^{(h)}) &= \mathbb{E}(\log(p(\mathbf{x}_1, \dots, \mathbf{x}_n, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n | \boldsymbol{\nu}, \boldsymbol{\Sigma}))) \\ &= \mathbb{E}(\log(p(\mathbf{x}_1, \dots, \mathbf{x}_n | \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n))) + \mathbb{E}(\log(p(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n | \boldsymbol{\nu}, \boldsymbol{\Sigma}))) + \text{const.} \\ &= \sum_{i=1}^n \mathbb{E}(\log(p(\boldsymbol{\theta}_i | \boldsymbol{\nu}, \boldsymbol{\Sigma}))) + \text{const.} \end{aligned} \quad (7)$$

where the expectation is taken with respect to the conditional distribution $\boldsymbol{\theta}_i | \mathbf{x}_i, \boldsymbol{\nu}^{(h)}, \boldsymbol{\Sigma}^{(h)}$ of the parameters $\boldsymbol{\theta}_i = (\theta_{i1}, \dots, \theta_{iN})^T$ of the latent densities for all $i = 1, \dots, n$ given the current estimates $\boldsymbol{\nu}^{(h)}$ and $\boldsymbol{\Sigma}^{(h)}$ for $\boldsymbol{\nu}$ and $\boldsymbol{\Sigma}$. However, in the given setting, the conditional distribution needed to compute the expectation of the complete-data log-likelihood is not directly available, making the computation intractable. To address this challenge, Wei and Tanner (1990) introduce the Monte Carlo Expectation Maximization (MCEM) algorithm, which uses a Monte Carlo approach to approximate the expected value in Q . Thus, for our particular use case, we need to generate samples of $\boldsymbol{\theta}_i | \mathbf{x}_i, \boldsymbol{\nu}^{(h)}, \boldsymbol{\Sigma}^{(h)}$ for all $i = 1, \dots, n$. In the following, we outline the procedure for obtaining these samples and implementing the MCEM algorithm in our use case. Subsections 2.3.1 and 2.3.2 detail the E- and the M-steps, respectively, 2.3.3 the selection of model space and initial values, and 2.3.4 summarizes the complete algorithm.

2.3.1 E-step

For all $i = 1, \dots, n$ we approximate the conditional expectation $\mathbb{E}(\log(p(\boldsymbol{\theta}_i | \boldsymbol{\nu}, \boldsymbol{\Sigma})))$ where the expectation is taken with respect to $\boldsymbol{\theta}_i | \mathbf{x}_i, \boldsymbol{\nu}^{(h)}, \boldsymbol{\Sigma}^{(h)}$ using importance sampling, which is a method for estimating properties of a target distribution by sampling from another, auxiliary distribution. In our case we sample for all $i = 1, \dots, n$, r replicates of the parameters $\boldsymbol{\theta}_i$ of the latent densities from an auxiliary distribution with density $p_i^*(\boldsymbol{\theta}_i)$ instead of $\boldsymbol{\theta}_i | \mathbf{x}_i, \boldsymbol{\nu}^{(h)}, \boldsymbol{\Sigma}^{(h)}$. This yields replicates $\boldsymbol{\theta}_i^{(1)}, \dots, \boldsymbol{\theta}_i^{(r)}$, $r \in \mathbb{N}$, and we approximate

$$\mathbb{E}(\log(p(\boldsymbol{\theta}_i | \boldsymbol{\nu}, \boldsymbol{\Sigma}))) \approx \sum_{t=1}^r \frac{\omega_{it}}{\sum_{t=1}^r \omega_{it}} \log(p(\boldsymbol{\theta}_i^{(t)} | \boldsymbol{\nu}, \boldsymbol{\Sigma})) \quad (8)$$

with weights ω_{it} , $t = 1, \dots, r$ given as $\omega_{it} = \frac{p(\boldsymbol{\theta}_i^{(t)} | \mathbf{x}_i, \boldsymbol{\nu}^{(h)}, \boldsymbol{\Sigma}^{(h)})}{p_i^*(\boldsymbol{\theta}_i^{(t)})}$ for all $i = 1, \dots, n$. For details on this method and a comprehensive treatment of several related Monte Carlo methods, see the book by Hammersley and Handscomb (1964).

The key to this method lies in selecting an appropriate auxiliary distribution. To achieve this, we use the eigen decomposition of $\boldsymbol{\Sigma}^{(h)}$ with sorted eigenvalues $\sigma_1^2{}^{(h)} \geq \dots \geq \sigma_N^2{}^{(h)}$ and corresponding eigenvectors $\mathbf{v}_1^{(h)}, \dots, \mathbf{v}_N^{(h)}$. Then $\boldsymbol{\theta}_i \sim \mathcal{N}(\boldsymbol{\nu}^{(h)}, \boldsymbol{\Sigma}^{(h)})$ is equivalent to $\mathbf{z}_i = \mathbf{V}^{(h)}(\boldsymbol{\theta}_i - \boldsymbol{\nu}^{(h)}) \sim \mathcal{N}(\mathbf{0}, \text{diag}(\sigma_1^2{}^{(h)}, \dots, \sigma_N^2{}^{(h)}))$, where $\mathbf{V}^{(h)} = (\mathbf{v}_1^{(h)}, \dots, \mathbf{v}_N^{(h)})$ is a matrix whose columns are the eigenvectors of $\boldsymbol{\Sigma}^{(h)}$. Thus sampling from $\boldsymbol{\theta}_i | \mathbf{x}_i, \boldsymbol{\nu}^{(h)}, \boldsymbol{\Sigma}^{(h)}$ is equivalent to sampling from the conditional distribution of the scores $\mathbf{z}_i = (z_{i1}, \dots, z_{iN})$ given as

$$\begin{aligned}
 p(\mathbf{z}_i | \mathbf{x}_i, \boldsymbol{\nu}^{(h)}, \boldsymbol{\Sigma}^{(h)}) &\propto p(\mathbf{x}_i | \mathbf{z}_i, \boldsymbol{\nu}^{(h)}, \boldsymbol{\Sigma}^{(h)}) p(\mathbf{z}_i | \boldsymbol{\Sigma}^{(h)}) = p(\mathbf{x}_i | \boldsymbol{\theta}_i = \mathbf{V}^{(h)T} \mathbf{z}_i + \boldsymbol{\nu}^{(h)}) \prod_{k=1}^N p(z_{ik} | \sigma_k^2)^{(h)} \\
 &= \prod_{j=1}^{m_i} \text{clr}^{-1} \left(\sum_{k=1}^N \nu_k^{(h)} e_k + \mathbf{z}_i^T \mathbf{v}_k^{(h)} e_k \right) (x_{ij}) \prod_{k=1}^N p(z_{ik} | \sigma_k^2)^{(h)} \\
 &= \frac{\exp \left(\sum_{j=1}^{m_i} \left(\mu^{(h)}(x_{ij}) + \sum_{k=1}^N \mathbf{z}_i^T \mathbf{v}_k^{(h)} e_k(x_{ij}) \right) \right)}{\left(\int_I \exp \left(\mu^{(h)}(x) + \sum_{k=1}^N \mathbf{z}_i^T \mathbf{v}_k^{(h)} e_k(x) \right) dx \right)^{m_i}} \prod_{k=1}^N p(z_{ik} | \sigma_k^2)^{(h)} \tag{9}
 \end{aligned}$$

where $z_{ik} | \sigma_k^2 \sim \mathcal{N}(0, \sigma_k^2)$ for all $i = 1, \dots, n$ and $k = 1, \dots, N$. Here, $\mu^{(h)} = \sum_{k=1}^N \nu_k^{(h)} e_k$ is the current estimate for the mean function and $g_i = \mu^{(h)} + \sum_{k=1}^N \mathbf{z}_i^T \mathbf{v}_k^{(h)} e_k$, $i = 1, \dots, n$ are the current predictions for the latent clr transformed densities. Here we again take the Bayesian perspective, where $p(\mathbf{z}_i | \mathbf{x}_i, \boldsymbol{\nu}^{(h)}, \boldsymbol{\Sigma}^{(h)})$ is the posterior distribution for the prior $\mathcal{N}(\boldsymbol{\nu}^{(h)}, \boldsymbol{\Sigma}^{(h)})$. Note that this posterior is a proper distribution if the prior $\mathcal{N}(\boldsymbol{\nu}^{(h)}, \boldsymbol{\Sigma}^{(h)})$ is proper, that is if all eigenvalues of $\boldsymbol{\Sigma}^{(h)}$ are finite.

Lemma 2.5. *Let $\sigma_1^2 < \infty$. Then $\int_{\mathbb{R}^N} p(\mathbf{z}_i | \mathbf{x}_i, \boldsymbol{\nu}^{(h)}, \boldsymbol{\Sigma}^{(h)}) d\mathbf{z}_i < \infty$ for all i .*

A proof for this statement can be found in Appendix A.4. This also implies that $p(\mathbf{z}_i | \mathbf{x}_i, \boldsymbol{\nu}^{(h)}, \boldsymbol{\Sigma}^{(h)})$ is decreasing as $\|\mathbf{z}_i\| \rightarrow \infty$ and since it is also continuous, it attains its mode $\mathbf{z}_i^* \in \mathbb{R}^n$. This is not necessarily the case if $\mathcal{N}(\boldsymbol{\nu}^{(h)}, \boldsymbol{\Sigma}^{(h)})$ is improper (see Appendix A.5 for a counterexample).

Since we assume here that the prior distribution $\mathcal{N}(\boldsymbol{\nu}^{(h)}, \boldsymbol{\Sigma}^{(h)})$ is proper, the mode of the posterior distribution $\mathbf{z}_i^* = \operatorname{argmax}_{\mathbf{z}_i \in \mathbb{R}^N} p(\mathbf{z}_i | \mathbf{x}_i, \boldsymbol{\nu}^{(h)}, \boldsymbol{\Sigma}^{(h)})$ is attained. Hence we can choose a multivariate normal distribution centered around the mode as an auxiliary distribution for the scores \mathbf{z}_i . We further choose the variances to be proportional to the prior variances $\sigma_1^2, \dots, \sigma_N^2$. This means that for a tuning parameter $\lambda > 0$, we choose the auxiliary distribution $p_i^*(\mathbf{z}_i)$ to be $\mathcal{N}(\mathbf{z}_i^*, \lambda \operatorname{diag}(\sigma_1^2, \dots, \sigma_N^2))$. Usually we set $\lambda = 1$, but if one wants to explore the parameter space for \mathbf{z} more or less, one can also set λ larger or smaller. Consequently, once we compute the mode \mathbf{z}_i^* , sampling from the auxiliary distribution p_i^* reduces to independently sampling each element of the vector \mathbf{z}_i from a univariate normal distribution. To numerically compute the mode, i.e., the maximizer of (9) with respect to \mathbf{z}_i , it is useful to derive the gradient of its log transformation to apply a gradient descent algorithm.

Lemma 2.6. *The gradient of the log conditional density of the scores $\mathbb{R}^N \rightarrow \mathbb{R}$, $\mathbf{z}_i \mapsto \log(p(\mathbf{z}_i | \mathbf{x}_i, \boldsymbol{\nu}^{(h)}, \boldsymbol{\Sigma}^{(h)}))$ is given as*

$$\nabla \log(p(\mathbf{z}_i | \mathbf{x}_i, \boldsymbol{\nu}^{(h)}, \boldsymbol{\Sigma}^{(h)})) = \sum_{k=1}^N \mathbf{v}_k^{(h)} \left(\sum_{j=1}^{m_i} e_k(x_{ij}) - m_i \langle \mathbf{f}_{\mathbf{z}_i}, e_k \rangle_{\mathbb{L}_2} \right) - \left(\frac{z_{il}}{\sigma_l^2} \right)_{l=1, \dots, N}$$

where $\mathbf{f}_{\mathbf{z}_i} = \text{clr}^{-1} \left(\mu^{(h)} + \sum_{k=1}^N \mathbf{z}_i^T \mathbf{v}_k^{(h)} e_k \right)$ for all $\mathbf{z}_i = (z_{i1}, \dots, z_{iN})^T \in \mathbb{R}^N$.

A detailed derivation can be found in Appendix A.6. With this readily available gradient, finding the mode becomes numerically feasible and we can obtain i.i.d. samples for the scores \mathbf{z}_{it} and corresponding weights $\omega_{it} \in \mathbb{N}$ for all $t = 1, \dots, r$ using the importance sampling described above.

The equivalence of conditionally sampling $\boldsymbol{\theta}_i$ or \mathbf{z}_i also yields samples $\boldsymbol{\theta}_i^{(t)}$ from $\boldsymbol{\theta}_i | \mathbf{x}_i, \boldsymbol{\nu}^{(h)}, \boldsymbol{\Sigma}^{(h)}$ for all $t = 1, \dots, r$ via $\boldsymbol{\theta}_i^{(t)} = \boldsymbol{\nu}^{(h)} + \mathbf{V}^{(h)T} \mathbf{z}_{it}$. Hence, we approximate the expected complete-data log-likelihood given in (7) by

$$Q(\boldsymbol{\nu}, \boldsymbol{\Sigma} | \boldsymbol{\nu}^{(h)}, \boldsymbol{\Sigma}^{(h)}) \approx \sum_{i=1}^n \sum_{t=1}^r \frac{\omega_{it}}{\sum_{t=1}^r \omega_{it}} \log(p(\boldsymbol{\theta}_i^{(t)} | \boldsymbol{\nu}, \boldsymbol{\Sigma})) + \text{const.} \tag{10}$$

using the Monte-Carlo approximation given in (8) for the conditional expectation.

2.3.2 M-step

The M-step updates the parameters ν and Σ by maximizing this approximation (10) of the Q -function. The new estimates are then given by

$$\left(\nu^{(h+1)}, \Sigma^{(h+1)}\right) = \operatorname{argmax}_{\nu, \Sigma} Q(\nu, \Sigma | \nu^{(h)}, \Sigma^{(h)}) \approx \operatorname{argmax}_{\nu, \Sigma} \sum_{i=1}^n \sum_{t=1}^r \frac{\omega_{it}}{\sum_{t=1}^r \omega_{it}} \log(p(\theta_i^{(t)} | \nu, \Sigma)).$$

Since for the latent density model (5) we assume that $\theta_i^{(t)} | \nu, \Sigma$ follows a multivariate normal distribution with mean ν and covariance matrix Σ , this optimization problem corresponds to a weighted maximum likelihood estimation of the mean and the covariance matrix. Remarkably, this maximization problem also arises when the EM algorithm is used to estimate a Gaussian mixture distribution, allowing us to derive the solution based on the computations performed for this problem (e.g. Bishop and Nasrabadi, 2006). Hence, we compute the updates for the parameters of our model as

$$\begin{aligned} \nu^{(h+1)} &= \frac{1}{\sum_{i=1}^n \sum_{t=1}^r \omega_{it}} \sum_{i=1}^n \sum_{t=1}^r \omega_{it} \theta_i^{(t)} \\ \Sigma^{(h+1)} &= \frac{1}{\sum_{i=1}^n \sum_{t=1}^r \omega_{it}} \sum_{i=1}^n \sum_{t=1}^r \omega_{it} (\theta_i^{(t)} - \nu^{(h+1)}) (\theta_i^{(t)} - \nu^{(h+1)})^T. \end{aligned}$$

These are the weighted mean and weighted covariance matrix of the samples of the principal component scores $\theta_i^{(t)}$, $i = 1, \dots, n$, $t = 1, \dots, r$.

2.3.3 Selection of model space and initial values

In order to apply our method to real-world problems, we first need to find a suitable model space as well as suitable initial values for the MCEM algorithm. We suggest using piecewise constant spline functions for modeling. Since the piecewise constant functions are dense in \mathbb{L}^2 , they allow us to approximate any function in \mathbb{L}_0^2 , and thus any density in \mathcal{B} , with arbitrary accuracy if the nodes are chosen to be on a fine grid. Therefore, we fix a fine grid for the knots $\kappa_1, \dots, \kappa_{N+1}$ and choose as a basis the indicator functions which are one between two neighboring knots and zero elsewhere. That is $e_k = \mathbb{1}_{[\kappa_k, \kappa_{k+1}]}$ for all $k = 1, \dots, N$.

To obtain suitable initial values for $\nu^{(0)}$ and $\Sigma^{(0)}$, we propose to first estimate the latent densities $\hat{f}_1, \dots, \hat{f}_n$ by kernel density estimation. We then develop their clr transformations $\hat{g}_1, \dots, \hat{g}_n$ in our basis e_1, \dots, e_N . Subsequently, we estimate $\nu^{(0)}$ and $\Sigma^{(0)}$ as the empirical mean and covariance of the coefficients $\theta_1, \dots, \theta_n$ of $\hat{g}_1, \dots, \hat{g}_n$, respectively. This approach effectively restricts the model space to the span of the kernel density estimates. This could be a problem, for example, if only a small sample of densities is available, or if the kernel density estimates are close to zero in some parts of the support.

In this case, an alternative would be to initially select a lower dimensional, smooth spline space for modeling. If we choose an orthonormal basis $e_1, \dots, e_N \in \mathbb{L}_0^2$, such as normalized versions of the orthogonal compositional splines suggested by Machalová et al. (2021), we can choose arbitrary values for the initial mean and covariance of the coefficients, for instance, $\nu^{(0)} = \mathbf{0} \in \mathbb{R}^N$ and $\Sigma^{(0)} = \mathbb{I}_N \in \mathbb{R}^{N \times N}$, the identity matrix.

On the other hand, when the number of densities n is large, which results also in a $N = n$ basis functions by the procedure given above, not only does computing the mode become a high-dimensional optimization problem, which is computationally demanding, also calculating the weights ω_{it} becomes unstable as in this case typically many of the variances $\sigma_k^{2(h)}$ will be very small, thus the product $\prod_{k=1}^N p(z_k | \sigma_k^{2(h)})$ will be close to zero. In this case, we suggest reducing the dimensionality of $GP(\hat{\mu}^{(h)}, \hat{K}^{(h)})$ in each step h by setting the variances $\sigma_k^{2(h)} = 0$ for $k > N'$, where

N' is a chosen value such that $N' \ll N$. The specific value of N' can be determined based on the desired proportion of variance explained, while the proportion of variance explained should be considerably greater than desired for the final PCA.

2.3.4 Estimation of the density PCA using the MCEM algorithm

Subsections 2.3.1-2.3.3 derived all necessary ingredients to now use the MCEM algorithm to estimate the PCA. First, we initialize $\boldsymbol{\nu}^{(0)}$ and $\boldsymbol{\Sigma}^{(0)}$ according to 2.3.3. Then we iterate the following two steps until the convergence criteria $\|\boldsymbol{\nu}^{(h+1)} - \boldsymbol{\nu}^{(h)}\|_{\mathbb{L}^2} < \epsilon$ and $\|\boldsymbol{\Sigma}^{(h+1)} - \boldsymbol{\Sigma}^{(h)}\|_{\mathbb{L}^2} < \epsilon$ for a threshold $\epsilon > 0$ are reached.

- E-step following 2.3.1,

- M-step following 2.3.2.

Estimates at convergence then approximate the maximum likelihood estimates of the observed data likelihood, that is $\hat{\boldsymbol{\nu}} \approx \boldsymbol{\nu}^{(h+1)}$ and $\hat{\boldsymbol{\Sigma}} \approx \boldsymbol{\Sigma}^{(h+1)}$. Finally, the equivalence given in Corollary 2.4 yields estimates for μ , φ_k and σ_k^2 in (2) using the eigendecomposition of $\hat{\boldsymbol{\Sigma}}$, where $\hat{\boldsymbol{v}}_1, \dots, \hat{\boldsymbol{v}}_N$ are the eigenvectors of $\boldsymbol{\Sigma}$ with corresponding eigenvalues $\hat{\sigma}_1^2, \dots, \hat{\sigma}_N^2$. Then, the estimate for the mean is obtained as $\hat{\boldsymbol{\mu}} = \sum_{k=1}^N \hat{\nu}_k e_k$, where $\hat{\boldsymbol{\nu}} = (\hat{\nu}_1, \dots, \hat{\nu}_N)^T$. The estimates for the eigenfunctions are given as $\hat{\varphi}_l = \sum_{k=1}^N \hat{v}_{lk} e_k$, where $\hat{\boldsymbol{v}}_l = (\hat{v}_{l1}, \dots, \hat{v}_{lN})$ for all $l = 1, \dots, N$ with corresponding eigenvalues $\hat{\sigma}_1^2, \dots, \hat{\sigma}_N^2$.

The scores Z_{ik} , $k = 1, \dots, N$ for each density f_i , $i = 1, \dots, n$ are then predicted as the posterior mode given the estimates $\hat{\boldsymbol{\nu}}$ for the mean and $\hat{\boldsymbol{\Sigma}}$ for the covariance of the coefficients, that is $\hat{\boldsymbol{z}}_i = \operatorname{argmax}_{\boldsymbol{z}_i \in \mathbb{R}^N} p(\boldsymbol{z}_i | \boldsymbol{x}_i, \hat{\boldsymbol{\nu}}, \hat{\boldsymbol{\Sigma}})$ with $\hat{\boldsymbol{z}}_i = (\hat{z}_{i1}, \dots, \hat{z}_{iN})$. This also yields predictions for the latent densities as $\hat{f}_i = \operatorname{clr}^{-1}(\hat{g}_i)$ with $\hat{g}_i = \hat{\boldsymbol{\mu}} + \sum_{k=1}^N \hat{z}_{ik} \hat{\varphi}_k$ for all $i = 1, \dots, n$.

2.4 PCA for compositional data as a special case

Although we have limited our considerations so far to densities with respect to the Lebesgue measure, it's important to recognize that our methodology can be seamlessly extended to densities with respect to arbitrary measures. In the following, we illustrate this via showing how our approach can be used to derive Principal Component Analysis (PCA) for compositional data, namely densities with respect to the discrete measure. Notably, our method has the advantage of being applicable to "sparsely observed" compositional data, known as count compositions in the compositional data literature (Filzmoser et al., 2018), even when some of the categories have no observations and without imputations.

The discrete measure on the power set $\mathcal{P}(\{A_1, \dots, A_k\})$ of a finite set of disjoint outcomes A_1, \dots, A_k is given as $\eta = \sum_{k=1}^N \delta_{A_k}$, where $\delta_{A_k}(B) = 1$, if $B = A_k$ and $\delta_{A_k}(B) = 0$ else, for all $B \in \mathcal{P}(\{A_1, \dots, A_k\})$. Hence, every probability measure on $\mathcal{P}(\{A_1, \dots, A_k\})$ is given by a discrete density, i.e. probability mass function $f : \{A_1, \dots, A_k\} \rightarrow \mathbb{R}$ with respect to η by the Radon-Nikodym Theorem. Here, the density f is characterized solely by the values $\pi_k = f(A_k)$ for all $k = 1, \dots, N$, and since we consider only probability measures, it must hold that $\sum_{k=1}^N \pi_k = 1$. That means \mathcal{B} , the set of densities with respect to the discrete measure on $\mathcal{P}(\{A_1, \dots, A_k\})$ can be identified with the simplex $\{\boldsymbol{\pi} \in \mathbb{R}^N | \sum_{k=1}^N \pi_k = 1, \pi_k \geq 0 \forall k = 1, \dots, N\}$ and via the discrete centered log-ratio transformation $\boldsymbol{\rho} = \operatorname{clr}(\boldsymbol{\pi}) = (\log(\pi_1) - \frac{1}{N} \sum_{k=1}^N \log(\pi_k), \dots, \log(\pi_N) - \frac{1}{N} \sum_{k=1}^N \log(\pi_k))$ with the $N - 1$ dimensional Hilbert space $\mathcal{H} = \mathbb{R}_0^N = \{\boldsymbol{\rho} \in \mathbb{R}^N | \sum_{k=1}^N \rho_k = 0\}$ (Aitchison, 1982). If we equip this space with the standard scalar product on \mathbb{R}^N , we obtain the Aitchison geometry on \mathcal{B} , which defines a Hilbert space structure for compositional data, i.e. for densities with respect to the discrete measure.

In order to use our method to perform principal component analysis for observed count compositions $\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_n$, we need to find a suitable basis representation of \mathcal{H} , as well as appropriate initial values for the mean and covariance of the corresponding basis coefficients. For densities with respect to the discrete measure, kernel density estimation cannot be employed due to the usual lack of an order relation and thus neighborhood structure to smooth over. We therefore rely on the alternative described in Subsubsection 2.3.3 and use an orthonormal basis of $\mathcal{H} = \mathbb{R}_0^N$.

There are many options since an orthonormal basis can be obtained from any basis of \mathbb{R}_0^N using Gram-Schmidt orthogonalization. For example, Egozcue et al. (2003) suggest to use the following such basis

$$e_k = \sqrt{\frac{k}{k+1}} (\overbrace{k^{-1}, \dots, k^{-1}}^{k \text{ times}}, -1, 0, \dots, 0)^T$$

with $k = 1, \dots, N$ as an orthonormal basis of \mathbb{R}_0^N , which yields an orthonormal basis of \mathcal{B} via the inverse clr transform. With this basis choice we propose to set $\boldsymbol{\nu}^{(0)} = \mathbf{0} \in \mathbb{R}^{N-1}$ and $\boldsymbol{\Sigma}^{(0)} = \mathbb{I}_N \in \mathbb{R}^{(N-1) \times (N-1)}$ as initial values for the mean and covariance of the basis coefficients and proceed with the MCEM algorithm as in the continuous case described above.

3 Applications

In this section, we demonstrate the applicability and advantages of our latent density-based PCA. We consider two different applications. In the first application, we analyze densities describing distributions of summer daily maximum temperatures in Berlin. In the second application, the latent densities describe the distribution of rental prices for each district in Munich.

3.1 Distributions of daily maximum temperature in summer months per year

According to the Copernicus Climate Change Service (C3S) report (<https://climate.copernicus.eu/esotc/2022/temperature>), 2022 was the second warmest year on record in Europe, with temperatures 0.9°C above the long-term average. In particular, the summer of that year set a new mark as the hottest on record, with temperatures 1.4°C above average, topping the previous warmest summer in 2021 by 0.3°C.

These findings are consistent with the observed trend of increasing temperatures, an indicator of climate change. To refine the analysis of summer temperature trends, we focus not on average temperatures, but on trends in the entire distribution of daily maximum temperatures during the summer months of June, July, and August each year. That is, we consider as observational units densities that describe the distributions of daily maximum temperatures in summer. Thus, for every year, we treat the daily maximum temperature measured for the 92 days in June, July, and August as observations from these densities per year. The daily maximum temperature data we use in this application have been collected from 1951 to 2022 at a single weather observatory, which is Berlin Tempelhof. Data and metadata are available at <https://www.ecad.eu>, provided by the ECA&D project (Klein Tank et al., 2002).

To visualize and describe how the distribution of daily maximum temperature has changed over the period from 1951 to 2022, we proceed as follows. First, we obtain a low-dimensional representation of the latent daily maximum temperature densities using our latent density PCA and with kernel density estimates as initial estimates (Figure 6 in the appendix) For technical details of the estimation please refer to Appendix B.1). Then, in Figure 1, we visualize the first four principal components on clr level (top row), and transformed back to density level (middle row). In the bottom row, we plot the temporal trend of the corresponding predicted scores, overlaying a scatterplot smoother and pointwise confidence bands based on Wood (2017).

This shows that adding a multiple of the first principal component, which explains 40% of the total variability in latent densities, to the estimated mean mainly causes a rightward shift of the density. In particular, a positive value of the first principal component implies that high temperatures are more likely than in average years. Looking at the effect on clr level, we notice that this is especially true for temperatures above 35°C, which are more likely to occur in years with high first principal component scores. Notably, these first principal component scores show a clear increase over the time course from 1951 to 2022 (Figure 1, bottom left), meaning that hot and also very hot daily maximal temperatures in summer became more likely.

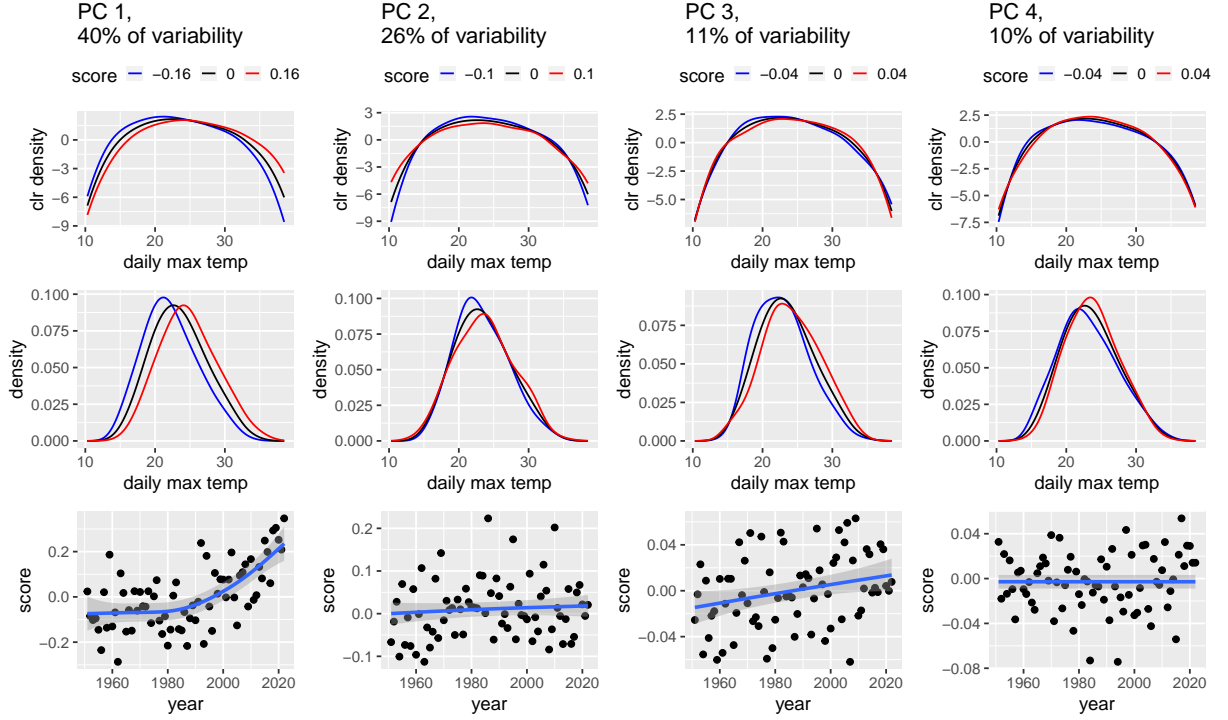


Figure 1: Latent density PCA for daily maximum temperature ($^{\circ}\text{C}$). Top: Effect of adding/subtracting $\hat{\sigma}_k \hat{\varphi}_k$ to the clr transformed mean density μ , where $\hat{\varphi}_k$ is the k th principal component, with corresponding eigenvalue $\hat{\sigma}_k^2$, $k = 1, 2, 3, 4$. Middle: Effect on the density level, i.e. clr^{-1} transformations of the functions in the top row. Bottom: Temporal trend of the corresponding predicted scores per year, with scatterplot smoother and pointwise confidence bands overlaid.

In contrast, the scores associated with the second and fourth principal components show no or almost no visible temporal trend. Adding these principal components to the estimated mean results in smaller changes in the shape of the density, with subtler shifts to the right in certain areas of the density. However, adding the third principal component to the estimated mean shifts the density towards experiencing moderate to hot temperatures (25°C - 35°C) more frequently, while decreasing the likelihood of milder temperatures (15°C - 25°C). The scores associated with this third principal component also show an increase over time.

This means that the trends of both the first and third principal component scores indicate that hot and very hot days are becoming more likely. When we plot both scores together (Figure 2) we see that all early years, corresponding to (dark) blue points, tend to lie in the bottom left corner, while recent years (yellow and orange points) are predominantly in the top right corner. Thus, recent years have high first principal component scores and/or high third principal component scores, which means the likelihood for hot and/or very hot days has been higher in these years than in earlier years.

This application shows that the estimation of the latent density model (5) is suitable for identifying a small number of principal directions of variation in the data for densities when only discrete observations of each density are available. These principal components can then be used to visualize the data and/or for further analysis, such as to relate them to other scalar variables (in our case, the year of observation).

However, in this homogeneous and only mildly sparse setting with 92 observations per density, differences to the simpler two-step approach with pre-smoothing (our starting values), a PCA obtained from the clr transformations of the kernel density estimates (see Figure 7), are relatively small. One notable difference, though, is that in the latent density model, the trend of the score associated with the first principal component changes more rapidly around 1990, with a

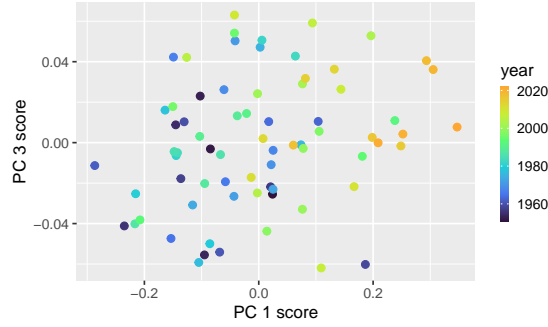


Figure 2: Scores associated with the first and third principal components for all years.

steeper increase since that time, than that of the first principal component scores obtained in the two-step approach. The trend in the third PC score is somewhat more pronounced as well, which may be due to unstable kernel density estimates in areas where the densities are close to zero. In contrast to this first application, our second application has more heterogeneous sample sizes and we will observe how this unbalanced setting affects the estimation.

3.2 Distributions of rental prices in the districts of Munich

We consider data that was collected in Munich in 2019 (the most recent year) to construct an official rent index by estimating the rent given certain covariates of the apartments in a regression model. For details on the data collection, the available variables, estimation and interpretation of the rent index, see (Windmann and Kauermann, 2019, in German). A sub-dataset, which contains data on all 3255 apartments but only a subset of the covariates used for the official Munich rent index, is provided in the supplemental material of Fahrmeir et al. (2023) and is also used here.

The goal of the analysis here is to describe the differences in the distributions of rental prices across the districts of Munich. To do this, we model the distribution of net rents per square meter, assuming a latent density for each district. Figure 8 in Appendix B.2 displays the histogram estimates and kernel density estimates (Gaussian kernel, bandwidth = 2) for the densities in each district. Table 1 shows that the number of observations m_i , $i = 1, \dots, 25$ per district varies considerably in this dataset. It ranges from 29 observations in district 23-Allach-Untermenzing to 261 observations in district 9-Neuhausen-Nymphenburg. Below, we will compare the estimation of our latent density model (5) for this heterogeneous sampling scheme to a two-step approach, where the densities are first estimated and then PCA is performed.

district i	name	m_i	district i	name	m_i
1	Altstadt-Lehel	79	14	Berg am Laim	75
2	Ludwigsvorstadt-Isarvorstadt	217	15	Trudering-Riem	92
3	Maxvorstadt	219	16	Ramersdorf-Perlach	111
4	Schwabing-West	241	17	Obergiesing-Fasangarten	125
5	Au-Haidhausen	230	18	Untergiesing-Harlaching	151
6	Sendling	136	19	Thalkirchen-Obersendling- Forstenried-Fürstenried-Solln	160
7	Sendling-Westpark	95	20	Hadern	64
8	Schwanthalerhöhe	110	21	Pasing-Obermenzing	107
9	Neuhausen-Nymphenburg	261	22	Aubing-Lochhausen-Langwied	35
10	Moosach	81	23	Allach-Untermenzing	29
11	Milbertshofen-Am Hart	101	24	Feldmoching-Hasenbergl	38
12	Schwabing-Freimann	198	25	Laim	135
13	Bogenhausen	165			

Table 1: Munich districts: number i , name and number of observations per district m_i

We first consider in Figure 3a the results of estimating the latent density model with kernel density estimates as initial estimates. Details for the estimation can be found in Appendix B.2. In this model, the first principal component causes a shift towards more expensive apartments, and looking at the effect on clr level, we see that this principal component also primarily describes whether the occurrence of very expensive apartments (more than 25€ per square meter) is likely. Looking at which districts have a positive score for the first principal component, we see that these districts are mainly in the city center and correspond closely to the districts that have an increasing influence on the expected net rent in the regression model estimated in Fahrmeir et al. (2023).

The second principal component describes the presence of very cheap and expensive apartments. However, this principal component has only a small influence on the shape of the rent per square meter density. In the latent density model the variance of the principal components, and thus their importance, is measured at the clr level. Since in this case the largest effect is on parts of the clr transformed densities with negative values, the multiplicative effect on the actual densities is in parts where the density is close to zero and thus the effect, corresponding to heavier tails, is hardly visible. It is therefore not surprising that the spatial distribution of the corresponding scores has little structure, since the second principal component describes the occurrence of a few extreme observations in the districts.

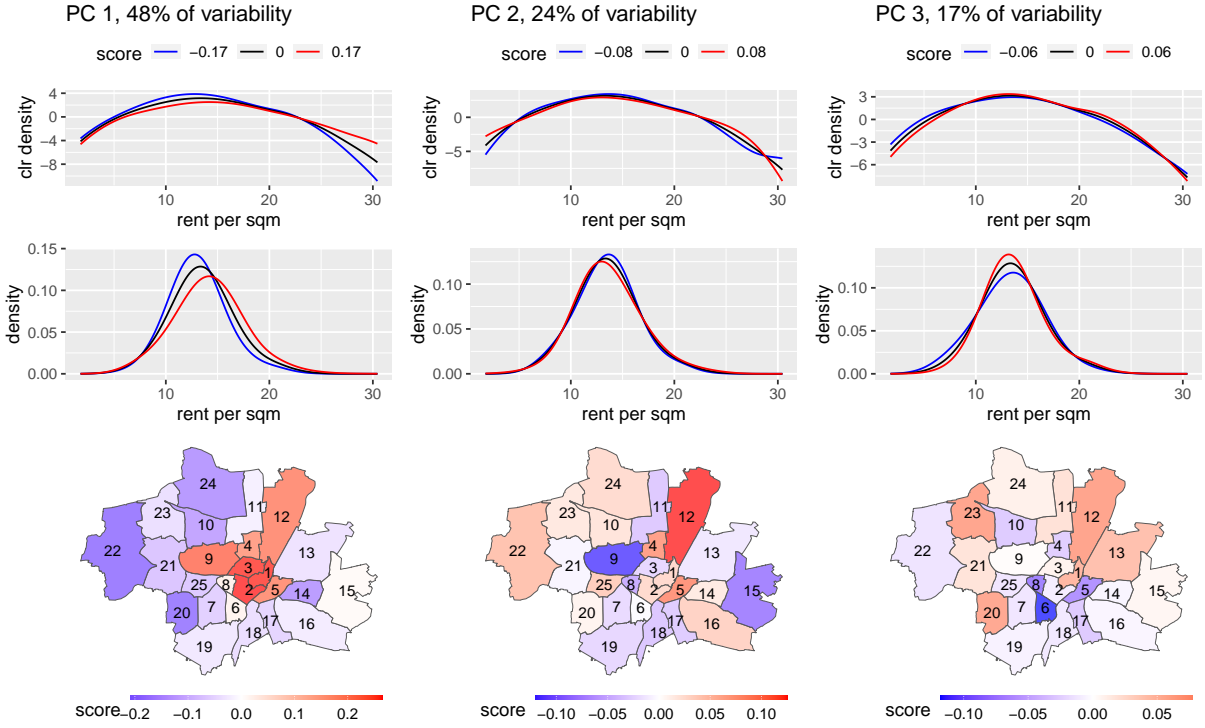
The effect of the third principal component on the densities appears to be larger than the effect of the second principal component, although the variability explained is smaller, as it affects areas around the mode of the distribution. For this component, a negative score mostly describes a larger share of affordable housing (5€ to 10€ per square meter). These negative scores are mainly predicted for districts in the south of Munich, i.e. 6-Sendling and neighboring districts.

Comparing the effects of the first three principal components estimated with the latent density model with the estimates obtained using a two-step approach based on pre-smoothing with kernel density estimates (Gaussian kernel, bandwidth = 2), we see that the estimates differ considerably. In particular, for the two-step approach, the main variation in the scores is caused by density estimates of districts with only a small number of observations m_i . That is, the most extreme scores are estimated for district 22 with $m_{22} = 35$, district 23 with $m_{23} = 29$, and district 24 with $m_{24} = 38$, for the first three principal components, respectively. These districts benefit from the shrinkage effect of the latent density model, which divides the total variance into a part which is due to the underlying stochastic process for the latent densities and a part due to sampling from them. This shrinkage causes the predictions for the densities in these districts to be closer to the overall mean. Note also that the percentages of variance explained for the PCs shown in figures 3a and 3b, respectively, are correspondingly relative to different total variances.

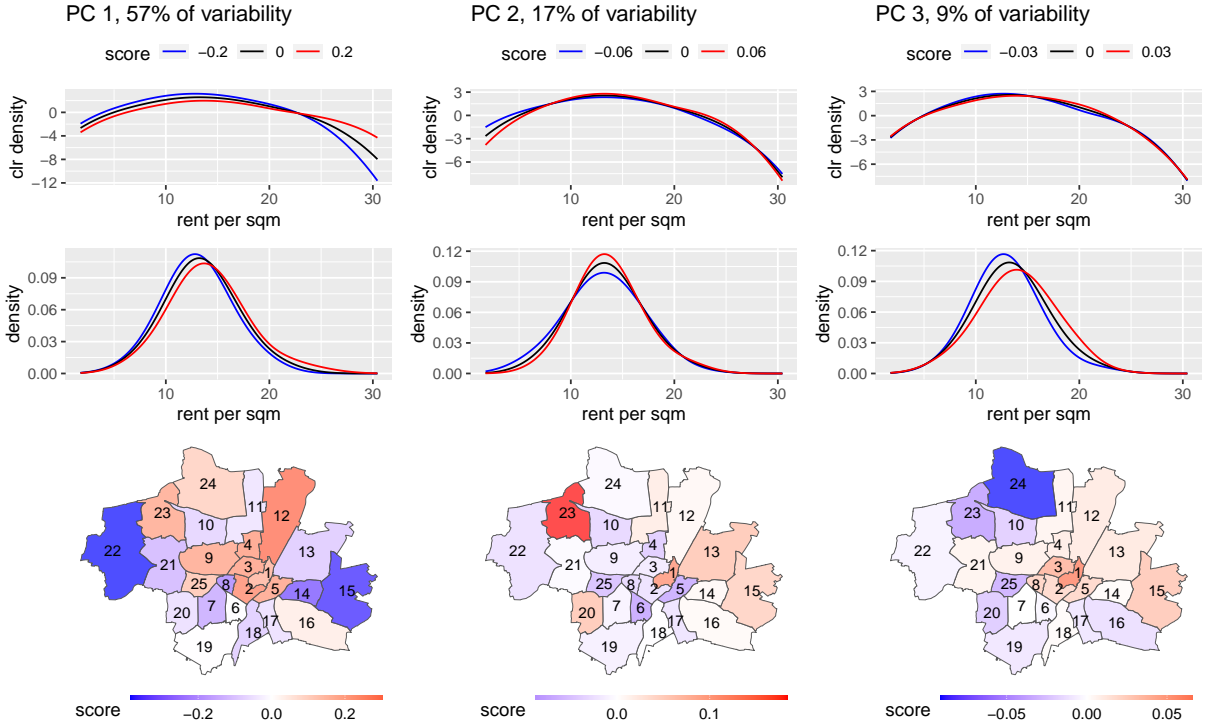
Figure 8 in the appendix shows that for these densities, the latent density predictions appear more plausible than the kernel density estimates, especially in areas where there are few observations, i.e. for very cheap or expensive housing. However, nearly all predicted densities appear to better reflect the underlying data using our approach, as the kernel density estimates generally underestimate the modes. The reason for this behavior is that we had to choose a relatively large bandwidth for the kernels to avoid estimating close to zero densities in other parts of the domain, especially for small samples. While we use the same kernel density estimates for our approach as starting values, the influence of such problems and the choice of the bandwidth in general is much smaller due to the later updates of the model. In order to systematically investigate the influence of the number of observations per density on the model estimation, a simulation is carried out in the following subsection.

4 Simulation

This subsection aims to evaluate how well our latent density model can recover the mean and covariance structure of the latent process for a varying number of observations per density, ranging from very few observations to a moderate number. We also compare the performance of the model with two-step approaches, where the density estimates are obtained first, and then the PCA is performed after applying the clr transformation to each density. This corresponds to the simplicial PCA proposed by Hron et al. (2016). For the first comparison, we obtain kernel density estimates with a Gaussian kernel and then perform PCA on the clr transformed densities, which is also used as the initial estimate



(a) PCA based on the latent density model.



(b) Two-step approach using kernel density estimates and then applying PCA after clr transformation.

Figure 3: Comparison of the latent density model with a two-step approach using kernel density estimates as preprocessing on the Munich rent dataset. For each method, the first row shows the effect of adding a principal component times the corresponding standard deviation to the mean on clr level, the second row shows the same effect on density level, and the third row shows a map of Munich districts, where the color represents the predicted scores.

for our method. Second, we use a compositional spline estimate for the clr transformed densities, as suggested by Machalová et al. (2021), before performing the PCA.

To this end we choose the following simulation setting. In each simulation run we simulate $n = 30$ densities on the interval $I = [0, 1]$ from a Gaussian process with true clr transformed mean function $\mu(x) = -20(x - \frac{1}{2})^2 + \frac{5}{3}$ and only two principal components. These are given on clr level as $g_1(x) = \frac{1}{5} \sin(10(x - \frac{1}{2}))$ with corresponding factor $Z_1 \sim \mathcal{N}(0, 0.5)$ and $g_2(x) = \frac{1}{10} \cos(2\pi(x - \frac{1}{2}))$ with corresponding factor $Z_2 \sim \mathcal{N}(0, 0.2)$. Note that these functions satisfy $\mu, g_1, g_2 \in \mathbb{L}_0^2$ and $g_1 \perp g_2$. The samples for the densities $f_i, i = 1, \dots, 30$ are then obtained as $f_i = \text{clr}^{-1}(\mu + z_{i1}g_1 + z_{i2}g_2)$ where $z_{i1} \stackrel{i.i.d.}{\sim} Z_1$ and $z_{i2} \stackrel{i.i.d.}{\sim} Z_2$. The resulting densities are shown in Figure 5 in the top row on the left for a simulation run with $m_i = 40$. Finally, we sample observations $x_{ij} \stackrel{i.i.d.}{\sim} f_i$ with $j = 1, \dots, m_i$ from each density $f_i, i = 1, \dots, n$. For the number of observations per density we consider $m_i \in \{20, 40, 80, 160\}$ and repeat the simulation 100 times for each m_i .

For the two-step approaches we then estimate the densities, either with kernel density estimates using Gaussian kernels with bandwidths 0.12, 0.09, 0.08, 0.07 for the different setting with $m_i = 20, 40, 80, 160$, respectively, or using cubic compositional splines with five knots. The kernel density estimates are also used as initial estimates for our latent density model. The number of Monte Carlo samples in the E step 2.3.1 is chosen to be $r = 10h$, where h is the iteration index, i.e., the number of samples increases over the iterations. The parameter λ for the proposal density is set to 1. In each iteration, the dimension reduction is at most 0.0001, i.e. we keep as many principal components as necessary to explain at least 99.999 % of the variance.

The performance of the different methods is evaluated as the distance to the oracle estimates, that are the pointwise estimates of the mean and the covariance function based on the true underlying densities f_1, \dots, f_n (Figure 4) evaluated on a equidistant grid with 200 grid points. More precisely, we compute the distance of the mean estimates as $\sqrt{\int_0^1 (\tilde{\mu}(x) - \hat{\mu}(x))^2 dx}$, where $\tilde{\mu}$ is the oracle estimate for the mean and $\hat{\mu}$ is the estimate of each method. Analogously, the distance of the covariance functions is obtained as $\sqrt{\int_0^1 \int_0^1 (\tilde{C}(x_1, x_2) - \hat{C}(x_1, x_2))^2 dx_1 dx_2}$, where \tilde{C} is the oracle estimate for the covariance and \hat{C} is the estimate of each method.

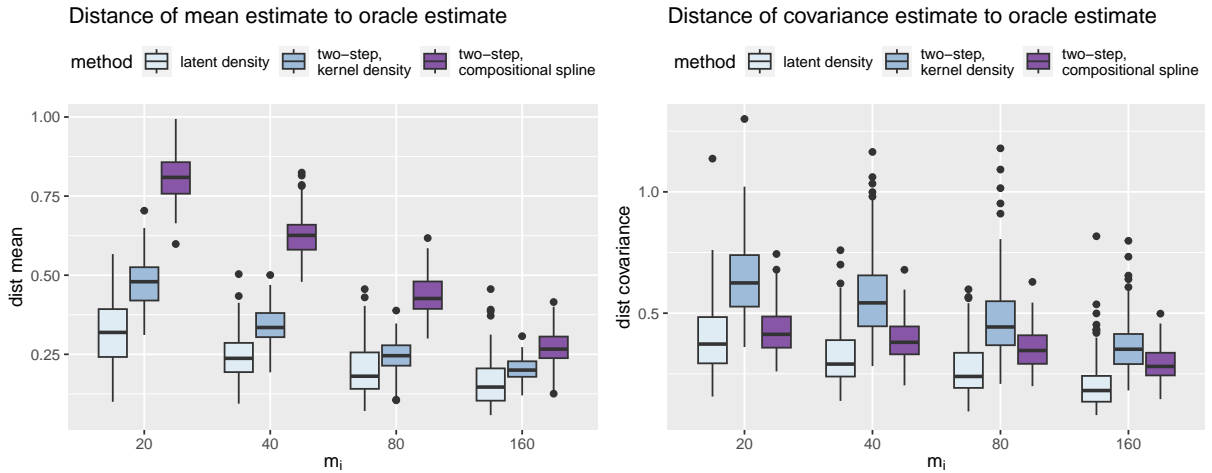


Figure 4: \mathbb{L}^2 distance of the estimated mean and covariance functions for each method to the oracle estimates based on the true underlying densities f_1, \dots, f_n .

As expected, the performance of all three methods improves as the number of observations per density increases. Still, over all values of $m_i \in \{20, 40, 80, 160\}$, and for both the mean and the covariance function, our latent density model has the smallest average distance to the oracle estimate over the 100 replicates. This shows that our method outperforms both two-step approaches in this scenario. However, when comparing the two-step approaches, it is worth

noting that kernel density estimates seem to be better for estimating the mean, while compositional splines excel at capturing the covariance structure. For a more concrete picture of how the mean and covariance estimates for the three methods behave relative to the oracle estimate, Figure 5 shows each for an example with $m_i = 40$ observations per density.

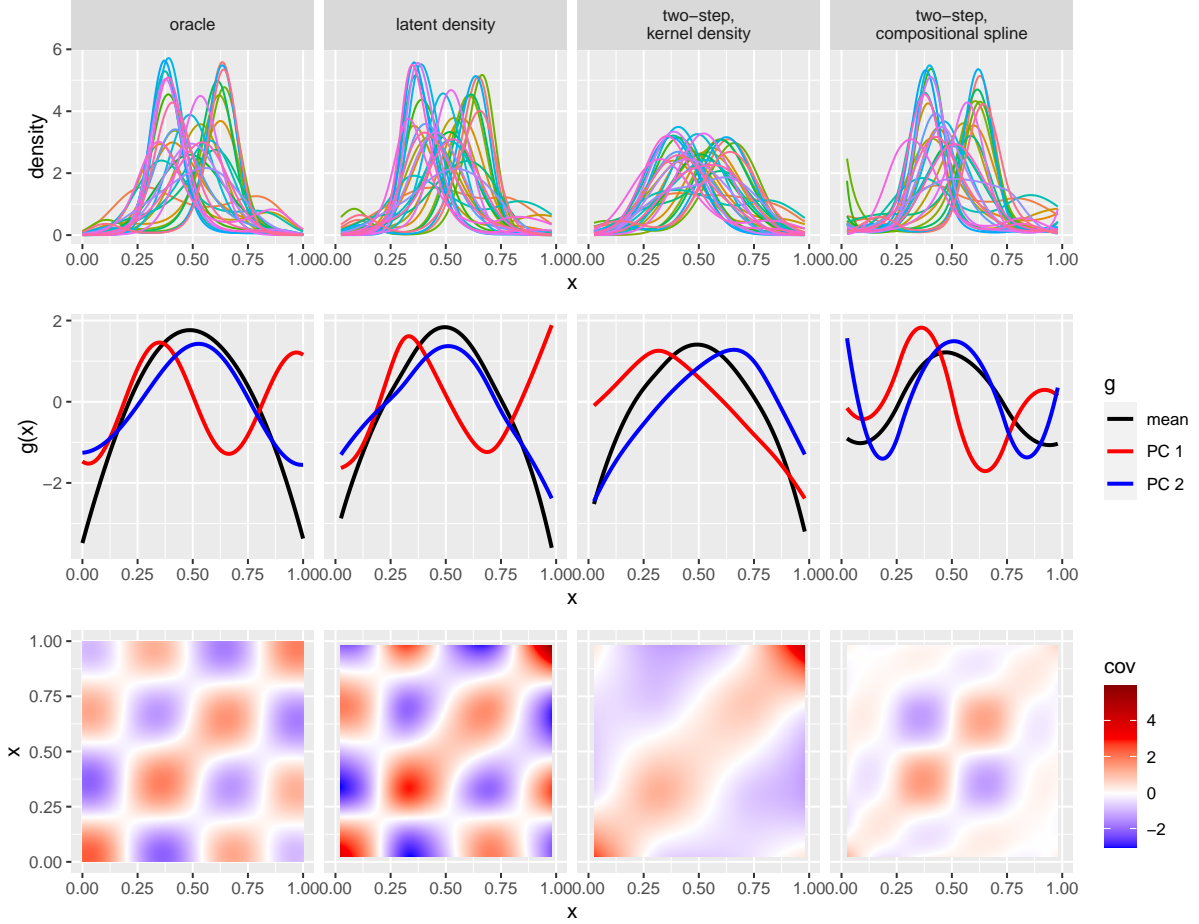


Figure 5: One randomly selected simulation run with $m_i = 40$ observations per density. Here 'oracle' refers to estimation based on the true underlying densities. Top: Density estimates/predictions. Middle: Mean and first two principal components. Bottom: Estimated covariance structure.

In this example, the estimates for the mean function, the first two principal components, and the covariance function of the latent density model (second column) appear similar to the corresponding oracle estimates (first column). Also, the pattern of predicted latent densities (first row, third column) is similar to the pattern of true underlying densities. For the two-step approach based on kernel density estimates (third row), the modes of the estimated densities appear to be too low, and although the mean has a similar shape as the oracle mean, the first two principal components, as well as the covariance function, differ substantially from the oracle estimates.

The compositional splines used for the two-stage approach in the fourth row appear to behave similarly to the true densities near the center of the observed interval (in $\approx [0.2, 0.8]$), but show implausible characteristics near the boundary, i.e., near 0 and 1. This is also evident in the estimates for the mean and the first two principal components. Similarly, the estimate for the covariance function seems to be close to the oracle estimate in the center of the $I \times I$ domain, but not at the boundaries.

5 Discussion

We have proposed improvements to existing PCA methods for densities in the Bayes Hilbert space, by explicitly incorporating in a maximum likelihood approach that there are usually only discrete samples of the densities available. This differs from two-step approaches, where the densities are first estimated in a preprocessing step and then PCA is performed for these estimates in the Bayes Hilbert space, ignoring uncertainty arising from the first density estimation step. We confirmed in applications and in a simulation that our latent density model can be successfully estimated employing an MCEM algorithm, and that the estimation of the mean and covariance structure for the densities is superior to the estimation using two-step approaches. While improvements are particularly pronounced for small samples per density, differences persist even for moderately large samples.

Consequently, resulting estimates for the principal components using our approach are better suited for understanding the variation in the underlying densities, and the predicted scores can be better used for dimension reduction and for further analyses, such as to describe differences and trends in the densities. In addition, given the importance of principal components for dimension reduction in functional data analysis (Ramsay and Silverman, 2005; Chiou and Li, 2007), the principal components could also be used as the basis for further subsequent functional data analysis methods, such as as a basis in which to expand model terms in a functional regression model (as for example in Yao et al., 2005; Scheipl et al., 2015; Volkmann et al., 2023) in the Bayes Hilbert space. Finally, the predicted latent densities could also be used for subsequent analysis, providing more reliable reconstructions of the underlying true densities than the density estimates obtained using preprocessing with kernel density estimation or compositional splines.

To allow maximum likelihood estimation of the principal components, we assume a normal distribution of the scores, i.e. a Gaussian process (prior) for the latent densities, as common in many statistical methods ranging from mixed models (Guo, 2002) to Gaussian process regression (Rasmussen and Williams, 2005). This distributional assumption may be too restrictive in some applications, however, in such cases leading to unjustified shrinkage of the latent densities. In future research, it would thus be appealing to extend our approach further by including a choice of different distributions for the scores in the latent density model, e.g. to account for heavy tails.

In this paper, our primary focus was on continuous densities with respect to the Lebesgue measure on a bounded interval. Furthermore, we provided a brief outline of a potential extension to discrete measures, which incorporates compositional data. In addition to an application of our model to the compositional data case often encountered in practical data scenarios, an interesting direction for future research would be a generalization of the approach and implementation to other measures. This could include mixed discrete and continuous measures as in Maier et al. (2022), the inclusion of unbounded domains as well as the multivariate case.

Acknowledgements

We gratefully acknowledge funding by grants GR 3793/3-1 ‘Flexible regression methods for curve and shape data’ and GR 3793/8-1 ‘Flexible density regression methods’ from the German Research Foundation (DFG). We would also like to thank the data providers in the ECA&D project for providing the temperature data for the first application and Michael Windmann for sharing the Munich rent data for the second application.

References

- J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982. doi: <https://doi.org/10.1111/j.2517-6161.1982.tb01195.x>.
- J eremie Bigot, Ra ul Gouet, Thierry Klein, and Alfredo Lopez. Geodesic PCA in the Wasserstein space. *Annales de l’Institut Henri Poincar e, Probabilit es et Statistiques*, 53, 07 2013. doi: 10.1214/15-AIHP706.
- Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.
- 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.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. ISSN 00359246. URL <http://www.jstor.org/stable/2984875>.
- Juan Jos e Egozcue, Vera Pawlowsky-Glahn, Gl oria Mateu-Figueras, and Carles Barcelo-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical geology*, 35(3):279–300, 2003.
- Juan Jos e Egozcue, Jos e Luis D ıaz-Barrero, and Vera Pawlowsky-Glahn. Hilbert space of probability density functions based on Aitchison geometry. *Acta Mathematica Sinica*, 22:1175–1182, 2006.
- Ludwig Fahrmeir, G oran Kauermann, Gerhard Tutz, and Michael Windmann. Spatial smoothing revisited: An application to rental data in Munich. *Statistical Modelling*, 0(0):online before print, 2023.
- Peter Filzmoser, Karel Hron, Matthias Templ, Peter Filzmoser, Karel Hron, and Matthias Templ. Analyzing compositional data using R. *Applied Compositional Data Analysis: With Worked Examples in R*, pages 17–34, 2018.
- Wensheng Guo. Functional mixed effects models. *Biometrics*, 58(1):121–128, 2002.
- J.M. Hammersley and D.C. Handscomb. *Monte Carlo Methods*. Methuen’s monographs on applied probability and statistics. Methuen, 1964. ISBN 9780416523409. URL <https://books.google.de/books?id=Kk40AAAAQAAJ>.
- Karel Hron, Alessandra Menafoglio, Matthias Templ, Klara Hruzova, and Peter Filzmoser. Simplicial principal component analysis for density functions in Bayes spaces. *Comput. Stat. Data Anal.*, 94:330–350, 2016.
- Tailen Hsing and Randall Eubank. *Theoretical foundations of functional data analysis, with an introduction to linear operators*, volume 997. John Wiley & Sons, 2015.
- Kim Huynh and David Jacho-Chavez. Firm size distributions through the lens of functional principal components analysis. *Journal of Applied Econometrics*, 25:1211–1214, 11 2010. doi: 10.1002/jae.1200.
- Kari Karhunen. Zur Spektraltheorie stochastischer Prozesse. In *Annales Academiae Scientiarum Fennicae Series A*, volume 1, page 34, 1946.
- A. M. G. Klein Tank, J. B. Wijngaard, G. P. K onnen, R. B ohm, G. Demar ee, A. Gocheva, M. Mileta, S. Pashiardis, L. Hejkrlik, C. Kern-Hansen, R. Heino, P. Bessemoulin, G. M uller-Westermeier, M. Tzanakou, S. Szalai, T. Palisd ottir, D. Fitzgerald, S. Rubin, M. Capaldo, M. Maugeri, A. Leitass, A. Bukantis, R. Aberfeld, A. F. V. van Engelen, E. Forland, M. Mielus, F. Coelho, C. Mares, V. Razuvaev, E. Nieplova, T. Cegnar, J. Antonio L opez, B. Dahlstr om, A. Moberg, W. Kirchhofer, A. Ceylan, O. Pachaliuk, L. V. Alexander, and P. Petrovic. Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. *International Journal of Climatology*, 22(12):1441–1453, 2002.
- Alois Kneip and Klaus Utikal. Inference for density families using functional principal component analysis. *Journal of the American Statistical Association*, 96:519–542, 02 2001. doi: 10.1198/016214501753168235.
- M Lo eve. Fonctions al eatoires  a d ecomposition orthogonale exponentielle. *La Revue Scientifique*, 84:159–162, 1946.

- Jitka Machalová, Renáta Talská, Karel Hron, and Aleš Gába. Compositional splines for representation of density functions. *Computational Statistics*, 36:1–34, 06 2021. doi: 10.1007/s00180-020-01042-7.
- Eva-Maria Maier, Almond Stöcker, Bernd Fitzenberger, and Sonja Greven. Additive density-on-scalar regression in Bayes hilbert spaces with an application to gender economics, 2022. arXiv:2110.11771.
- Marcos Matabuena, Alexander Petersen, Juan C Vidal, and Francisco Gude. Glucodensities: a new representation of glucose profiles using distributional data analysis. *Statistical methods in medical research*, 30(6):1445–1464, 2021.
- David Nerini and Badih Ghattas. Classifying densities using functional regression trees: Applications in oceanology. *Computational Statistics & Data Analysis*, 51(10):4984–4993, 2007. ISSN 0167-9473. doi: <https://doi.org/10.1016/j.csda.2006.09.028>. URL <https://www.sciencedirect.com/science/article/pii/S0167947306003550>.
- Victor M. Panaretos and Yoav Zemel. Statistical aspects of Wasserstein distances. *Annual Review of Statistics and Its Application*, 6(1):405–431, 2019. doi: 10.1146/annurev-statistics-030718-104938. URL <https://doi.org/10.1146/annurev-statistics-030718-104938>.
- Alexander Petersen and Hans-Georg Müller. Functional data analysis for density functions by transformation to a Hilbert space. *Annals of Statistics*, 44:183–218, 2016.
- Jiaming Qiu, Xiongtao Dai, and Zhengyuan Zhu. Nonparametric estimation of repeated densities with heterogeneous sample sizes. *Journal of the American Statistical Association*, 0(0):1–13, 2022. doi: 10.1080/01621459.2022.2104728. URL <https://doi.org/10.1080/01621459.2022.2104728>.
- James O. Ramsay and Bernhard W. Silverman. *Functional Data Analysis*. Springer New York, 2005.
- C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning series. MIT Press, 2005. ISBN 9780262182539.
- Fabian Scheipl, Ana-Maria Staicu, and Sonja Greven. Functional additive mixed models. *Journal of Computational and Graphical Statistics*, 24(2):477–501, 2015. doi: 10.1080/10618600.2014.901914. PMID: 26347592.
- Riccardo Scimone, Alessandra Menafoglio, Laura Sangalli, and Piercesare Secchi. A look at the spatio-temporal mortality patterns in Italy during the COVID-19 pandemic through the lens of mortality densities. *Spatial Statistics*, 49:100541, 09 2021. doi: 10.1016/j.spasta.2021.100541.
- Anuj Srivastava, Ian Jermyn, and Shantanu Joshi. Riemannian analysis of probability density functions with applications in vision. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. doi: 10.1109/CVPR.2007.383188.
- Karl Gerald van den Boogaart, Juan José Egozcue, and Vera Pawlowsky-Glahn. Bayes hilbert spaces. *Australian & New Zealand Journal of Statistics*, 56(2):171–194, 2014. doi: <https://doi.org/10.1111/anzs.12074>.
- Alexander Volkmann, Almond Stöcker, Fabian Scheipl, and Sonja Greven. Multivariate functional additive mixed models. *Statistical Modelling*, 23(4):303–326, 2023. doi: 10.1177/1471082X211056158.
- Greg C. G. Wei and Martin A. Tanner. A Monte Carlo implementation of the EM algorithm and the poor man’s data augmentation algorithms. *Journal of the American Statistical Association*, 85:699–704, 1990.
- M Windmann and G Kauermann. Statistische Analyse der Nettomieten. In *Mietspiegel für München 2019. Statistik, Dokumentation und Analysen*, pages 19–70. Landeshauptstadt München, Sozialreferat, Amt für Wohnen und Migration, 2019. Chapter 2.
- S.N. Wood. *Generalized Additive Models: An Introduction with R, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press, 2017. ISBN 9781498728348. URL <https://books.google.de/books?id=HL-PDwAAQBAJ>.
- Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, 33(6):2873–2903, 2005.

A Proofs and Computations

A.1 Proof of Lemma 2.2

Proof. The clr transformation is well defined as

$$\left| \int_I \log(f(x)) dx \right| \leq |I| \|\log(f)\|_{\mathbb{L}^2} < \infty,$$

due to the Cauchy-Schwarz inequality and $\log(f) \in \mathbb{L}^2$. We further compute

$$\begin{aligned} \text{clr}([\alpha f]) &= \log(\alpha f) - \frac{1}{|I|} \int_I \log(\alpha f(x)) dx \\ &= \log(\alpha) + \log(f) - \frac{1}{|I|} \int_I \log(\alpha) + \log(f(x)) dx \\ &= \log(f) - \frac{1}{|I|} \int_I \log(f(x)) dx = \text{clr}([f]) \end{aligned}$$

for all $\alpha \in \mathbb{R}$ and $f \in B$. Since clr^{-1} is clearly well defined as well, we show that clr is bijective via showing

$$\text{clr}(\text{clr}^{-1}(g)) = \text{clr}([\exp(g)]) = g - \frac{1}{|I|} \int_I g(x) dx = g$$

for all $g \in \mathbb{L}_0^2$, and

$$\text{clr}^{-1}(\text{clr}(f)) = \left[\exp \left(\log(f) - \frac{1}{|I|} \int_I \log(f(x)) dx \right) \right] = \left[f \exp \left(-\frac{1}{|I|} \int_I \log(f(x)) dx \right) \right] = [f]$$

for all $f \in B$. The clr transformation is also an isometry as for all $g_1, g_2 \in \mathbb{L}_0^2$ holds

$$\begin{aligned} \langle [\text{clr}^{-1}(g_1)], [\text{clr}^{-1}(g_2)] \rangle_{\mathcal{B}} &= \frac{1}{2|I|} \int_I \int_I \log \left(\frac{\exp(g_1(x))}{\exp(g_1(y))} \right) \log \left(\frac{\exp(g_2(x))}{\exp(g_2(y))} \right) dx dy \\ &= \frac{1}{2|I|} \int_I \int_I (g_1(x) - g_1(y)) (g_2(x) - g_2(y)) dx dy \\ &= \frac{1}{2|I|} \int_I \int_I (g_1(x)g_2(x) - g_1(y)g_2(x) - g_1(x)g_2(y) + g_1(y)g_2(y)) dx dy \\ &= \frac{1}{2} \int_I g_1(x)g_2(x) dx - \frac{1}{2|I|} \int_I g_1(y) dy \int_I g_2(x) dx - \frac{1}{2|I|} \int_I g_1(x) dx \int_I g_2(y) dy \\ &\quad + \frac{1}{2} \int_I g_1(y)g_2(y) dy \\ &= \int_I g_1(x)g_2(x) dx = \langle g_1, g_2 \rangle_{\mathbb{L}^2} \end{aligned}$$

since $\int_I g_1(x) dx = \int_I g_2(x) dx = 0$. □

A.2 Proof of Corollary 2.4

Proof. It is trivial to see that if $G = \sum_{k=1}^N \theta_k e_k$ with $\theta \stackrel{i.i.d.}{\sim} \mathcal{N}(\nu, \Sigma)$, then $G \stackrel{i.i.d.}{\sim} GP(\mu, K)$ with $\mu = \sum_{k=1}^N \nu_k e_k$ and $K(x_1, x_2) = \sum_{k=1}^N \sum_{l=1}^N e_k(x_1) e_l(x_2) \Sigma_{kl}$ since

$$\begin{aligned} \mathbb{E}(G(x)) &= \mathbb{E}\left(\sum_{k=1}^N \theta_k e_k(x)\right) = \sum_{k=1}^N \nu_k e_k \quad \text{and} \\ \text{Cov}(G(x_1), G(x_2)) &= \text{Cov}\left(\sum_{k=1}^N \theta_k e_k(x_1), \sum_{l=1}^N \theta_l e_l(x_2)\right) = \sum_{k=1}^N \sum_{l=1}^N e_k(x_1) e_l(x_2) \text{Cov}(\theta_k, \theta_l) \\ &= \sum_{k=1}^N \sum_{l=1}^N e_k(x_1) e_l(x_2) \Sigma_{kl}. \end{aligned} \tag{11}$$

The other direction is an implication of the Karhunen-Loève decomposition. If \mathcal{H} is N' -dimensional, $G \stackrel{i.i.d.}{\sim} GP(\mu, K)$ can be decomposed as $G = \mu + \sum_{k=1}^{N'} Z_k \varphi_k$ with $\varphi_k, k = 1, \dots, N'$ being the orthonormal eigenfunctions and uncorrelated scores Z_k with $\mathbb{E}(Z_k) = 0$ and $\text{Var}(Z_k) = \sigma_k^2$.

Choose $\nu_k = \langle \mu, e_k \rangle_{\mathbb{L}_2}$ to be the orthonormal projection on e_k and $\Sigma_{kl} = \sum_{j=1}^{N'} \sigma_j^2 \langle \varphi_j, e_k \rangle_{\mathbb{L}_2} \langle \varphi_j, e_l \rangle_{\mathbb{L}_2}$ for all $k, l = 1, \dots, N$. Then we compute

$$\sum_{k=1}^N \nu_k e_k = \sum_{k=1}^N \langle \mu, e_k \rangle_{\mathbb{L}_2} e_k = \mu,$$

since this gives the orthogonal projection of μ on $\text{span}\{e_1, \dots, e_N\}$ and $\mu \in \mathcal{H} \subseteq \text{span}\{e_1, \dots, e_N\}$. We further compute using the same identity

$$\begin{aligned} \text{Cov}(G(x_1), G(x_2)) &= \text{Cov}\left(\sum_{k=1}^{N'} Z_k \varphi_k(x_1), \sum_{l=1}^{N'} Z_l \varphi_l(x_2)\right) = \sum_{k=1}^{N'} \sum_{l=1}^{N'} \varphi_k(x_1) \varphi_l(x_2) \text{Cov}(Z_k, Z_l) \\ &= \sum_{k=1}^{N'} \sigma_k^2 \varphi_k(x_1) \varphi_k(x_2) \end{aligned}$$

which is identical to

$$\begin{aligned} \sum_{k=1}^N \sum_{l=1}^N e_k(x_1) e_l(x_2) \Sigma_{kl} &= \sum_{k=1}^N e_k(x_1) \sum_{l=1}^N e_l(x_2) \sum_{j=1}^{N'} \sigma_j^2 \langle \varphi_j, e_k \rangle_{\mathbb{L}_2} \langle \varphi_j, e_l \rangle_{\mathbb{L}_2} \\ &= \sum_{j=1}^{N'} \sigma_j^2 \sum_{k=1}^N e_k(x_1) \langle \varphi_j, e_k \rangle_{\mathbb{L}_2} \sum_{l=1}^N e_l(x_2) \langle \varphi_j, e_l \rangle_{\mathbb{L}_2} \\ &= \sum_{j=1}^{N'} \sigma_j^2 \varphi_j(x_1) \varphi_j(x_2). \end{aligned}$$

To show the correspondence of the eigenvalue decompositions we need to show that if $v_l = (v_{l1}, \dots, v_{lN})$ is an eigenvector of Σ with corresponding eigenvalue σ_l^2 then $\varphi_l = \sum_{m=1}^N v_{lm} e_m$ is an eigenfunction of K with the same eigenvalue σ_l^2 . This is true since if we plug in the formula obtained in (11) for K we obtain

$$\begin{aligned}
 \int_I K(x_1, \cdot) \varphi_l(x_1) dx_1 &= \int_I K(x_1, \cdot) \sum_{m=1}^N v_{lm} e_m(x_1) dx_1 = \sum_{k=1}^N e_k \int_I \left(\sum_{j=1}^N e_j(x_1) \Sigma_{kj} \right) \sum_{m=1}^N v_{lm} e_m(x_1) dx_1 \\
 &= \sum_{k=1}^N e_k \sum_{j=1}^N \Sigma_{kj} \sum_{m=1}^N v_{lm} \int_I e_j(x_1) e_m(x_1) dx_1 = \sum_{k=1}^N e_k \sum_{j=1}^N \Sigma_{kj} v_{lj} \\
 &= \sum_{k=1}^N e_k (\Sigma \mathbf{v}_l)_k = \sum_{k=1}^N e_k (\sigma_l^2 \mathbf{v}_l)_k = \sigma_l^2 \sum_{k=1}^N e_k v_{lk} = \sigma_l^2 \varphi_l.
 \end{aligned}$$

A.3 Derivation of the likelihood

$$\begin{aligned}
 L(\mu, K | \mathbf{x}_i, \dots, \mathbf{x}_n) &= \prod_{i=1}^n p(\mathbf{x}_i | \mu, K) = \prod_{i=1}^n \int_{\mathbb{R}^N} p(\mathbf{x}_i | \boldsymbol{\theta}_i) p(\boldsymbol{\theta}_i | \mu, K) d\boldsymbol{\theta}_i \\
 &= \prod_{i=1}^n \int_{\mathbb{R}^N} \left(\prod_{j=1}^{m_i} p(x_{ij} | \boldsymbol{\theta}_i) \right) p(\boldsymbol{\theta}_i | \mu, K) d\boldsymbol{\theta}_i \\
 &= \prod_{i=1}^n \int_{\mathbb{R}^N} \left(\prod_{j=1}^{m_i} \frac{\exp(\sum_{k=1}^N \theta_{ik} e_k(x_{ij}))}{\int_I \exp(\sum_{k=1}^N \theta_{ik} e_k(x)) dx} \right) p(\boldsymbol{\theta}_i | \mu, K) d\boldsymbol{\theta}_i \\
 &= \prod_{i=1}^n \int_{\mathbb{R}^N} \frac{\exp(\sum_{j=1}^{m_i} \sum_{k=1}^N \theta_{ik} e_k(x_{ij})) p(\boldsymbol{\theta}_i | \mu, K)}{\left(\int_I \exp(\sum_{k=1}^N \theta_{ik} e_k(x)) dx \right)^{m_i}} d\boldsymbol{\theta}_i
 \end{aligned}$$

□

A.4 Proof of Lemma 2.5

Proof. To show this statement we consider the three non constant additive parts of the logarithm of the conditional distribution $\log(p(\mathbf{z}_i | \mathbf{x}_i, \boldsymbol{\nu}^{(h)}, \boldsymbol{\Sigma}^{(h)}))$.

- The first part $\sum_{j=1}^{m_i} (\mu^{(h)}(x_{ij}) + \sum_{k=1}^N \mathbf{z}_i^T \mathbf{v}_k^{(h)} e_k(x_{ij}))$ is linear in \mathbf{z}_i , which means there is a constant $M_1 \in \mathbb{R}$ such that it is $M_1 \mathbf{z}_i$.
- The second part is always negative, as

$$-m_i \log \left(\int_I \exp \left(\mu^{(h)}(x) + \sum_{k=1}^N \mathbf{z}_i^T \mathbf{v}_k^{(h)} e_k(x) \right) dx \right) \leq -m_i \int_I \mu^{(h)}(x) + \sum_{k=1}^N \mathbf{z}_i^T \mathbf{v}_k^{(h)} e_k(x) dx = 0$$

where the inequality is due to Jensen's inequality and the integral is equal to zero as all functions are in \mathbb{L}_0^2 .

- The third part consists of normal densities and is therefore quadratic in \mathbf{z}_i . More precisely, we have

$$\log \left(\prod_{k=1}^N p(z_{ik} | \sigma_k^{2(h)}) \right) = \sum_{k=1}^N \frac{-z_{ik}^2}{2\sigma_k^{2(h)}} + \text{const.} \leq -\frac{1}{2\sigma_1^{2(h)}} \|\mathbf{z}_i\|^2 + \text{const.}$$

Taking these three parts together this shows that there are constants $M_1, M_2 \in \mathbb{R}$ and $M_3 > 0$ such that $\log(p(\mathbf{z}_i | \mathbf{x}_i, \boldsymbol{\nu}^{(h)}, \boldsymbol{\Sigma}^{(h)})) \leq -M_3 \|\mathbf{z}_i\|^2 + M_1 \mathbf{z}_i + M_2$, which implies that $p(\mathbf{z}_i | \mathbf{x}_i, \boldsymbol{\nu}^{(h)}, \boldsymbol{\Sigma}^{(h)})$ is a proper density. □

A.5 Example: The posterior mode will not necessarily be attained if the prior is improper

Proof. Consider the 1-dimensional case $N = 1$ with densities defined on the unit interval $I = [0, 1]$ and only one basis function for the clr transformed densities given as $e_1 = \mathbb{1}_{[0,0.5]} - \mathbb{1}_{]0.5,1]}$. This means also the parameter space \mathbb{R} in this case and assuming for this parameter $z_i \in \mathbb{R}$ an improper 1-dimensional normal distribution is equivalent to assuming a flat prior.

If we further assume there is only one observation $x_{i1} = 0.2$ and for the prior mean holds $\boldsymbol{\nu}^{(h)} = 0$ we compute

$$\begin{aligned} p(z_i | x_{i1}, \boldsymbol{\nu}^{(h)}, \boldsymbol{\Sigma}^{(h)}) &\propto \frac{\exp(z_i e_1(x_{i1}))}{\int_{[0,1]} \exp(z_i e_1(x)) dx} = \frac{\exp(z_i)}{\int_{[0,0.5]} \exp(z_i) dx + \int_{]0.5,1]} \exp(-z_i) dx} \\ &= \frac{2 \exp(z_i)}{\exp(z_i) + \exp(-z_i)} = \frac{2}{1 + \exp(-2z_i)}, \end{aligned}$$

which is monotonously increasing in $z_i \in \mathbb{R}$. Hence it does not attain its maximum and therefore $p(z_i | x_{i1}, \boldsymbol{\nu}^{(h)}, \boldsymbol{\Sigma}^{(h)})$ does not define a proper distribution in this case. \square

A.6 Proof of Lemma 2.6

Like in the proof of Lemma 2.5 (Appendix A.4) we consider the three non constant additive parts of the logarithm of the conditional distribution $\log(p(\mathbf{z}_i | \mathbf{x}_i, \boldsymbol{\nu}^{(h)}, \boldsymbol{\Sigma}^{(h)}))$.

- For the linear part $\sum_{j=1}^{m_i} \left(\mu^{(h)}(x_{ij}) + \sum_{k=1}^N \mathbf{z}_i^T \mathbf{v}_k^{(h)} e_k(x_{ij}) \right)$ the gradient with respect to \mathbf{z}_i is given as

$$\frac{\partial}{\partial \mathbf{z}_i} \sum_{j=1}^{m_i} \sum_{k=1}^N \mathbf{z}_i^T \mathbf{v}_k^{(h)} e_k(x_{ij}) = \sum_{j=1}^{m_i} \sum_{k=1}^N \mathbf{v}_k^{(h)} e_k(x_{ij}) = \sum_{k=1}^N \mathbf{v}_k^{(h)} \sum_{j=1}^{m_i} e_k(x_{ij})$$

- For the second part we compute the gradient as

$$\begin{aligned} &\frac{\partial}{\partial \mathbf{z}_i} - m_i \log \left(\int_I \exp \left(\mu^{(h)}(x) + \sum_{k=1}^N \mathbf{z}_i^T \mathbf{v}_k^{(h)} e_k(x) \right) dx \right) \\ &= \frac{-m_i \int_I \frac{\partial}{\partial \mathbf{z}_i} \exp \left(\mu^{(h)}(x) + \sum_{k=1}^N \mathbf{z}_i^T \mathbf{v}_k^{(h)} e_k(x) \right) dx}{\int_I \exp \left(\mu^{(h)}(x) + \sum_{k=1}^N \mathbf{z}_i^T \mathbf{v}_k^{(h)} e_k(x) \right) dx} \\ &= \frac{-m_i \int_I \exp \left(\mu^{(h)}(x) + \sum_{k=1}^N \mathbf{z}_i^T \mathbf{v}_k^{(h)} e_k(x) \right) \sum_{k=1}^N \mathbf{v}_k^{(h)} e_k(x) dx}{\int_I \exp \left(\mu^{(h)}(x) + \sum_{k=1}^N \mathbf{z}_i^T \mathbf{v}_k^{(h)} e_k(x) \right) dx} \\ &= -m_i \int_I \frac{\exp \left(\mu^{(h)}(x) + \sum_{k=1}^N \mathbf{z}_i^T \mathbf{v}_k^{(h)} e_k(x) \right)}{\int_I \exp \left(\mu^{(h)}(x) + \sum_{k=1}^N \mathbf{z}_i^T \mathbf{v}_k^{(h)} e_k(x) \right) dx} \sum_{k=1}^N \mathbf{v}_k^{(h)} e_k(x) dx \\ &= -m_i \int_I \text{clr}^{-1} \left(\mu^{(h)} + \sum_{k=1}^N \mathbf{z}_i^T \mathbf{v}_k^{(h)} e_k \right) (x) \sum_{k=1}^N \mathbf{v}_k^{(h)} e_k(x) dx \\ &= -m_i \sum_{k=1}^N \mathbf{v}_k^{(h)} \int_I f_{\mathbf{z}_i}(x) e_k(x) dx \end{aligned}$$

where, in the first equation, we can interchange differentiation and integration applying the Leibniz rule, since we have assumed that all clr transformed densities are bounded.

- The third part is the sum of logarithms of normal densities. Therefore, for all $l = 1, \dots, N$ we compute the partial derivative with respect to z_{il} as

$$\frac{\partial}{\partial z_{il}} \log\left(\prod_{k=1}^N p(z_{ik} | \sigma_k^2(h))\right) = \frac{\partial}{\partial z_{il}} \frac{-z_{il}^2}{2\sigma_l^2(h)} = \frac{-z_{il}}{\sigma_l^2(h)}.$$

Adding these three parts together gives the gradient of the logarithm of the conditional density of the scores.

B Additional plots for the applications in Section 3

B.1 Temperature data

For the temperature data, the latent density model is estimated using kernel density estimates with a Gaussian kernel and bandwidth = 1.5 as initial estimates. The number of Monte Carlo samples in the E step 2.3.1 is chosen to be $r = 50h$, where h is the iteration index, i.e., the number of samples increases over the iterations. The tuning parameter λ for the proposal density is set to the default value 1. In each iteration, the dimension reduction is at most 0.001, i.e. we keep as many principal components as necessary to explain at least 99.99 % of the variance.

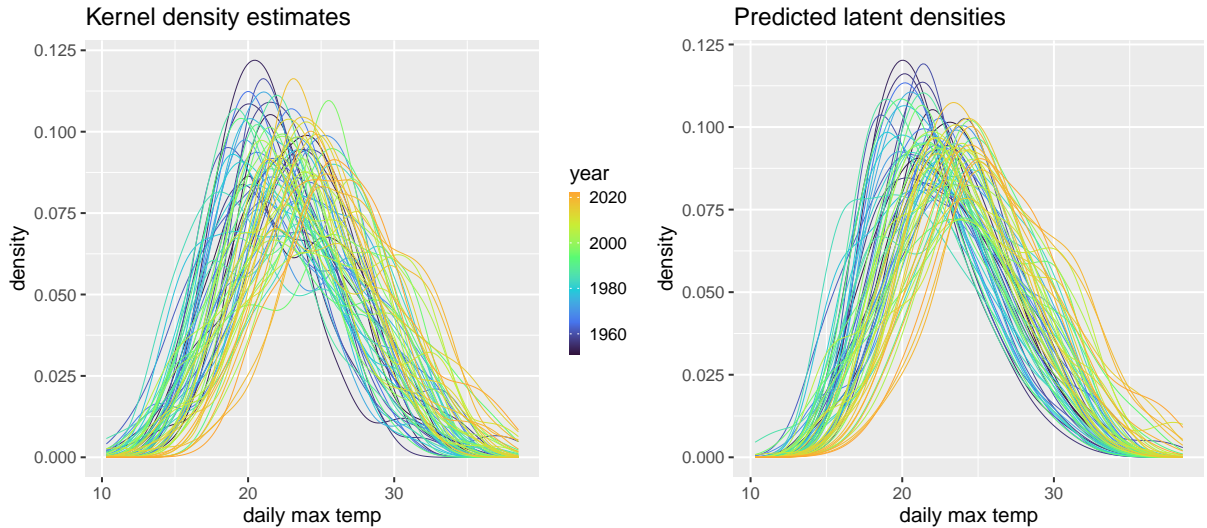


Figure 6: Kernel density estimates (Gaussian kernel, bandwidth = 1.5) and predicted latent densities estimated from the latent density model (5) for the daily maximum summer temperatures in Berlin Tempelhof from 1951 to 2022. The kernel densities are also used as initial estimates in the latent density model.

In Figure 6 we show the kernel density estimates and the predicted latent densities obtained from the latent density model. For both, the trend towards higher temperatures is evident. Note that the variance of the predicted latent densities is smaller than the variance of the kernel density estimates, since the latent density model effectively splits the total variance into the variance due to the underlying stochastic process for the latent densities and the variance due to sampling from them. Correspondingly, percentages variance explained for the PCs depicted in Figures 1 and 7, respectively, are relative to different total variances.

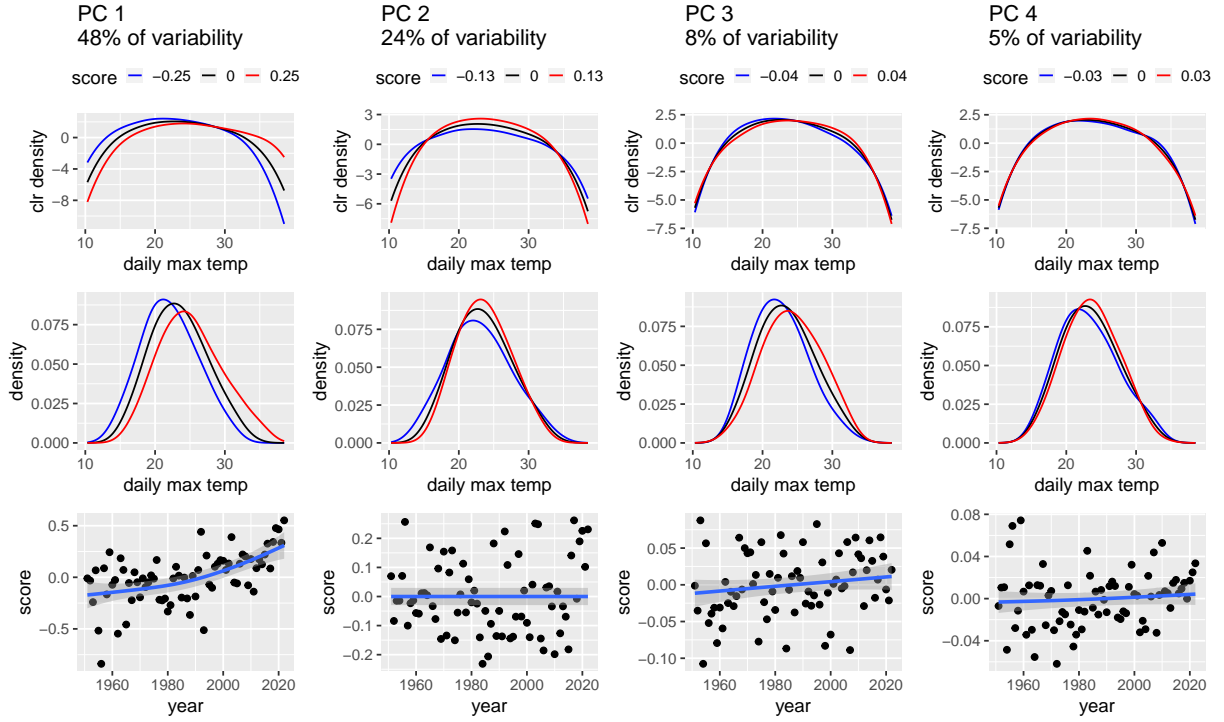


Figure 7: PCA based on the clr transformations of the kernel density estimates (Gaussian kernel, bandwidth = 1.5) for daily maximum temperature (°C). Top: Effect of adding/subtracting $\hat{\sigma}_k \hat{\varphi}_k$ to the clr transformed mean density μ , where $\hat{\varphi}_k$ is the k th principal component, with corresponding eigenvalue $\hat{\sigma}_k^2$, $k = 1, 2, 3, 4$. Middle: Effect on the density level, i.e. clr^{-1} transformations of the functions in the top row. Bottom: Temporal trend of the corresponding predicted scores per year, with scatterplot smoother and pointwise confidence bands overlaid.

B.2 Rental prices

For the rent index data, the latent density model is estimated using kernel density estimates with a Gaussian kernel and bandwidth = 2 as initial estimates. The number of Monte Carlo samples in the E step 2.3.1 is chosen to be $r = 100h$, where h is the iteration index, i.e., the number of samples increases over the iterations. The parameter λ for the proposal density is set to 2. In each iteration, the dimension reduction is at most 0.0005, i.e. we keep as many principal components as necessary to explain at least 99.995 % of the variance.

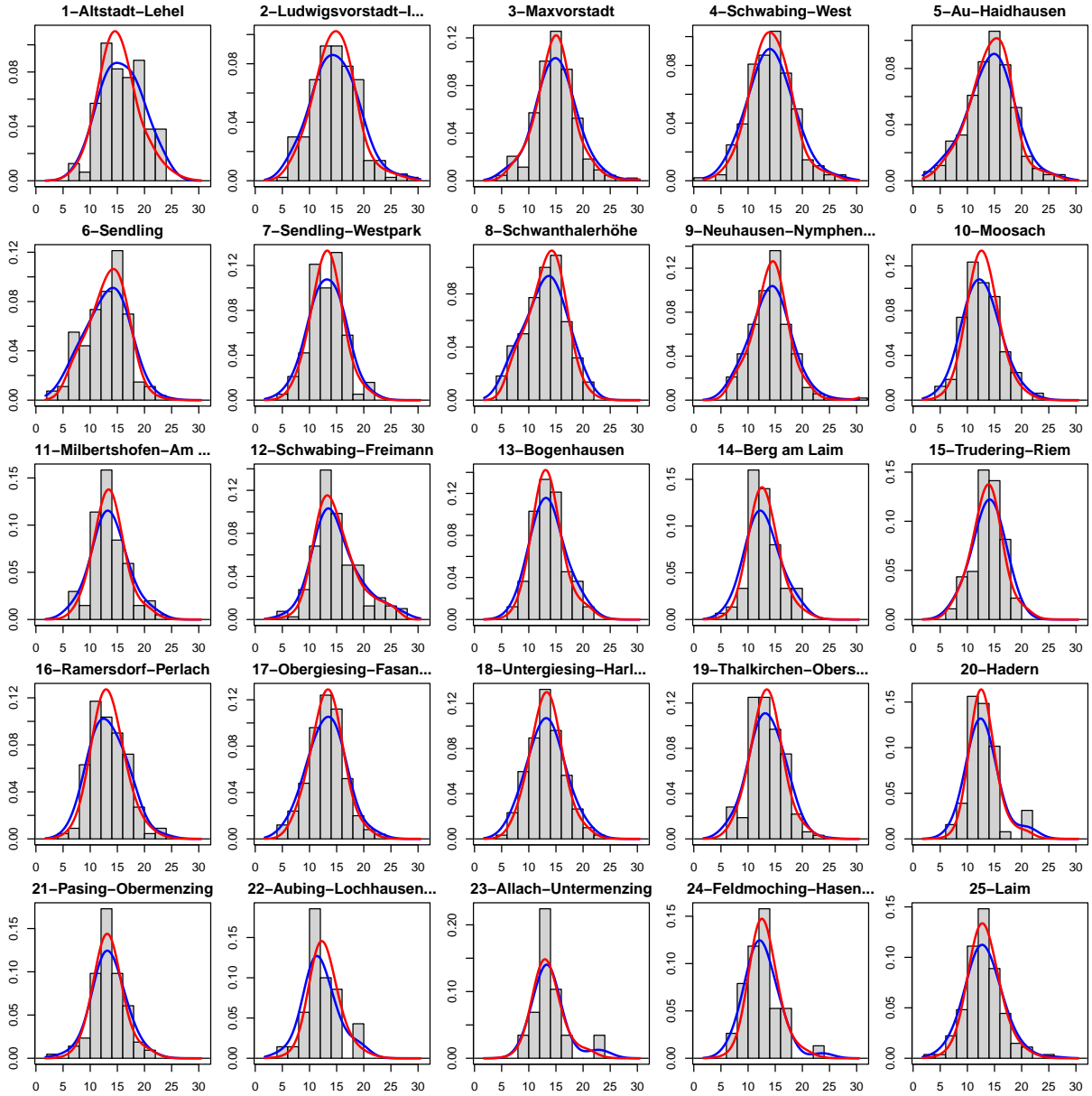


Figure 8: Histograms of rent per square meter for each district, with overlaid kernel density estimates (Gaussian kernel, bandwidth = 2) in blue and predicted latent densities based on our proposed approach in red.

Appendix

A. Online Supplement for Paper I

“Elastic Analysis of Irregularly or Sparsely Sampled Curves”

Supporting information for the contribution:

Steyer, L., Stöcker, A., and Greven, S. (2023). Elastic analysis of irregularly or sparsely sampled curves. *Biometrics*, 79:2103–2115. DOI: [10.1111/biom.13706](https://doi.org/10.1111/biom.13706)

Supporting Information for *Elastic analysis of irregularly or sparsely sampled curves*

by Lisa Steyer, Almond Stöcker and Sonja Greven

Web Appendix A Adaptations for closed curves	1
Web Appendix B Proofs and computations	3
B.1 Proof of Lemma 1	3
B.2 Gradient of the objective function function in Lemma 1	5
B.3 Closed form solution for the coordinate wise maximization	6
B.4 Proof of Theorem 1	7
B.5 Integral approximation for SRV-spline mean computation	10
B.6 Proof of Theorem 2	11
B.7 Proof of Corollary 1	12
B.8 Proof of Lemma 2	13
Web Appendix C Examples and counterexamples	13
C.1 Optimal warping and Fréchet means are not unique	13
C.2 Identifiability of constant SRV splines	16
C.3 Splines of degree four are not identifiable	16
Web Appendix D Further simulations and supplementary plots	17
D.1 Simulation: Aligning sparsely and irregularly sampled curves	17
D.2 Simulation: Convergence of spline mean coefficients	18
D.3 Simulation: Misspecified spline model	21
Web Appendix E Classifying spiral curve drawings for detecting Parkinson’s disease	21
E.1 Classification based on the elastic distance to a template	22
E.2 Warping functions of misclassified subjects	23
E.3 Influence of down-sampling on the accuracy	24
E.4 Comparison with the package “fdasrvf”	25

Web Appendix A Adaptations for closed curves

Analogously to the warping problem for open curves (or closed curves with known start and end point) we can formulate a similar criterion for closed curves, using a different set of warping functions. Here we assume $\gamma : [0, 1] \rightarrow [0, 1]$ such that there exists $t_0 \in [0, 1]$ with

$$\gamma(t_0) = 0, \quad \lim_{t \nearrow 1} \gamma(t) = \gamma(0), \quad \lim_{t \nearrow t_0} \gamma(t) = 1,$$

and γ monotonically increasing and differentiable on $[0, t_0[$ and on $[t_0, 1]$. This allow us to obtain a similar result as in Lemma 1 for closed curves.

Corollary A.1 (Optimization problem for closed curves). *Let \mathbf{p} and \mathbf{q} be as in Lemma 1 and additionally let them be the SRV transformations of closed curves. Let \mathbf{p}^* be the periodic extension of \mathbf{p} to the whole real line, that is $\mathbf{p}^*(t) = \mathbf{p}(t - [t])$ for all $t \in \mathbb{R}$. Then the optimization problem for closed curves is equivalent to the following problem.*

$$\begin{aligned} \text{Maximize} \quad & \Phi^*(\mathbf{t}) = \Phi^*(t_0, t_1, \dots, t_{m-1}) = \sum_{j=0}^{m-1} \sqrt{(s_{j+1} - s_j) \int_{t_j}^{t_{j+1}} \langle \mathbf{p}^*(t), \mathbf{q}_j \rangle_+^2 dt} \\ \text{w.r.t} \quad & t_0 \leq t_1 \leq \dots \leq t_m = t_0 + 1. \end{aligned} \quad (1)$$

For a maximizer $(t_0, t_1, \dots, t_{m-1})$ of (1) there is a $\gamma : [0, 1] \rightarrow [0, 1]$ with $\gamma(t_j - [t_j]) = s_j$ for all $j = 0, \dots, m-1$ which is a minimizer of the corresponding warping problem for closed curves.

A further advantage of Algorithm 1 is that it can be easily adapted to closed curves, which has not been explicitly addressed by Lahiri et al. (2015). We adjust our algorithm for open polygons via appropriately updating t_0 and t_m .

Algorithm 3: Elastic distance for two closed polygons

Input: piecewise constant SRV-curves \mathbf{p}, \mathbf{q} ; convergence tolerance $\epsilon > 0$;

starting values $0 \leq t_1^{(0)} \leq \dots \leq t_{m-1}^{(0)} \leq t_m^{(0)} = t_0^{(0)} + 1$; // e.g. relative arc length

for $k \in \mathbb{N}$ **do**

for $j = 1, \dots, m-1$ **do**

if $j - k$ even **then**

$$\left[t_j^{(k)} = \operatorname{argmax}_{t_j \in [t_{j-1}^{(k-1)}, t_{j+1}^{(k-1)}]} \Phi \Big|_{\{t_{j'} = t_{j'}^{(k-1)}, j' \neq j\}} \right]$$

else if $j - k$ odd **then**

$$\left[t_j^{(k)} = t_j^{(k-1)} \right]$$

if k even **then**

$$\left[\begin{aligned} t_0^{(k)} &= \operatorname{argmax}_{t_0 \in [t_{m-1}^{(k)} - 1, t_1^{(k)}]} \Phi^* \Big|_{\{t_{j'} = t_{j'}^{(k)}, j' \neq 0\}}; \\ t_m^{(k)} &= t_0^{(k)} + 1 \end{aligned} \right]$$

if $\|\mathbf{t}^{(k)} - \mathbf{t}^{(k-2)}\| < \epsilon$ **and** $\|\mathbf{t}^{(k-1)} - \mathbf{t}^{(k-3)}\| < \epsilon$ **then**

return $\mathbf{t}^{(k)} = (t_1^{(k)}, \dots, t_{m-1}^{(k)})$

The optimal warping function $\gamma : [0, 1] \rightarrow [0, 1]$ then fullfills $\gamma(t_j^{(k+2)} - [t_j^{(k+2)}]) = s_j$ for all $j = 0, \dots, m-1$. Moreover, the algorithm for computing an elastic spline mean needs to be adjusted accordingly as well.

Remark A.2 (Smooth elastic mean for closed curves). *We replace the warping step of Algorithm 2, i.e. updating the optimal parametrizations γ_i , by considering the corresponding minimization problem for closed curves (1) via gradient descent or Algorithm 3 depending on the spline degree. For updating the least-squares estimate for given parametrizations, we use a penalty function method to deal with the non-linear constraint of closedness for $\bar{\mathbf{p}}$ (see for example Sun and Yuan (2006)). Thus, we add a cost function penalizing openness with increasing weight. Precisely, in the k -th iteration step, we consider the loss function*

$$\sum_{i=1}^n \inf_{\gamma_i} \left\| \bar{\mathbf{p}} - (\mathbf{q}_i \circ \gamma_i) \sqrt{\gamma_i} \right\|_{L_2}^2 + \lambda_k \left\| \int_0^1 \bar{\mathbf{p}}(t) \|\bar{\mathbf{p}}(t)\| dt \right\|^2,$$

with $\lambda_k \rightarrow \infty$ for $k \rightarrow \infty$. Since $\int_0^1 \bar{\mathbf{p}}(t) \|\bar{\mathbf{p}}(t)\| dt = \bar{\beta}(1) - \bar{\beta}(0)$, if $\bar{\mathbf{p}}$ is the SRV of $\bar{\beta}$, the penalty term vanishes if and only if $\bar{\beta}$ is closed.

Figure 1 shows three iterations of this adapted algorithm for calculating a smooth mean of four, irregularly sampled, closed heart shapes. The initial mean (iteration 0) was computed as a least-squares-estimate assuming the curves were parametrized by relative arc length. The sequence $(\lambda_k)_{k \in \mathbb{N}}$ was chosen as $\lambda_k = 10^{-3}k$ for all $k \in \mathbb{N}$.

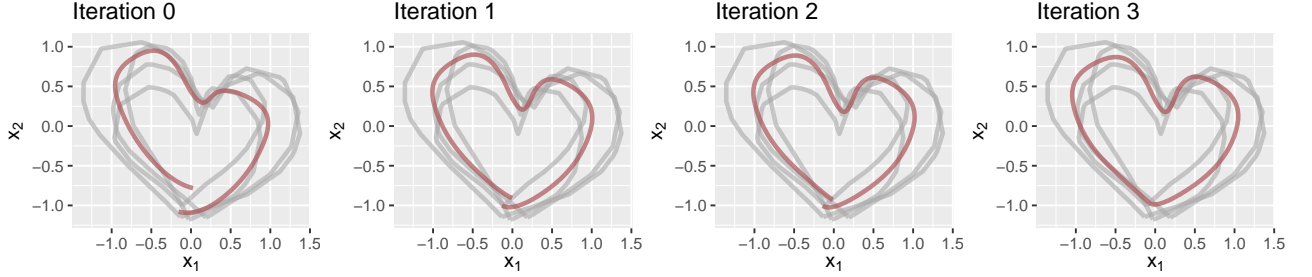


Figure 1: First three iterations of the algorithm for closed mean curves on a toy dataset.

Web Appendix B Proofs and computations

In this part of the appendix we provide proofs to all statements presented in the main article.

B.1 Proof of Lemma 1

Proof. We will prove this statement for optimization with respect to $\bar{\Gamma}$, the set of absolutely continuous curves $\gamma : [0, 1] \rightarrow [0, 1]$, onto and with $\dot{\gamma} \geq 0$ almost everywhere (a.e.). The statement for Γ follows as Γ is dense in $\bar{\Gamma}$ and the warping action of $\bar{\Gamma}$ continuous (Bruveris, 2016). Hence, to compute the elastic distance between two absolutely continuous curves with SRV transformations $\mathbf{p}, \mathbf{q} : [0, 1] \rightarrow \mathbb{R}^d$ we need to consider the following minimization problem

$$\begin{aligned} \text{Minimize} \quad & \int_0^1 \|\mathbf{p}(t) - \mathbf{q}(\gamma(t))\sqrt{\dot{\gamma}(t)}\|^2 dt \\ \text{w.r.t.} \quad & \gamma : [0, 1] \rightarrow [0, 1] \text{ absolutely continuous, onto and with } \dot{\gamma} \geq 0 \text{ a.e.} \end{aligned}$$

The objective function can be written as

$$\begin{aligned} \int_0^1 \|\mathbf{p}(t) - \mathbf{q}(\gamma(t))\sqrt{\dot{\gamma}(t)}\|^2 dt &= \int_0^1 \|\mathbf{p}(t)\|^2 dt - 2 \int_0^1 \langle \mathbf{p}(t), \mathbf{q}(\gamma(t)) \rangle \sqrt{\dot{\gamma}(t)} dt \\ &\quad + \int_0^1 \|\mathbf{q}(\gamma(t))\|^2 \dot{\gamma}(t) dt \\ &= \|\mathbf{p}\|_{L_2}^2 - 2 \int_0^1 \langle \mathbf{p}(t), \mathbf{q}(\gamma(t)) \rangle \sqrt{\dot{\gamma}(t)} dt + \|\mathbf{q}\|_{L_2}^2. \end{aligned}$$

Hence the minimization problem stated above is equivalent to

$$\begin{aligned} \text{Maximize} \quad & \int_0^1 \langle \mathbf{p}(t), \mathbf{q}(\gamma(t)) \rangle \sqrt{\dot{\gamma}(t)} dt \\ \text{w.r.t.} \quad & \gamma : [0, 1] \rightarrow [0, 1] \text{ absolutely continuous, onto and with } \dot{\gamma} \geq 0 \text{ a.e.} \end{aligned}$$

We assume that \mathbf{q} is the square root velocity curve of a polygon (for example a polygon with observations at its corners). Hence \mathbf{q} is piecewise constant, which means there exist time points $0 = s_0 < s_1 < \dots < s_{m-1} < s_m = 1$ such that $\mathbf{q}|_{[s_j, s_{j+1}]} = \mathbf{q}_j \in \mathbb{R}^d$ for all $j = 0, \dots, m-1$. Since γ is increasing and onto, this gives time points $0 = t_0 < \dots < t_m = 1$ such that $\gamma(t_j) = s_j$ for all $j = 1, \dots, m$. Hence the optimization problem becomes equivalently

$$\begin{aligned} \text{Maximize} \quad & \sum_{j=0}^{m-1} \int_{t_j}^{t_{j+1}} \langle \mathbf{p}(t), \mathbf{q}_j \rangle \sqrt{\dot{\gamma}(t)} dt \\ \text{w.r.t.} \quad & \gamma : [0, 1] \rightarrow [0, 1] \text{ absolutely continuous, onto, with } \dot{\gamma} \geq 0 \text{ a.e.} \\ & \text{and } \gamma(t_j) = s_j \quad \forall j = 1, \dots, m-1. \end{aligned}$$

We can split this optimization problem into an outer maximization over t_1, \dots, t_{m-1} and an inner one, where for fixed $j = 0, \dots, m-1$, the following maximization problem needs to be solved.

$$\begin{aligned} \text{Maximize} \quad & \int_{t_j}^{t_{j+1}} \langle \mathbf{p}(t), \mathbf{q}_j \rangle \sqrt{\dot{\gamma}(t)} dt \\ \text{w.r.t.} \quad & \dot{\gamma} : [t_j, t_{j+1}] \rightarrow \mathbb{R}_0^+ \text{ and } \int_{t_j}^{t_{j+1}} \dot{\gamma}(t) dt = s_{j+1} - s_j. \end{aligned} \tag{2}$$

We obtain an upper bound for these objective functions using the Cauchy-Schwarz inequality. We have

$$\begin{aligned} \int_{t_j}^{t_{j+1}} \langle \mathbf{p}(t), \mathbf{q}_j \rangle \sqrt{\dot{\gamma}(t)} dt &\leq \int_{t_j}^{t_{j+1}} \langle \mathbf{p}(t), \mathbf{q}_j \rangle_+ \sqrt{\dot{\gamma}(t)} dt \\ &\stackrel{C.S.}{\leq} \sqrt{\int_{t_j}^{t_{j+1}} \langle \mathbf{p}(t), \mathbf{q}_j \rangle_+^2 dt} \sqrt{\int_{t_j}^{t_{j+1}} \dot{\gamma}(t) dt} \\ &= \sqrt{(s_{j+1} - s_j) \int_{t_j}^{t_{j+1}} \langle \mathbf{p}(t), \mathbf{q}_j \rangle_+^2 dt} \end{aligned} \tag{3}$$

To show this upper bound is actually the supremum over all feasible functions $\dot{\gamma}$ we consider two distinct cases.

i) If $\int_{t_j}^{t_{j+1}} \langle \mathbf{p}(t), \mathbf{q}_j \rangle_+^2 dt > 0$ we can choose

$$\dot{\gamma}(t) = \frac{(s_{j+1} - s_j) \langle \mathbf{p}(t), \mathbf{q}_j \rangle_+^2}{\int_{t_j}^{t_{j+1}} \langle \mathbf{p}(t), \mathbf{q}_j \rangle_+^2 dt}. \tag{4}$$

This choice of $\dot{\gamma}$ is feasible as it attains only non-negative values and $\int_{t_j}^{t_{j+1}} \dot{\gamma}(t) dt = s_{j+1} - s_j$ for all $j = 0, \dots, m-1$. We calculate

$$\begin{aligned} \int_{t_j}^{t_{j+1}} \langle \mathbf{p}(t), \mathbf{q}_j \rangle \sqrt{\dot{\gamma}(t)} dt &= \int_{t_j}^{t_{j+1}} \langle \mathbf{p}(t), \mathbf{q}_j \rangle \frac{\sqrt{s_{j+1} - s_j} \langle \mathbf{p}(t), \mathbf{q}_j \rangle_+}{\sqrt{\int_{t_j}^{t_{j+1}} \langle \mathbf{p}(t), \mathbf{q}_j \rangle_+^2 dt}} dt \\ &= \frac{\sqrt{s_{j+1} - s_j}}{\sqrt{\int_{t_j}^{t_{j+1}} \langle \mathbf{p}(t), \mathbf{q}_j \rangle_+^2 dt}} \int_{t_j}^{t_{j+1}} \langle \mathbf{p}(t), \mathbf{q}_j \rangle \langle \mathbf{p}(t), \mathbf{q}_j \rangle_+ dt \\ &= \sqrt{s_{j+1} - s_j} \sqrt{\int_{t_j}^{t_{j+1}} \langle \mathbf{p}(t), \mathbf{q}_j \rangle_+^2 dt}, \end{aligned}$$

where the last equality is due to $\langle \mathbf{p}(t), \mathbf{q}_j \rangle \langle \mathbf{p}(t), \mathbf{q}_j \rangle_+ = \langle \mathbf{p}(t), \mathbf{q}_j \rangle_+^2$, since $\langle \mathbf{p}(t), \mathbf{q}_j \rangle < 0$ implies

$\langle \mathbf{p}(t), \mathbf{q}_j \rangle_+ = 0$. Hence $\dot{\gamma}$ is a maximizing function.

ii) If $\int_{t_j}^{t_{j+1}} \langle \mathbf{p}(t), \mathbf{q}_j \rangle_+^2 dt = 0$, the objective function is bounded above by 0 due to (3) and we construct a sequence $(\dot{\gamma}_k)_{k \in \mathbb{N}}$ of feasible functions to reach that upper bound. For all $k \in \mathbb{N}$ let

$$\dot{\gamma}_k = (s_{j+1} - s_j)k \mathbb{1}_{[t_j, t_j + \frac{1}{k}]} \geq 0.$$

Hence we have for sufficiently large $k \in \mathbb{N}$

$$\int_{t_j}^{t_{j+1}} \dot{\gamma}_k(t) dt = (s_{j+1} - s_j) \int_{t_j}^{t_j + \frac{1}{k}} k dt = s_{j+1} - s_j,$$

which shows that the functions $\dot{\gamma}_k$ are feasible for $k \geq \frac{1}{t_{j+1} - t_j}$.

Since $\|\mathbf{p}\|_\infty < \infty$ we have for sufficiently large $k \in \mathbb{N}$

$$\begin{aligned} \left| \int_{t_j}^{t_{j+1}} \langle \mathbf{p}(t), \mathbf{q}_j \rangle \sqrt{\dot{\gamma}_k(t)} dt \right| &\leq \int_{t_j}^{t_{j+1}} |\langle \mathbf{p}(t), \mathbf{q}_j \rangle| \sqrt{\dot{\gamma}_k(t)} dt \\ &\leq \int_{t_j}^{t_{j+1}} \|\mathbf{p}(t)\| \|\mathbf{q}_j\| \sqrt{\dot{\gamma}_k(t)} dt \\ &\leq \|\mathbf{p}\|_\infty \|\mathbf{q}_j\| \int_{t_j}^{t_j + \frac{1}{k}} \sqrt{(s_{j+1} - s_j)k} dt \\ &= \|\mathbf{p}\|_\infty \|\mathbf{q}_j\| \sqrt{s_{j+1} - s_j} \frac{\sqrt{k}}{k} \xrightarrow{k \rightarrow \infty} 0. \end{aligned}$$

This shows that $(\dot{\gamma}_k)$ is a maximizing sequence of warping functions since

$$0 \geq \int_{t_j}^{t_{j+1}} \langle \mathbf{p}(t), \mathbf{q}_j \rangle \sqrt{\dot{\gamma}_k(t)} dt \xrightarrow{k \rightarrow \infty} 0.$$

In this case, we do not find a maximizing warping function γ but a sequence of maximizing warping functions γ_k .

In both cases i) and ii), the inner optimization (2) takes the value $\sqrt{(s_{j+1} - s_j) \int_{t_j}^{t_{j+1}} \langle \mathbf{p}(t), \mathbf{q}_j \rangle_+^2 dt}$ for given $j = 0, \dots, m-1$. The overall optimization thus becomes the outer optimization over the sum of these terms with respect to t_1, \dots, t_{m-1} , i.e. takes the form (3). \square

B.2 Gradient of the objective function function in Lemma 1

The simplified objective function function given in Lemma 1,

$$\Phi(\mathbf{t}) = \Phi(t_1, \dots, t_{m-1}) = \sum_{j=0}^{m-1} \sqrt{(s_{j+1} - s_j) \int_{t_j}^{t_{j+1}} \langle \mathbf{p}(t), \mathbf{q}_j \rangle_+^2 dt}$$

is differentiable if \mathbf{p} is at least continuous. In this case the partial derivatives can be computed as

$$\begin{aligned}
\frac{\partial}{\partial t_j} \Phi(\mathbf{t}) &= \frac{\partial}{\partial t_j} \sum_{k=0}^{m-1} \sqrt{(s_{k+1} - s_k) \int_{t_k}^{t_{k+1}} \langle \mathbf{p}(t), \mathbf{q}_k \rangle_+^2 dt} \\
&= \frac{\partial}{\partial t_j} \sqrt{(s_j - s_{j-1}) \int_{t_{j-1}}^{t_j} \langle \mathbf{p}(t), \mathbf{q}_{j-1} \rangle_+^2 dt} + \frac{\partial}{\partial t_j} \sqrt{(s_{j+1} - s_j) \int_{t_j}^{t_{j+1}} \langle \mathbf{p}(t), \mathbf{q}_j \rangle_+^2 dt} \\
&= \frac{\frac{1}{2}(s_j - s_{j-1}) \langle \mathbf{p}(t_j), \mathbf{q}_{j-1} \rangle_+^2}{\sqrt{(s_j - s_{j-1}) \int_{t_{j-1}}^{t_j} \langle \mathbf{p}(t), \mathbf{q}_{j-1} \rangle_+^2 dt}} - \frac{\frac{1}{2}(s_{j+1} - s_j) \langle \mathbf{p}(t_j), \mathbf{q}_j \rangle_+^2}{\sqrt{(s_{j+1} - s_j) \int_{t_j}^{t_{j+1}} \langle \mathbf{p}(t), \mathbf{q}_j \rangle_+^2 dt}} \\
&= \frac{1}{2} \left(\frac{\sqrt{s_j - s_{j-1}} \langle \mathbf{p}(t_j), \mathbf{q}_{j-1} \rangle_+^2}{\sqrt{\int_{t_{j-1}}^{t_j} \langle \mathbf{p}(t), \mathbf{q}_{j-1} \rangle_+^2 dt}} - \frac{\sqrt{s_{j+1} - s_j} \langle \mathbf{p}(t_j), \mathbf{q}_j \rangle_+^2}{\sqrt{\int_{t_j}^{t_{j+1}} \langle \mathbf{p}(t), \mathbf{q}_j \rangle_+^2 dt}} \right)
\end{aligned}$$

for all $j = 1, \dots, m-1$. If \mathbf{p} is piecewise linear, $t \mapsto \langle \mathbf{p}(t), \mathbf{q}_j \rangle_+^2$ is piecewise quadratic and one can compute the integral in the denominator exactly.

B.3 Closed form solution for the coordinate wise maximization

For fixed $j \in \{1, \dots, m-1\}$ and fixed $0 = t_0 \leq \dots \leq t_{j-1} \leq t_{j+1} \leq \dots \leq t_m = 1$ we need to solve

$$\begin{aligned}
\text{Maximize } L(t_j) &= \sum_{k=j}^{j+1} \sqrt{(s_k - s_{k-1}) \int_{t_{k-1}}^{t_k} \langle \mathbf{p}(t), \mathbf{q}_{k-1} \rangle_+^2 dt} \\
\text{w.r.t } &t_{j-1} \leq t_j \leq t_{j+1}.
\end{aligned} \tag{5}$$

Since \mathbf{p} is assumed to be piecewise constant on $[t_{j-1}, t_{j+1}]$ there exists $t_{j-1} = r_0 < \dots < r_l = t_{j+1}$ such that $\mathbf{p}|_{[r_\ell, r_{\ell+1}]} = \mathbf{p}_\ell \in \mathbb{R}^d$ for all $\ell = 0, \dots, l-1$. Hence the objective function restricted to $[r_\ell, r_{\ell+1}]$ can be written as

$$\begin{aligned}
L|_{[r_\ell, r_{\ell+1}]}(t_j) &= \sqrt{(s_j - s_{j-1}) \left((t_j - r_\ell) \langle \mathbf{p}_\ell, \mathbf{q}_{j-1} \rangle_+^2 + \sum_{k=0}^{\ell-1} (r_{k+1} - r_k) \langle \mathbf{p}_k, \mathbf{q}_{j-1} \rangle_+^2 \right)} \\
&\quad + \sqrt{(s_{j+1} - s_j) \left((r_{\ell+1} - t_j) \langle \mathbf{p}_\ell, \mathbf{q}_j \rangle_+^2 + \sum_{k=\ell+1}^{l-1} (r_{k+1} - r_k) \langle \mathbf{p}_k, \mathbf{q}_j \rangle_+^2 \right)}.
\end{aligned}$$

This shows that for all $\ell = 0, \dots, l-1$ there are constant values

$$\begin{aligned}
A_{\ell 1} &= (s_j - s_{j-1}) \langle \mathbf{p}_\ell, \mathbf{q}_{j-1} \rangle_+^2 \\
A_{\ell 2} &= (s_{j+1} - s_j) \langle \mathbf{p}_\ell, \mathbf{q}_j \rangle_+^2 \\
B_{\ell 1} &= (s_j - s_{j-1}) \left(r_\ell \langle \mathbf{p}_\ell, \mathbf{q}_{j-1} \rangle_+^2 - \sum_{k=0}^{\ell-1} (r_{k+1} - r_k) \langle \mathbf{p}_k, \mathbf{q}_{j-1} \rangle_+^2 \right) \\
B_{\ell 2} &= (s_{j+1} - s_j) \left(r_{\ell+1} \langle \mathbf{p}_\ell, \mathbf{q}_j \rangle_+^2 + \sum_{k=\ell+1}^{l-1} (r_{k+1} - r_k) \langle \mathbf{p}_k, \mathbf{q}_j \rangle_+^2 \right)
\end{aligned}$$

such that

$$L|_{[r_\ell, r_{\ell+1}]}(t_j) = \sqrt{A_{\ell 1} t_j - B_{\ell 1}} + \sqrt{B_{\ell 2} - A_{\ell 2} t_j}$$

with $A_{l1}t_j - B_{l1} \geq 0$ and $B_{l2} - A_{l2}t_j \geq 0$ for all $t_j \in [r_l, r_{l+1}]$. Without loss of generality we assume $A_{l1}, A_{l2} > 0$ since otherwise the objective function is monotonic, hence attains its maximum on the boundary. This case can be included separately below. Thus $L|_{[r_l, r_{l+1}]}$ is twice continuously differentiable on $]r_l, r_{l+1}[$ with

$$\begin{aligned}\frac{\partial}{\partial t_j} L|_{[r_l, r_{l+1}]}(t_j) &= \frac{1}{2} \left(\frac{A_{l1}}{\sqrt{A_{l1}t_j - B_{l1}}} - \frac{A_{l2}}{\sqrt{B_{l2} - A_{l2}t_j}} \right), \\ \frac{\partial^2}{\partial t_j^2} L|_{[r_l, r_{l+1}]}(t_j) &= -\frac{1}{4} \left(\frac{A_{l1}^2}{\sqrt{A_{l1}t_j - B_{l1}}^3} + \frac{A_{l2}^2}{\sqrt{B_{l2} - A_{l2}t_j}^3} \right) < 0.\end{aligned}$$

Therefore, every maximizer t_j within $]r_l, r_{l+1}[$ fulfills

$$\begin{aligned}\frac{A_{l1}}{\sqrt{A_{l1}t_j - B_{l1}}} &= \frac{A_{l2}}{\sqrt{B_{l2} - A_{l2}t_j}} \quad \Leftrightarrow \quad A_{l1}^2(B_{l2} - A_{l2}t_j) = A_{l2}^2(A_{l1}t_j - B_{l1}) \\ &\Leftrightarrow \quad t_j = \frac{A_{l1}^2 B_{l2} + A_{l2}^2 B_{l1}}{A_{l1} A_{l2}^2 + A_{l1}^2 A_{l2}}.\end{aligned}$$

We conclude that every solution to the coordinate wise maximization problem (5) is contained in the set

$$\bigcup_{l=0}^l \{r_l\} \cup \bigcup_{l=0}^{l-1} \left\{ \frac{A_{l1}^2 B_{l2} + A_{l2}^2 B_{l1}}{A_{l1} A_{l2}^2 + A_{l1}^2 A_{l2}} \right\}$$

and can compare function values of L over this set to find the maximizer.

B.4 Proof of Theorem 1

Proof. Let Φ be defined as in Equation (3),

$$\Phi(\mathbf{t}) = \Phi(t_1, \dots, t_{m-1}) = \sum_{j=0}^{m-1} \sqrt{(s_{j+1} - s_j) \int_{t_j}^{t_{j+1}} \langle \mathbf{p}(t), \mathbf{q}_j \rangle_+^2 dt},$$

with \mathbf{p} being piecewise constant. Furthermore let $(\mathbf{t}^{(l)})_{l \in \mathbb{N}} = t^{(1)}, t^{(2)}, \dots$ be a sequence resulting from Algorithm 1 and \mathbf{t}^* an accumulation point of $(\mathbf{t}^{(l)})_{l \in \mathbb{N}}$.

We proof this main result in three steps. First, we show that the accumulation point $\mathbf{t}^* = (t_1^*, \dots, t_{m-1}^*)$ is a maximizer of Φ restricted to coordinate directions. Then we conclude that Φ is semi-differentiable at \mathbf{t}^* for every direction $\mathbf{u} \in \mathbb{R}^{m-1}$. Last we use Lemma B.1 below, which establishes local concavity of the objective function, to see that \mathbf{t}^* is a local maximum of Φ .

Since \mathbf{t}^* is an accumulation point, there is a subsequence $(\mathbf{t}^{(l_k)})_{k \in \mathbb{N}}$ with $\lim_{k \rightarrow \infty} \mathbf{t}^{(l_k)} = \mathbf{t}^*$. Denote by

$$\begin{aligned}\Phi_{\text{odd}}^{(k)} &:= \Phi|_{\{t_j = t_j^{(l_k)}, j \text{ even}\}} \\ \Phi_{\text{even}}^{(k)} &:= \Phi|_{\{t_j = t_j^{(l_k)}, j \text{ odd}\}}\end{aligned}$$

the restrictions of Φ at the current sequence value with either fixed odd or even coordinate entries. Φ is contin-

uous, hence we have pointwise limits

$$\begin{aligned}\lim_{k \rightarrow \infty} \Phi_{odd}^{(k)} &= \Phi|_{\{t_j = t_j^*, j \text{ even}\}} =: \Phi_{odd}^*, \\ \lim_{k \rightarrow \infty} \Phi_{even}^{(k)} &= \Phi|_{\{t_j = t_j^*, j \text{ odd}\}} =: \Phi_{even}^*,\end{aligned}$$

with $\Phi_{odd}^*, \Phi_{even}^*$ being the restrictions to odd and even coordinate directions at the accumulation point \mathbf{t}^* . Since at each step we either update all odd or all even entries, $\Phi_{odd}^{(k)}$ and $\Phi_{even}^{(k)}$ attain their maximum at either the current or the next sequence value. That is

$$\left\| \Phi_{odd}^{(k)} \right\|_{\infty}, \left\| \Phi_{even}^{(k)} \right\|_{\infty} \in \{ \Phi(\mathbf{t}^{(t_k)}), \Phi(\mathbf{t}^{(t_{k+1})}) \}$$

for all $k \in \mathbb{N}$. Thus, Φ_{odd}^* and Φ_{even}^* are bounded as well:

$$\left\| \Phi_{odd}^* \right\|_{\infty} = \lim_{k \rightarrow \infty} \left\| \Phi_{odd}^{(k)} \right\|_{\infty} \leq \lim_{k \rightarrow \infty} \Phi(\mathbf{t}^{(t_{k+1})}) = \Phi(\mathbf{t}^*),$$

since $\lim_{\iota \rightarrow \infty} \Phi(\mathbf{t}^{(t_{\iota})}) = \Phi(\mathbf{t}^*)$. We can conclude this as coordinate-wise maximization produces a monotonically increasing sequence $\Phi(\mathbf{t}^{(t_{\iota+1})}) \geq \Phi(\mathbf{t}^{(t_{\iota})})$ for all $\iota \in \mathbb{N}$ and the subsequence $\Phi(\mathbf{t}^{(t_{k})})$ converges to $\Phi(\mathbf{t}^*)$ due to Φ being continuous, which implies the whole sequence converges. Analogously we have $\left\| \Phi_{even}^* \right\|_{\infty} \leq \Phi(\mathbf{t}^*)$, hence \mathbf{t}^* is a maximizer of Φ restricted to any coordinate direction (i.e. t_j^* maximizes $\Phi(t_1^*, \dots, t_j, \dots, t_{m-1}^*)$ over t_j for all $j = 1, \dots, m-1$).

To show that this implies that Φ is partially semi-differentiable at \mathbf{t}^* , first note that Φ is partially semi-differentiable at every point $\mathbf{t} = (t_1, \dots, t_{m-1})$ with $(s_{j+1} - s_j) \int_{t_j}^{t_{j+1}} \langle \mathbf{p}(t), \mathbf{q}_j \rangle_+^2 dt > 0$ for all $j = 1, \dots, m-1$, since the square-root function is differentiable for strictly positive values and $\int_{t_j}^{t_{j+1}} \langle \mathbf{p}(t), \mathbf{q}_j \rangle_+^2 dt$ is piecewise linear, thus semi-differentiable.

Assume there is a $j \in \{1, \dots, m-1\}$ with $(s_{j+1} - s_j) \int_{t_j^*}^{t_{j+1}^*} \langle \mathbf{p}(t), \mathbf{q}_j \rangle_+^2 dt = 0$. We show that Φ is still partially semi-differentiable at \mathbf{t}^* in direction t_j . A similar argument shows differentiability in direction t_{j+1} .

Let L be the relevant part of the objective function Φ in direction t_j .

$$L(t_j) = \sqrt{(s_j - s_{j-1}) \int_{t_{j-1}^*}^{t_j} \langle \mathbf{p}(t), \mathbf{q}_{j-1} \rangle_+^2 dt} + \sqrt{(s_{j+1} - s_j) \int_{t_j}^{t_{j+1}^*} \langle \mathbf{p}(t), \mathbf{q}_j \rangle_+^2 dt}$$

We need to show that both, left and right derivatives of L at t_j^* exist.

- If $(s_j - s_{j-1}) \int_{t_{j-1}^*}^{t_j^*} \langle \mathbf{p}(t), \mathbf{q}_{j-1} \rangle_+^2 dt = 0$, we have $L(t_j^*) = 0$. This implies $L(t_j) = 0$ for all $t_j \in [t_{j-1}^*, t_{j+1}^*]$ since t_j^* is a maximizer (in t_j coordinate direction) and L is non-negative. Therefore $L = 0$ which means L is differentiable on its whole domain.
- If $(s_j - s_{j-1}) \int_{t_{j-1}^*}^{t_j^*} \langle \mathbf{p}(t), \mathbf{q}_{j-1} \rangle_+^2 dt > 0$, the left term of L is strictly positive in a neighborhood of t_j^* and consequently semi-differentiable in a neighborhood of t_j^* . The right term $H(t_j) := \sqrt{(s_{j+1} - s_j) \int_{t_j}^{t_{j+1}^*} \langle \mathbf{p}(t), \mathbf{q}_j \rangle_+^2 dt}$ is differentiable at t_j^* since it is 0 in a neighborhood of t_j^* . This is due to $t_j \mapsto (s_{j+1} - s_j) \int_{t_j}^{t_{j+1}^*} \langle \mathbf{p}(t), \mathbf{q}_j \rangle_+^2 dt$ being piecewise linear, non-negative and monotonically decreasing. Since it attains 0 at t_j^* , it is also 0 in a right neighborhood of t_j^* . If H were strictly positive in a neighborhood left of t_j^* , its left derivative would tend to $-\infty$ at t_j^* as $H(t_j^*) = 0$ and the derivative of the square-root tends to ∞ for values tending linearly to

0. But $\frac{\partial_-}{\partial t_j} H(t_j^*) = -\infty$ would imply $\frac{\partial_-}{\partial t_j} L(t_j^*) = -\infty$, which contradicts t_j^* being a maximizer.

Taking all those cases into account we conclude that Φ is partially semi-differentiable at the accumulation point \mathbf{t}^* produced by coordinate-wise maximization. Since we already know that t_j^* is the coordinate-wise maximizer of Φ for all $j = 1, \dots, m-1$ in coordinate directions, the left-sided partial derivatives need to be non-negative, the right-sided partial derivatives non-positive.

To show that this implies that \mathbf{t}^* is a local maximizer, consider sets $U = \times_{j=1}^{m-1} U_j \cap \{0 \leq t_1 \leq \dots \leq t_{m-1} \leq 1\}$ such that $\mathbf{t}^* \in U$ and \mathbf{p} is constant on the interior of the interval $U_j \neq \emptyset$ for all $j = 1, \dots, m-1$.

We prove that \mathbf{t}^* is the maximizer of $\Phi|_U$ by contradiction. Assume there is a $\mathbf{u} \in U$ such that $\Phi(\mathbf{u}) > \Phi(\mathbf{t}^*)$. Let $\alpha(s) = s\mathbf{u} + (1-s)\mathbf{t}^*$ for all $s \in [0, 1]$. Since the square-root is improperly differentiable on $[0, \infty[$, with the derivative at 0 being ∞ , this implies that $\Phi \circ \alpha$ is improperly differentiable on $[0, 1]$ with

$$(\Phi \circ \alpha)'(s) = \left\langle \frac{\partial \Phi}{\partial \mathbf{t}}(\alpha(s)), \mathbf{u} - \mathbf{t}^* \right\rangle = \sum_{j=1}^{m-1} (u_j - t_j^*) \frac{\partial \Phi}{\partial t_j}(\alpha(s)).$$

Considering the limit $s \searrow 0$ yields

$$\lim_{s \searrow 0} \frac{\partial \Phi}{\partial t_j}(\alpha(s)) = \begin{cases} \frac{\partial_+ \Phi}{\partial t_j}(t_j^*) \leq 0 & \text{if } u_j - t_j^* > 0, \\ \frac{\partial_- \Phi}{\partial t_j}(t_j^*) \geq 0 & \text{if } u_j - t_j^* < 0. \end{cases}$$

Hence the right-sided derivative will be attained if $u_j - t_j^*$ is positive and the left-sided derivative if $u_j - t_j^*$ is negative. This implies $(u_j - t_j^*) \lim_{s \searrow 0} \frac{\partial \Phi}{\partial t_j}(\alpha(s)) \leq 0$ for all $j = 1, \dots, m-1$ and therefore,

$$(\Phi \circ \alpha)'(0) = \lim_{s \searrow 0} (\Phi \circ \alpha)'(s) = \sum_{j=1}^{m-1} (u_j - t_j^*) \lim_{s \searrow 0} \frac{\partial \Phi}{\partial t_j}(\alpha(s)) \leq 0.$$

But since U is a convex set, Φ is concave on the interior of U (see Lemma B.1) and therefore, as Φ is continuous, it is concave on U . We compute

$$\begin{aligned} (\Phi \circ \alpha)'(0) &= \lim_{s \searrow 0} \frac{(\Phi \circ \alpha)(s) - (\Phi \circ \alpha)(0)}{s} \\ &= \lim_{s \searrow 0} \frac{\Phi(s\mathbf{u} + (1-s)\mathbf{t}^*) - \Phi(\mathbf{t}^*)}{s} \\ &\geq \lim_{s \searrow 0} \frac{s\Phi(\mathbf{u}) + (1-s)\Phi(\mathbf{t}^*) - \Phi(\mathbf{t}^*)}{s} \\ &= \Phi(\mathbf{u}) - \Phi(\mathbf{t}^*) > 0, \end{aligned}$$

which contradicts $(\Phi \circ \alpha)'(0) \leq 0$.

Thus, \mathbf{t}^* is a maximum of $\Phi|_U$. This means it is a maximum on the union of such U 's, whose interior is a relatively open neighbourhood of \mathbf{t}^* with respect to the relative topology on $\{0 \leq t_1 \leq \dots \leq t_{m-1} \leq 1\}$. Hence \mathbf{t}^* is a local maximizer of Φ . \square

Lemma B.1. *Let Φ be the objective function defined in Equation (3), \mathbf{p} piecewise constant and $U \subset \mathbb{R}^{m-1}$ a convex set such that $\mathbf{p}(t_j)$ is constant for all $j = 1, \dots, m$ and all $(t_1, \dots, t_{m-1}) \in U$. Then $\Phi|_U$ is concave.*

Proof. Note that Φ is twice continuously differentiable on the interior $\overset{\circ}{U}$ of U . We show that all second directional derivatives $\partial_{\mathbf{u}\mathbf{u}}^2 \Phi$ are non positive. This implies the Hessian H is negative semi-definite, since $\mathbf{u}^T H \mathbf{u} = \partial_{\mathbf{u}\mathbf{u}}^2 \Phi$ for all $\mathbf{u} \in \mathbb{R}^{m-1}$. Hence $\Phi|_U$ is concave.

To show that the second derivative at $\mathbf{t} = (t_1, \dots, t_{m-1}) \in \mathring{U}$ is non-positive in any direction, let $\alpha \in \mathbb{R}$ and $\mathbf{u} = (u_1, \dots, u_{m-1}) \in \mathbb{R}^{m-1}$. Define

$$Q_j(\alpha) = (s_{j+1} - s_j) \int_{t_j + \alpha u_j}^{t_{j+1} + \alpha u_{j+1}} \langle \mathbf{p}(t), \mathbf{q}_j \rangle_+^2 dt.$$

Q_j is linear around $\alpha = 0$ and therefore differentiable with constant derivative $Q'_j(\alpha) =: c_j \in \mathbb{R}$. If $Q_j(0) \neq 0$ for all $j \in 1, \dots, m-1$ we compute the directional derivative of the objective function as

$$\partial_{\mathbf{u}} \Phi(\mathbf{t}) = \frac{\partial}{\partial \alpha} \sum_{j=1}^{m-1} \sqrt{Q_j(\alpha)} \Big|_{\alpha=0} = \frac{1}{2} \sum_{j=1}^{m-1} \frac{Q'_j(\alpha)}{\sqrt{Q_j(\alpha)}} \Big|_{\alpha=0} = \frac{1}{2} \sum_{j=1}^{m-1} \frac{c_j}{\sqrt{Q_j(\alpha)}} \Big|_{\alpha=0},$$

and the second derivative becomes

$$\partial_{\mathbf{u}\mathbf{u}}^2 \Phi(\mathbf{t}) = \frac{\partial^2}{\partial \alpha^2} \sum_{j=1}^{m-1} \sqrt{Q_j(\alpha)} \Big|_{\alpha=0} = -\frac{1}{4} \sum_{j=1}^{m-1} \frac{c_j^2}{\sqrt{Q_j(\alpha)}^3} \Big|_{\alpha=0} \leq 0.$$

If $Q_j(0) = 0$ for some $j = \dots, m-1$, we have in particular $\langle \mathbf{p}(t_j), \mathbf{q}_j \rangle_+^2 = 0$ and $\langle \mathbf{p}(t_{j+1}), \mathbf{q}_j \rangle_+^2 = 0$, which means Q_j is zero in a neighborhood of $\alpha = 0$. Hence the second derivative of $\sqrt{Q_j(\alpha)}$ is zero as well and does not contribute to the sum. \square

B.5 Integral approximation for SRV-spline mean computation

In the L_2 -step of Algorithm 2, the integrals

$$\left\| \bar{\mathbf{p}} - (\mathbf{q}_i \circ \gamma_i) \sqrt{\dot{\gamma}_i} \right\|_{L_2}^2 = \int_0^1 \left\| \bar{\mathbf{p}}(t) - (\mathbf{q}_i(t) \circ \gamma_i(t)) \sqrt{\dot{\gamma}_i(t)} \right\|^2 dt \quad (6)$$

in the sum need to be approximated, since the curves β_i are only observed on a finite grid $0 = s_{i,0} \leq s_{i,1} \leq \dots \leq s_{i,m_i} = 1$, which means the SRV-curves $\mathbf{q}_1, \dots, \mathbf{q}_n$ are unobserved. One option is to assume that the SRVs \mathbf{q}_i of the observed curves are piecewise constant, like we do in the warping step. Since $\bar{\mathbf{p}}$ is piecewise linear (or even piecewise constant), $(\mathbf{q}_i \circ \gamma_i) \sqrt{\dot{\gamma}_i}$ will be piecewise linear as well (see proof of Lemma 1 in Appendix B), which leads to a closed form solution of the integral. If we use this approximation of the integral, the resulting mean tends to overfit the edges of the observed polygons (e.g. the dashed mean in Fig. 2 on the right).

Alternatively, we derive an approximation of the integrals in the L_2 fitting step of Algorithm 2 using the mean value theorem and the monotonicity of the warping. For all $j = 0, \dots, m_i - 1$, there is a $t_{i,j} \in [\gamma_i^{-1}(s_{i,j}), \gamma_i^{-1}(s_{i,j+1})]$ with $(\beta_i \circ \gamma_i)'(t_{i,j}) = \frac{\beta_i(s_{i,j+1}) - \beta_i(s_{i,j})}{\gamma_i^{-1}(s_{i,j+1}) - \gamma_i^{-1}(s_{i,j})}$ and therefore

$$\begin{aligned} (\mathbf{q}_i \circ \gamma_i(t_{i,j})) \sqrt{\dot{\gamma}_i(t_{i,j})} &= \frac{(\beta_i \circ \gamma_i)'(t_{i,j})}{\sqrt{\|(\beta_i \circ \gamma_i)'(t_{i,j})\|}} \\ &= \frac{\beta_i(s_{i,j+1}) - \beta_i(s_{i,j})}{\sqrt{\|\beta_i(s_{i,j+1}) - \beta_i(s_{i,j})\|} \sqrt{\gamma_i^{-1}(s_{i,j+1}) - \gamma_i^{-1}(s_{i,j})}}. \end{aligned}$$

While this is exact for unknown points $t_{i,j}$, we use an approximation by assuming this mean value of the derivative $(\beta_i \circ \gamma_i)'$ is attained in the middle of the interval $[\gamma_i^{-1}(s_{i,j}), \gamma_i^{-1}(s_{i,j+1})]$; hence we approximate $t_{i,j} \approx \frac{\gamma_i^{-1}(s_{i,j+1}) + \gamma_i^{-1}(s_{i,j})}{2}$ for all $j = 0, \dots, m_i - 1$. Thus, for $i = 1, \dots, n$, the integral in (6) is replaced by the weighted sum $\sum_{j=0}^{m_i-1} \omega_{i,j} \left\| \bar{\mathbf{p}}(t_{i,j}) - (\mathbf{q}_i \circ \gamma_i(t_{i,j})) \sqrt{\dot{\gamma}_i(t_{i,j})} \right\|^2$. This leaves us with a quadratic minimization

problem w.r.t. the spline coefficients in $\bar{\mathbf{p}}$, for which we compute the solution analytically as a generalized least squares estimate.

There are different options to choose the weights $\omega_{i,j}$ in this integral approximation. The weights $\omega_{i,j} = \gamma_i^{-1}(s_{i,j+1}) - \gamma_i^{-1}(s_{i,j})$ based on the trapezoidal rule for numerical integration give equal importance to each of the observed curves, independent of the number of points m_i observed on each of them. An alternative choice of $\omega_{i,j} = 1$ puts more weight on single observations on a specific curve. Consequently, curves or parts of curves with more observations have higher influence on the estimated mean than curves or parts of curves with fewer observations. The difference between this approximation (with $\omega_{i,j} = 1$) and the one based on assuming observed polygons also for the L_2 spline fitting step is displayed in Fig. 2 on the right. In this example, the estimated mean based on this discrete integral approximation (solid line) is closer to a proper spiral shape. In the following we will use weights $\omega_{i,j} = 1$ unless stated otherwise.

B.6 Proof of Theorem 2

Proof. Let $\mathbf{Q} = (Q_1, Q_2, \dots, Q_d)$. Without loss of generality we assume $d = 2$. For $d > 2$ perform a coordinate transformation such that (Q_1, Q_2) has a non-linear image between its knots and consider the first two coordinates.

Hence we assume $\mathbf{P} = \mathbf{Q} \circ \gamma$ with $\deg(\mathbf{P}), \deg(\mathbf{Q}) \in \{2, 3\}$ and \mathbf{Q} has non-linear image between its knots. First, we show that γ is piecewise polynomial, which implies γ is piecewise linear since $\deg(\gamma) \geq 2$ would imply $\deg(\mathbf{P}) = \deg(\mathbf{Q} \circ \gamma) \geq 4$.

Let $I \subseteq [0, 1]$ be an interval such that $\mathbf{P}|_I$ and $\mathbf{Q}|_{\gamma(I)}$ are polynomials of degree $\in \{2, 3\}$. That means we can denote

$$\begin{aligned} \mathbf{P}(t) &= \begin{pmatrix} P_1(t) \\ P_2(t) \end{pmatrix} = \begin{pmatrix} p_{10} + p_{11}t + p_{12}t^2 + p_{13}t^3 \\ p_{20} + p_{21}t + p_{22}t^2 + p_{23}t^3 \end{pmatrix} && \text{for all } t \in I, \\ \mathbf{Q}(t) &= \begin{pmatrix} Q_1(t) \\ Q_2(t) \end{pmatrix} = \begin{pmatrix} q_{10} + q_{11}t + q_{12}t^2 + q_{13}t^3 \\ q_{20} + q_{21}t + q_{22}t^2 + q_{23}t^3 \end{pmatrix} && \text{for all } t \in \gamma(I). \end{aligned}$$

We compute

$$\begin{aligned} q_{13}P_2(t) - q_{23}P_1(t) &= q_{13}Q_2(\gamma(t)) - q_{23}Q_1(\gamma(t)) \\ &= q_{13}q_{20} - q_{23}q_{10} + (q_{13}q_{21} - q_{23}q_{11})\gamma(t) + (q_{13}q_{22} - q_{23}q_{12})\gamma(t)^2. \end{aligned} \quad (7)$$

Note that either $\begin{vmatrix} q_{13} & q_{12} \\ q_{23} & q_{22} \end{vmatrix} = q_{13}q_{22} - q_{23}q_{12} \neq 0$ or $\begin{vmatrix} q_{13} & q_{11} \\ q_{23} & q_{21} \end{vmatrix} = q_{13}q_{21} - q_{23}q_{11} \neq 0$, because otherwise $\begin{pmatrix} q_{12} \\ q_{22} \end{pmatrix}$ and $\begin{pmatrix} q_{11} \\ q_{21} \end{pmatrix}$ are multiples of $\begin{pmatrix} q_{13} \\ q_{23} \end{pmatrix}$, which means \mathbf{Q} has a linear image on $\gamma(I)$. Thus we need to consider two cases.

i) If $q_{13}q_{22} - q_{23}q_{12} = 0$, this implies $(q_{13}q_{21} - q_{23}q_{11}) \neq 0$ and the claim follows via solving Equation (7) for $\gamma(t)$.

ii) If $c_1 := q_{13}q_{22} - q_{23}q_{12} \neq 0$ there exists a polynomial \tilde{P}_1 with $\deg(\tilde{P}_1) \leq 3$ and a constant $c_2 \in \mathbb{R}$ such

that $\gamma(t) = \sqrt{\tilde{P}_1} + c_2$ (derive this from Equation (7) by completing the square). Thus we observe that

$$\begin{aligned}
q_{12}P_2(t) - q_{22}P_1(t) &= q_{12}Q_2(\gamma(t)) - q_{22}Q_1(\gamma(t)) \\
&= q_{12}(q_{20} + q_{21}\gamma(t) + q_{23}\gamma(t)^3) - q_{22}(q_{10} + q_{11}\gamma(t) + q_{13}\gamma(t)^3) \\
&= q_{12}q_{20} - q_{22}q_{10} + (q_{12}q_{21} - q_{22}q_{11})(\sqrt{\tilde{P}_1} + c_2) - c_1 \left(\sqrt{\tilde{P}_1} + c_2 \right)^3 \\
&= c_3 + c_4c_2 + c_4\sqrt{\tilde{P}_1} - c_1 \left(\tilde{P}_1\sqrt{\tilde{P}_1} + 3c_2\tilde{P}_1 + 3c_2^2\sqrt{\tilde{P}_1} + c_2^3 \right) \\
&= c_3 + c_4c_2 - c_1(c_2^3 + 3c_2\tilde{P}_1) + (c_4 - c_1(3c_2^2 + \tilde{P}_1))\sqrt{\tilde{P}_1}
\end{aligned}$$

with additional constants $c_3 := q_{12}q_{20} - q_{22}q_{10}$ and $c_4 := q_{12}q_{21} - q_{22}q_{11}$. Thus,

$$\left(q_{12}P_2(t) - q_{22}P_1(t) - c_3 - c_4c_2 + c_1(c_2^3 + 3c_2\tilde{P}_1) \right)^2 = \left(c_4 - c_1(3c_2^2 + \tilde{P}_1) \right)^2 \tilde{P}_1,$$

which shows that either $c_4 - c_1(3c_2^2 + \tilde{P}_1) = 0$, which implies \tilde{P}_1 is constant (since $c_1 \neq 0$), or every (complex) root of \tilde{P}_1 has even multiplicity, which implies that $\sqrt{\tilde{P}_1}$ and therefore $\gamma(t) = \sqrt{\tilde{P}_1} + c_2$ are polynomial.

Together this shows that γ is polynomial and therefore linear on I . Hence $\gamma : [0, 1] \rightarrow [0, 1]$ is piecewise linear, that means γ is differentiable everywhere but at a finite number of breakpoints $0 = t_0 < t_1 < \dots < t_m = 1$. Thus, the k -th derivative of \mathbf{P} , $k < \deg(\mathbf{P})$, can be computed as

$$\frac{d^k}{dt^k} \mathbf{P}(t) = \frac{d^k}{dt^k} (\mathbf{Q} \circ \gamma)(t) = \left(\left(\frac{d^k}{dt^k} \mathbf{Q} \right) (\gamma(t)) \right) \left(\frac{d}{dt} \gamma(t) \right)^k,$$

since $\frac{d}{dt} \gamma(t)$ is piecewise constant. Assume $\gamma(t)$ is not differentiable at t_j , $j = 1, \dots, m-1$. Hence the (weak) derivative $\frac{d}{dt} \gamma(t)$ is not continuous at t_j and we need to have

$$\mathbf{Q}^{(k)}(\gamma(t_j)) = \left(\frac{d^k}{dt^k} \mathbf{Q} \right) (\gamma(t_j)) = 0 \quad \text{for all } k < \deg(\mathbf{P}),$$

since \mathbf{P} is $(\deg(\mathbf{P}) - 1)$ -times continuously differentiable on $[0, 1]$. Using a Taylor expansion of \mathbf{Q} around $\gamma(t_j)$, which is identical to \mathbf{Q} on $[\gamma(t_j), \gamma(t_{j+1})]$ since \mathbf{Q} is piecewise polynomial, we obtain:

$$\mathbf{Q}(s) = \left(\frac{Q_1^{(l)}(\gamma(t_j))}{l!} (s - \gamma(t_j))^l + Q_1(\gamma(t_j)) \right) = \left(\frac{Q_1^{(l)}(\gamma(t_j))}{Q_2^{(l)}(\gamma(t_j))} \right) (s - \gamma(t_j))^l + \left(\frac{Q_1(\gamma(t_j))}{Q_2(\gamma(t_j))} \right)$$

for all $s \in [\gamma(t_j), \gamma(t_{j+1})]$. Here we denote $l = \deg(\mathbf{P})$. This would mean that \mathbf{Q} has a linear image between $\gamma(t_j)$ and $\gamma(t_{j+1})$ in this case, which contradicts the assumptions. Hence γ needs to be differentiable on $[0, 1]$, which implies γ is linear. Since it is monotonically increasing and onto we conclude $\gamma = id$. \square

B.7 Proof of Corollary 1

Proof. Let \mathbf{q}_1^2 and \mathbf{q}_2^2 the component-wise squares of \mathbf{q}_1 and \mathbf{q}_2 , respectively. We compute

$$\mathbf{P}(s) := \int_0^s \mathbf{q}_2^2(t) dt = \int_0^s \mathbf{q}_1^2(\gamma(t)) \dot{\gamma}(t) dt = \int_0^{\gamma(s)} \mathbf{q}_1^2(t') dt' =: \mathbf{Q}(\gamma(s))$$

for all $s \in [0, 1]$ via substituting $\gamma(t) \mapsto t'$. Here we have cubic splines \mathbf{P} and \mathbf{Q} on both sides. Hence we deduce $\gamma = id$ by Theorem 2 and consequently $\mathbf{q}_2 = \mathbf{q}_1$. Note that the cubic spline curve $\mathbf{P}(s) = \int_0^s \mathbf{q}_2^2(t) dt$ is linear on any interval if and only if $\mathbf{q}_2(t)$ is constant on this interval, which is excluded by the assumptions. \square

B.8 Proof of Lemma 2

Proof. The embedding f is injective due to the previous results on identifiability (Theorem 2, Corollary 1, Remark 1) and continuous as the SRV transformation is continuous (Bruveris (2016)) and $\inf_{\gamma \in \Gamma} \|\mathbf{p} - (\mathbf{q} \circ \gamma)\sqrt{\tilde{\gamma}}\|_{L_2} \leq \|\mathbf{p} - \mathbf{q}\|_{L_2}$ for all $\mathbf{p}, \mathbf{q} \in L_2$.

The only part left to show is that f^{-1} (which exists if we restrict the co-domain of f to its image) is continuous as well. To prove this, let $(\boldsymbol{\xi}_n)_{n \in \mathbb{N}} \subseteq \Xi$ with $\boldsymbol{\beta}_n = f(\boldsymbol{\xi}_n)$ for all $n \in \mathbb{N}$ and $d(\boldsymbol{\beta}_n, \boldsymbol{\beta}) \xrightarrow{n \rightarrow \infty} 0$ for the elastic distance. Hence we have to show $\boldsymbol{\xi}_n \xrightarrow{n \rightarrow \infty} \boldsymbol{\xi}$ for $\boldsymbol{\xi} := f^{-1}(\boldsymbol{\beta})$.

Denote by \mathbf{p}_n the SRV transformation of $\boldsymbol{\beta}_n$ for all $n \in \mathbb{N}$ and by \mathbf{q} the SRV transformation of $\boldsymbol{\beta}$. Then

$$d(\boldsymbol{\beta}_n, \boldsymbol{\beta}) = \inf_{\gamma \in \Gamma} \|\mathbf{p}_n - (\mathbf{q} \circ \gamma)\sqrt{\tilde{\gamma}}\|_{L_2} \geq \inf_{\gamma \in \Gamma} \left(\|\mathbf{p}_n\|_{L_2} - \|(\mathbf{q} \circ \gamma)\sqrt{\tilde{\gamma}}\|_{L_2} \right) = \|\mathbf{p}_n\|_{L_2} - \|\mathbf{q}\|_{L_2},$$

which shows that $\|\dot{\boldsymbol{\beta}}_n\|_{L_2} = \|\mathbf{p}_n\|_{L_2}^2$ is bounded, as $d(\boldsymbol{\beta}_n, \boldsymbol{\beta})$ is bounded as a convergent sequence. Since $\|\dot{\boldsymbol{\beta}}_n\|_{L_2}$ or $\|\mathbf{p}_n\|_{L_2}$ induces a norm on Ξ , which is a subset of a finite vector space, $\|\boldsymbol{\xi}_n\|$ is bounded as well, as all norms are equivalent on finite vector spaces.

Consider an arbitrary subsequence of $(\boldsymbol{\xi}_n)_{n \in \mathbb{N}}$. Since this subsequence is bounded in $(\Xi, \|\cdot\|)$ as well, it contains a convergent subsequence $(\boldsymbol{\xi}_{n_k})_{k \in \mathbb{N}}$. Let $\boldsymbol{\xi}^* := \lim_{k \rightarrow \infty} \boldsymbol{\xi}_{n_k}$. Since the embedding f is continuous, we have $f(\boldsymbol{\xi}^*) = \lim_{k \rightarrow \infty} f(\boldsymbol{\xi}_{n_k}) = \lim_{k \rightarrow \infty} \boldsymbol{\beta}_{n_k} = \boldsymbol{\beta} = f(\boldsymbol{\xi})$ and therefore $\boldsymbol{\xi}^* = \boldsymbol{\xi}$ as f is injective. Hence, every subsequence has a subsequence which converges to $\boldsymbol{\xi}$ with respect to $\|\cdot\|$. Thus, $(\boldsymbol{\xi}_n)_{n \in \mathbb{N}}$ converges to $\boldsymbol{\xi}$ in $(\Xi, \|\cdot\|)$ and f^{-1} is hence continuous. \square

Web Appendix C Examples and counterexamples

C.1 Optimal warping and Fréchet means are not unique

We give an example that illustrates that both the optimal warping function minimizing the elastic distance and the Fréchet mean for a set of curves with respect to the elastic distance are not necessarily unique. Consider two piecewise linear curves $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ with respective piecewise constant SRV curves \mathbf{p} and \mathbf{q} given as

$$\mathbf{p}(t) = \begin{cases} (-3, 0)^T & \text{if } t \in [0, 0.25[\\ (2, -4)^T & \text{if } t \in [0.25, 0.5[\\ (-4, 2)^T & \text{if } t \in [0.5, 0.75[\\ (0, -3)^T & \text{if } t \in [0.75, 1] \end{cases} \quad \text{and} \quad \mathbf{q}(t) = \begin{cases} (-3, 1)^T & \text{if } t \in [0, 0.5[\\ (1, -3)^T & \text{if } t \in [0.5, 1] \end{cases}$$

The corresponding curves $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are displayed in Figure 2 on the left. The objective function Φ , which needs to be maximized in order to find the optimal warping of the second curve to the first, only depends on one parameter t_1 and is given as

$$\Phi(t_1) = \sqrt{0.5 \int_0^{t_1} \langle \mathbf{p}(t), \begin{pmatrix} -3 \\ 1 \end{pmatrix} \rangle_+^2 dt} + \sqrt{0.5 \int_{t_1}^1 \langle \mathbf{p}(t), \begin{pmatrix} 1 \\ -3 \end{pmatrix} \rangle_+^2 dt},$$

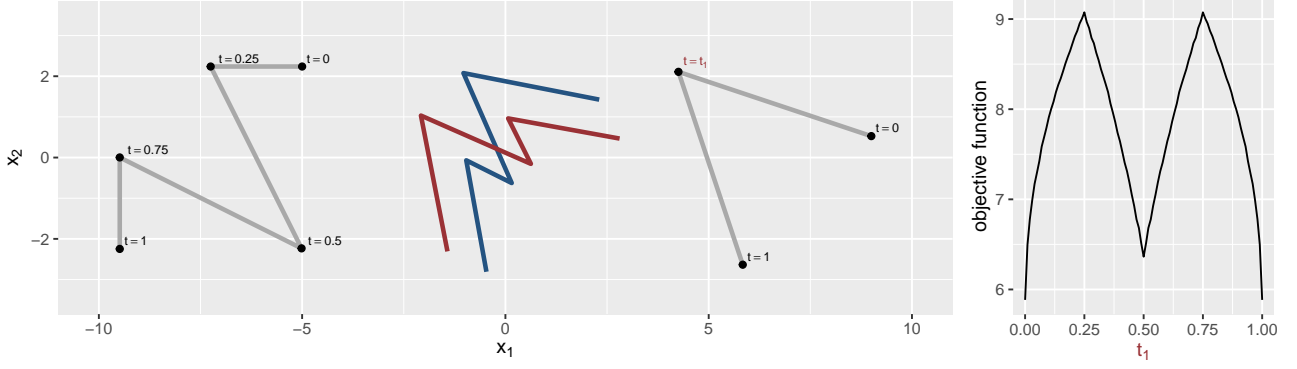


Figure 2: Left: Two piecewise linear curves in gray with Fréchet mean curves in red and blue. Right: Objective function with two modes. Both maximizer $t_1 = 0.25$ and $t_1 = 0.75$ correspond to optimal warping functions.

where

$$\langle \mathbf{p}(t), \begin{pmatrix} -3 \\ 1 \end{pmatrix} \rangle_+^2 = \begin{cases} 9^2 & \text{if } t \in [0, 0.25[\\ 0 & \text{if } t \in [0.25, 0.5[\\ 14^2 & \text{if } t \in [0.5, 0.75[\\ 0 & \text{if } t \in [0.75, 1] \end{cases} \quad \text{and} \quad \langle \mathbf{p}(t), \begin{pmatrix} 1 \\ -3 \end{pmatrix} \rangle_+^2 = \begin{cases} 0 & \text{if } t \in [0, 0.25[\\ 14^2 & \text{if } t \in [0.25, 0.5[\\ 0 & \text{if } t \in [0.5, 0.75[\\ 9^2 & \text{if } t \in [0.75, 1] \end{cases}. \quad (8)$$

With this we compute

$$\begin{aligned} \Phi(1 - t_1) &= \sqrt{0.5 \int_0^{1-t_1} \langle \mathbf{p}(t), \begin{pmatrix} -3 \\ 1 \end{pmatrix} \rangle_+^2 dt} + \sqrt{0.5 \int_{1-t_1}^1 \langle \mathbf{p}(t), \begin{pmatrix} 1 \\ -3 \end{pmatrix} \rangle_+^2 dt} \\ &= \sqrt{0.5 \int_{t_1}^1 \langle \mathbf{p}(1-t), \begin{pmatrix} -3 \\ 1 \end{pmatrix} \rangle_+^2 dt} + \sqrt{0.5 \int_0^{t_1} \langle \mathbf{p}(1-t), \begin{pmatrix} 1 \\ -3 \end{pmatrix} \rangle_+^2 dt} \\ &\stackrel{(8)}{=} \sqrt{0.5 \int_{t_1}^1 \langle \mathbf{p}(t), \begin{pmatrix} 1 \\ -3 \end{pmatrix} \rangle_+^2 dt} + \sqrt{0.5 \int_0^{t_1} \langle \mathbf{p}(t), \begin{pmatrix} -3 \\ 1 \end{pmatrix} \rangle_+^2 dt} \\ &= \Phi(t_1), \end{aligned}$$

which shows that Φ is symmetric around 0.5. Looking at the gradient of Φ given in Appendix B.2 we observe $\Phi'(t_1) > 0$ if $t_1 \in]0, 0.25[\cup]0.5, 0.75[$ and $\Phi'(t_1) < 0$ if $t_1 \in]0.25, 0.5[\cup]0.75, 1[$, which implies that both $t_1 = 0.25$ and $t_1 = 0.75$ are local maximizer and therefore global maximizer due to Φ being symmetric. For illustration of the objective function please refer to the right part of Figure 2.

The two maximiser of Φ correspond to two different optimal warping functions γ_1 and γ_2 of β_2 to β_1 . For $t_1 = 0.25$ we obtain $\hat{\gamma}_1$ according to (4) in Appendix B as

$$\hat{\gamma}_1(t) = \begin{cases} \frac{0.5 \langle \mathbf{p}(t), (-3, 1)^T \rangle_+^2}{\int_{i_j}^{i_{j+1}} \langle \mathbf{p}(t), (-3, 1)^T \rangle_+^2 dt} & \text{if } t \in [0, 0.25[\\ \frac{0.5 \langle \mathbf{p}(t), (1, -3)^T \rangle_+^2}{\int_{i_j}^{i_{j+1}} \langle \mathbf{p}(t), (1, -3)^T \rangle_+^2 dt} & \text{if } t \in [0.25, 1] \end{cases} = \begin{cases} \frac{0.5 \cdot 9^2}{0.25 \cdot 9^2} & \text{if } t \in [0, 0.25[\\ \frac{0.5 \langle \mathbf{p}(t), (1, -3)^T \rangle_+^2}{0.25 \cdot 14^2 + 0.25 \cdot 9^2} & \text{if } t \in [0.25, 1] \end{cases}.$$

Therefore, $\dot{\gamma}_1(t)$ for $t_1 = 0.25$ and analogously $\dot{\gamma}_2(t)$ for $t_1 = 0.75$ are piecewise constant with

$$\dot{\gamma}_1(t) = \begin{cases} 2 & \text{if } t \in [0, 0.25[\\ c_1 & \text{if } t \in [0.25, 0.5[\\ 0 & \text{if } t \in [0.5, 0.75[\\ c_2 & \text{if } t \in [0.75, 1] \end{cases} \quad \text{and } \dot{\gamma}_2(t) = \begin{cases} c_2 & \text{if } t \in [0, 0.25[\\ 0 & \text{if } t \in [0.25, 0.5[\\ c_1 & \text{if } t \in [0.5, 0.75[\\ 2 & \text{if } t \in [0.75, 1] \end{cases},$$

where the constant values are given as $c_1 = \frac{2 \cdot 14^2}{14^2 + 9^2}$ and $c_2 = \frac{2 \cdot 9^2}{14^2 + 9^2}$. Here, the form of the derivative of the second optimal warping function γ_2 of the second curve to the first curve is due to symmetry of this particular problem. Thus, both SRV-curves

$$\mathbf{q}(\gamma_1(t))\sqrt{\dot{\gamma}_1(t)} = \begin{cases} \sqrt{2}(-3, 1)^T & \text{if } t \in [0, 0.25[\\ \sqrt{c_1}(1, -3)^T & \text{if } t \in [0.25, 0.5[\\ 0 & \text{if } t \in [0.5, 0.75[\\ \sqrt{c_2}(1, -3)^T & \text{if } t \in [0.75, 1] \end{cases}$$

and

$$\mathbf{q}(\gamma_2(t))\sqrt{\dot{\gamma}_2(t)} = \begin{cases} \sqrt{c_2}(-3, 1)^T & \text{if } t \in [0, 0.25[\\ 0 & \text{if } t \in [0.25, 0.5[\\ \sqrt{c_1}(-3, 1)^T & \text{if } t \in [0.5, 0.75[\\ \sqrt{2}(1, -3)^T & \text{if } t \in [0.75, 1] \end{cases}$$

are SRV transformations of optimally aligned curves. This also means that both L_2 -means of \mathbf{p} and the SRV transformations $(\mathbf{q} \circ \gamma_i)\sqrt{\dot{\gamma}_i}$, $i = 1, 2$ of either optimally aligned β_2 are SRV transformations of Fréchet means of β_1 and β_2 (in red and blue in Figure 2).

To see this, let $\bar{\beta}$ be a curve with SRV transformation

$$\bar{\mathbf{p}} \in \left\{ \frac{1}{2}\mathbf{p} + \frac{1}{2}(\mathbf{q} \circ \gamma_i)\sqrt{\dot{\gamma}_i} \mid i = 1, 2 \right\}.$$

We compute for $i = 1, 2$

$$\begin{aligned} d([\beta_1], [\beta_2]) &\leq d([\beta_1], [\bar{\beta}]) + d([\bar{\beta}], [\beta_2]) \\ &= \inf_{\gamma} \left\| \frac{1}{2}\mathbf{p} + \frac{1}{2}(\mathbf{q} \circ \gamma_i)\sqrt{\dot{\gamma}_i} - (\mathbf{p} \circ \gamma)\sqrt{\dot{\gamma}} \right\|_{L_2} \\ &\quad + \inf_{\gamma} \left\| \frac{1}{2}\mathbf{p} + \frac{1}{2}(\mathbf{q} \circ \gamma_i)\sqrt{\dot{\gamma}_i} - (\mathbf{q} \circ \gamma_i \circ \gamma)\sqrt{\dot{\gamma}_i}\sqrt{\dot{\gamma}} \right\|_{L_2} \\ &\stackrel{\gamma=\text{id}}{\leq} \frac{1}{2} \left\| (\mathbf{q} \circ \gamma_i)\sqrt{\dot{\gamma}_i} - \mathbf{p} \right\|_{L_2} + \frac{1}{2} \left\| \mathbf{p} - (\mathbf{q} \circ \gamma_i)\sqrt{\dot{\gamma}_i} \right\|_{L_2} = \left\| \mathbf{p} - (\mathbf{q} \circ \gamma_i)\sqrt{\dot{\gamma}_i} \right\|_{L_2} \\ &= d([\beta_1], [\beta_2]), \end{aligned}$$

which shows that all inequalities have to be equalities and, therefore, γ the identity function. This also implies that $\bar{\beta}$ is optimally aligned to β_1 and $\beta_2 \circ \gamma_i$, $i = 1, 2$ and $d([\beta_1], [\bar{\beta}]) = d([\bar{\beta}], [\beta_2]) = \frac{1}{2}d([\beta_1], [\beta_2])$.

Hence, for every other curve $\tilde{\beta}$ it holds that

$$\begin{aligned} d([\tilde{\beta}], [\beta_1])^2 + d([\tilde{\beta}], [\beta_2])^2 &\geq 2 \left(\frac{d([\tilde{\beta}], [\beta_1]) + d([\tilde{\beta}], [\beta_2])}{2} \right)^2 \\ &\geq \frac{1}{2} d([\beta_1], [\beta_2])^2 \\ &= d([\bar{\beta}], [\beta_1])^2 + d([\bar{\beta}], [\beta_2])^2, \end{aligned}$$

where the first inequality is due to the square being convex and the second due to the triangle inequality. This shows that every $\bar{\beta}$ is a minimizer of the sum of squared distances and therefore a Fréchet mean. Hence, both $\frac{1}{2}\mathbf{p} + \frac{1}{2}(\mathbf{q} \circ \gamma_i)\sqrt{\gamma_i}$, $i = 1, 2$ are equivalently valid SRV transformations of Fréchet mean curves.

C.2 Identifiability of constant SRV splines

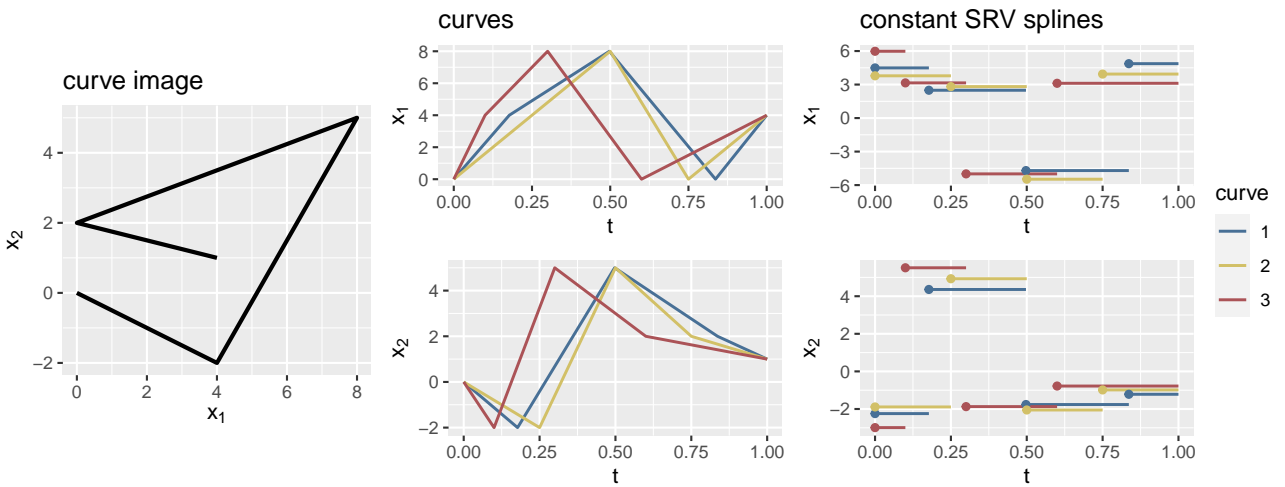


Figure 3: Three constant SRV splines (right) with corresponding linear spline curves (middle). All three of them have the same image displayed in black on the left.

Piecewise constant SRV-curves with varying knots are not identifiable. This means multiple constant SRV splines or equivalently linear spline curves can have the same image, as for example the curves displayed in Figure 3.

Fixing the set of knots determines the velocity between the knots and therefore the SRV transformation. Only if a knot is superfluous, i.e. the assignment of the knots to the corners of the polygonal image is not unique, there is more than one spline curve in each equivalence class for this case.

C.3 Splines of degree four are not identifiable

Spline curves with non-prime degree are not identifiable via their basis coefficients modulo warping. Consider the following counterexample with splines of degree four. Let

$$\mathbf{Q}(t) = \begin{pmatrix} 4t^4 - 2t^2 \\ 4t^4 \end{pmatrix} \text{ and } \mathbf{P}(t) = \begin{pmatrix} t^4 + 2t^3 - t \\ t^4 + 2t^3 + t^2 \end{pmatrix}.$$

Then $\gamma(t) = \sqrt{0.5(t^2 + t)}$ is a suitable warping function since it fulfills $\mathbf{P} = \mathbf{Q}(\gamma(t))$ and is monotonically increasing and onto, but monomial coefficients differ between \mathbf{P} and \mathbf{Q} and are thus not identifiable modulo

warping. Note that the counterexample for splines of degree 4 could similarly be constructed for all splines with any degree that is not a prime number. If the degree of the splines is a prime number, it seems possible that one can show a similar identifiability result as in Theorem 2. This would imply identifiability for quadratic SRV-curves using an analogous argument as in Corollary 1.

Web Appendix D Further simulations and supplementary plots

D.1 Simulation: Aligning sparsely and irregularly sampled curves

In this first simulation, we compare our methods (available in the R-package “elasdics” (Steyer, 2021)) for aligning sparsely and irregularly sampled curves to the implementation of the dynamic programming (DP) algorithm in the existing R package “fdasrvf” (Tucker, 2020) based on Srivastava et al. (2010). Since this DP implementation only allows for an equal number of observed points on both curves, we restrict the simulation to this case, although we developed our methods in particular for differing numbers of observed points per curve. In Figure 4, we present one simulated example for open and closed curves each.

For the open setting, we choose a parameterized curve $\beta(t) = \sin(t)(\cos(12t) + 2t, \sin(12t) + t)^T$, which we use as a template for both curves. The first curve β_1 (displayed in red in Figure 4) is obtained via sampling an unbalanced observation grid t_1, \dots, t_m with $m \in \{10, 30, 50\}$ and adding a Gaussian random walk error (with standard deviation $sd = 0.01$) to the evaluations $\beta_1(t_1), \dots, \beta_1(t_m)$. The second curve β_2 is re-sampled 100 times (displayed in grey in Figure 4) using the same sampling scheme as for β_1 .

For the closed setting we choose two butterfly shapes available in “fdasrvf” (Tucker, 2020). These are discretely observed curves with 100 observations each. We down-sample the curves such that $m \in \{30, 60, 90\}$ points per curve are left and such that points with high estimated curvature are more likely to be included. This way, the images of the curves are well preserved, as we are more likely to remove points on straight lines. Furthermore, we add an error term $\sin(\pi \frac{j-1}{m-1})\epsilon_j$ to the j -th remaining observation for all $j = 1, \dots, m$, where ϵ_j is distributed according to a Gaussian random walk with standard deviation $sd = 0.5$ and the modification with the sinus function ensures closedness. According to this sampling scheme, we draw one copy (plotted in red) of β_1 from the first butterfly shape and 100 copies (plotted in grey) of β_2 from the second butterfly shape.

For each of the settings we compare the optimal alignment for each copy of β_2 to the corresponding β_1 using our coordinate-wise-optimization (CWO) algorithm with the alignment produced by the dynamic programming (DP) from “fdasrvf” (Tucker, 2020). When looking at the coordinates separately, we visually observe slightly better alignment for our method CWO compared to DP. This is also evident in a smaller average elastic distance, e.g. on average 0.92 vs. 1.23 and 24.71 vs. 33.18 for $m = 30$ in the open and closed setting, respectively, and 0.67 vs. 1.19 (open, $m = 50$) and 17.9 vs. 21.37 (closed, $m = 60$) for moderate m . As expected, this difference decreases if 90 points of the butterfly shapes are selected (19.07 vs. 19.28 on average), as in this case the points are nearly observed on a regular, fairly dense grid, which is the setting the implementation in “fdasrvf” is designed for.

A highly unbalanced distribution of observed points on the curves described above causes difficulties for the mean computation in “fdasrvf” (Tucker, 2020) as well. Figure 5 demonstrates this for sets of partially densely and partially sparsely observed curves each, for which we compute means with respect to the elastic distance. The means in red, which are computed by the `curve_karcher_mean` function in “fdasrvf” (Tucker, 2020), do not capture the image of the observed curves as well as our methods (e.g. butterfly shape in blue) which are specifically developed for such unbalanced data. Visually, the blue butterfly shape captures small features, like the shape of the wings and the tail, better than the red one. Since the implementation in “fdasrvf” aims at computing a mean with respect to the geodesic shape distance, i.e. minimizes the geodesic distance on the sub-

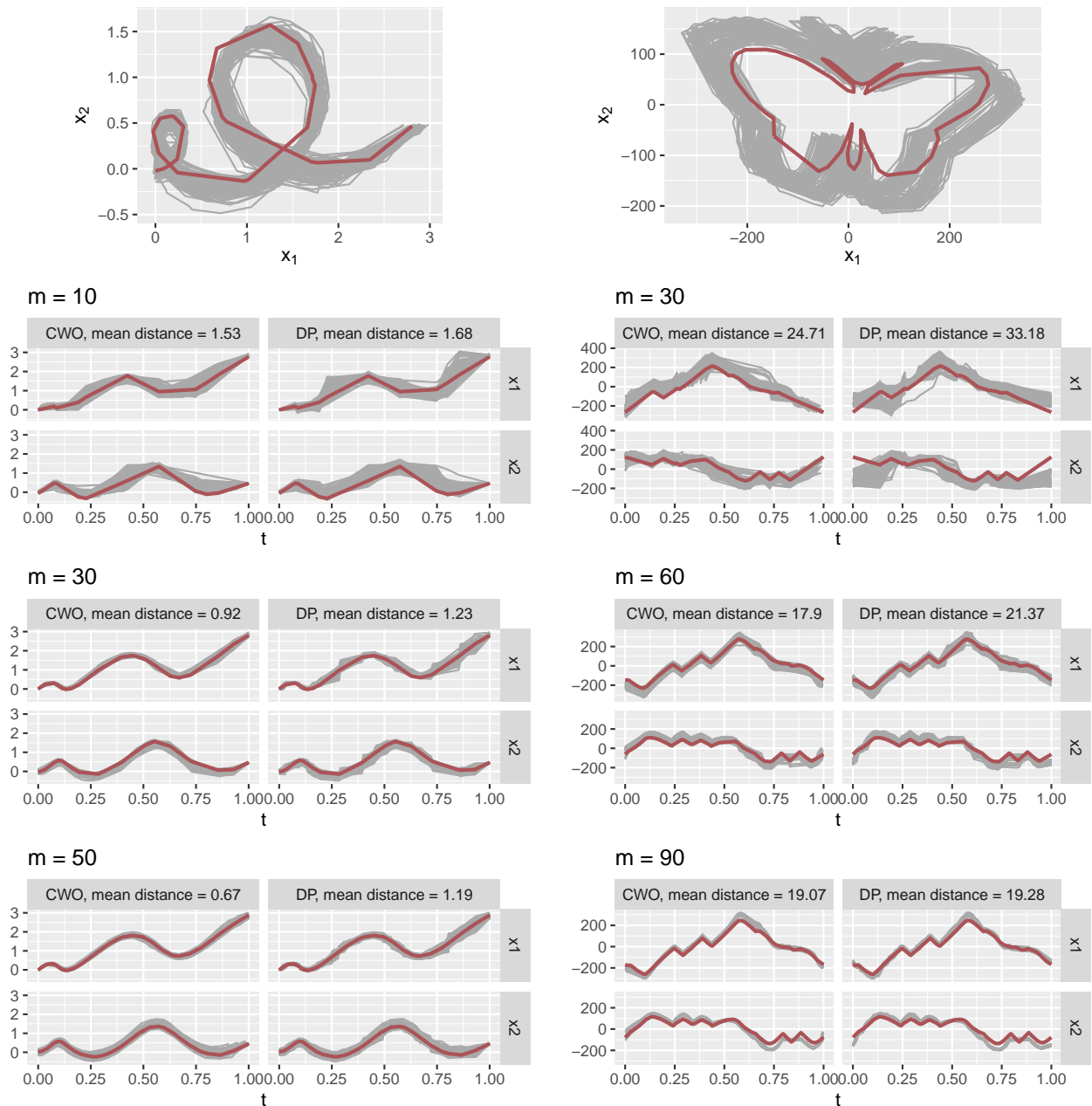


Figure 4: Comparison of the optimal alignment produced by our method CWO and the one computed with DP. The 100 gray curves are sampled with m points per curve and aligned to the red curve. The first row shows the sampled curves in the moderately sparse setting ($m = 30$ or $m = 60$ points per curve for the open or closed curve, respectively). The optimal alignments found by both methods are depicted in the lower rows, with the resulting mean elastic distances given in the headings. To make the alignment visually comparable, the aligned curves are evaluated at the observation grid of the red curve for DP.

manifold of (closed) curves with fixed curve length, the results are not completely comparable. Nevertheless, in particular for the open curves, which are of similar length, we expect the impact of this aspect to be relatively small compared to the warping.

D.2 Simulation: Convergence of spline mean coefficients

Additional to the simulation for the closed heart shaped template in the main document, we observe convergences of spline mean coefficients for two open templates here. The curves in Fig. 6 are sampled from linear splines on SRV level with three or nine equally spaced inner knots, using a standard deviation of $\sigma = 0.3$ or

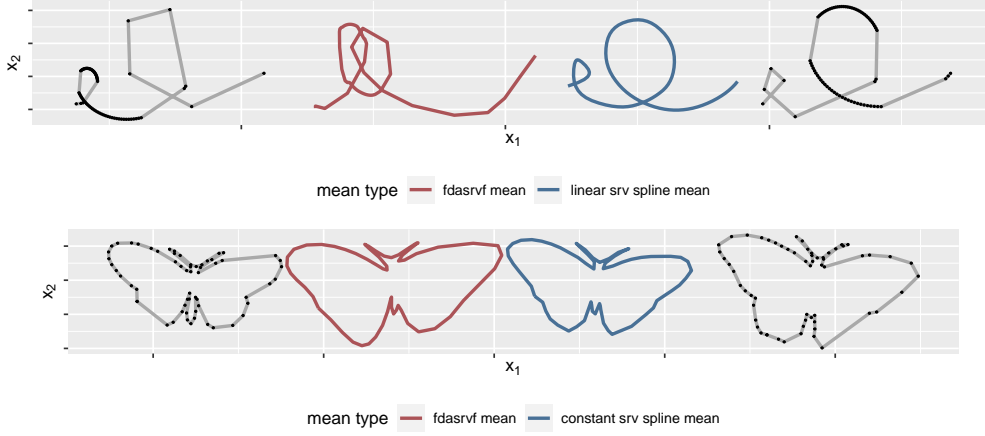


Figure 5: Elastic means for irregularly sampled curves. The observed curves are displayed in gray with black dots at the observed points. The red mean curves are computed with the “fdasrvf” package, the blue mean curves are computed using our methods and linear splines with 13 equally spaced inner knots or constant splines on SRV level with 68 equally spaced inner knots for the open curves and the closed butterfly shaped curves, respectively.

$\sigma = 0.4$ for the spline coefficients, respectively. Fig. 7 and Fig. 8 show the distribution of the estimated means and the corresponding coefficients for 100 repetitions of each setting. Increasing the number of repetitions from 20 to 40 and from 40 to 100 per setting had little effect on the overall picture, which is why we consider 100 simulation runs to be sufficient. For the first template, the estimation does not improve much if we select more points per curve. Since for this template a low number of coefficients has to be estimated, a small number of observed points $m_i \in [10, 15]$ per curve seems sufficient. The results for the second open template are similar to the closed heart-shaped template discussed in the main paper. In all three settings, our theoretical results on identifiability of spline coefficients (Corollary 1) and the continuity of the embedding (Lemma 2) are confirmed.

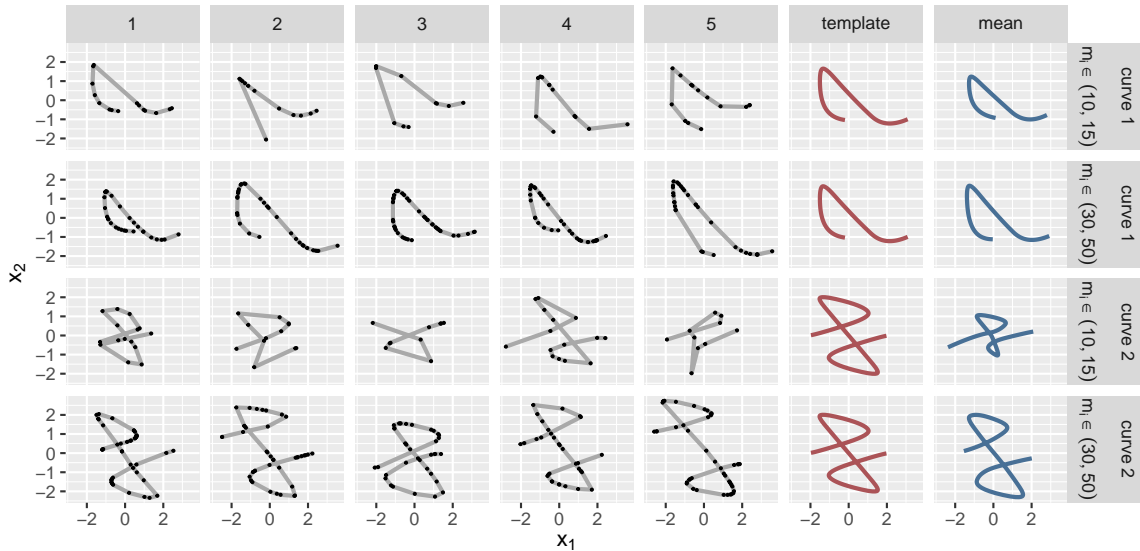


Figure 6: Example simulated data in gray with observed values marked as black dots and corresponding smooth elastic means over $n = 5$ observations in blue. The irregularly sampled curves are drawn from two different templates (in red) with varying number m_i of observed points per curve.

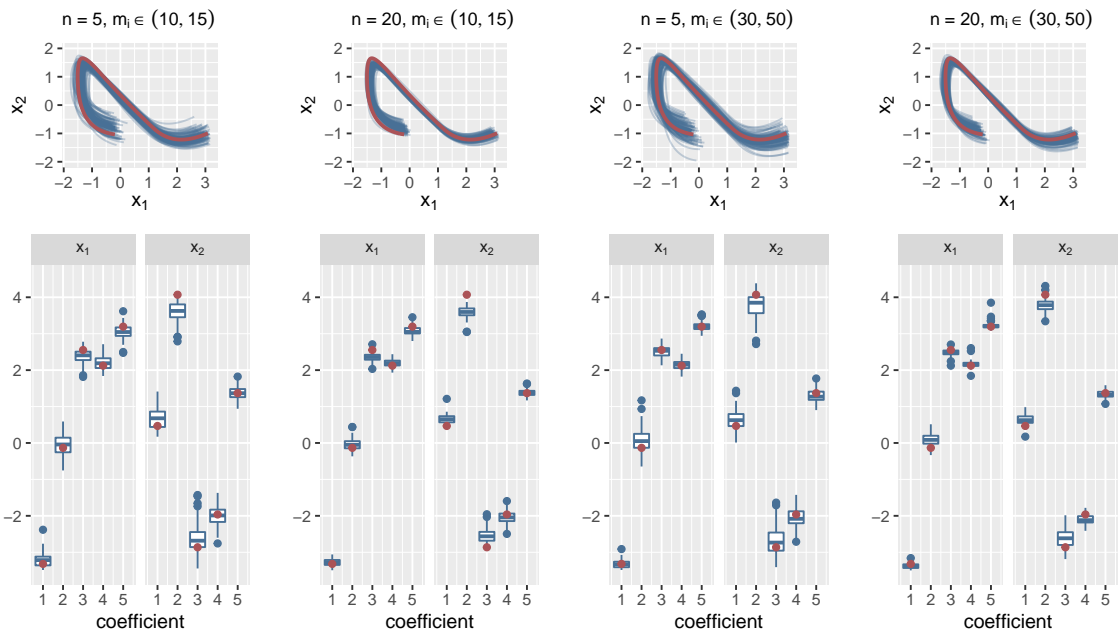


Figure 7: Top: Smooth means (in blue) computed for a set of n curves drawn from the open template curve (in red) via sampling its B-spline coefficients from a normal distribution with standard deviation $\sigma = 0.3$ and $m_i, i = 1, \dots, n$ points observed per curve. The means are computed using linear SRV splines and the same knot set as the template (three equally spaced inner knots)
 Bottom: Corresponding distribution of spline mean coefficients (in blue) and template coefficients (in red).

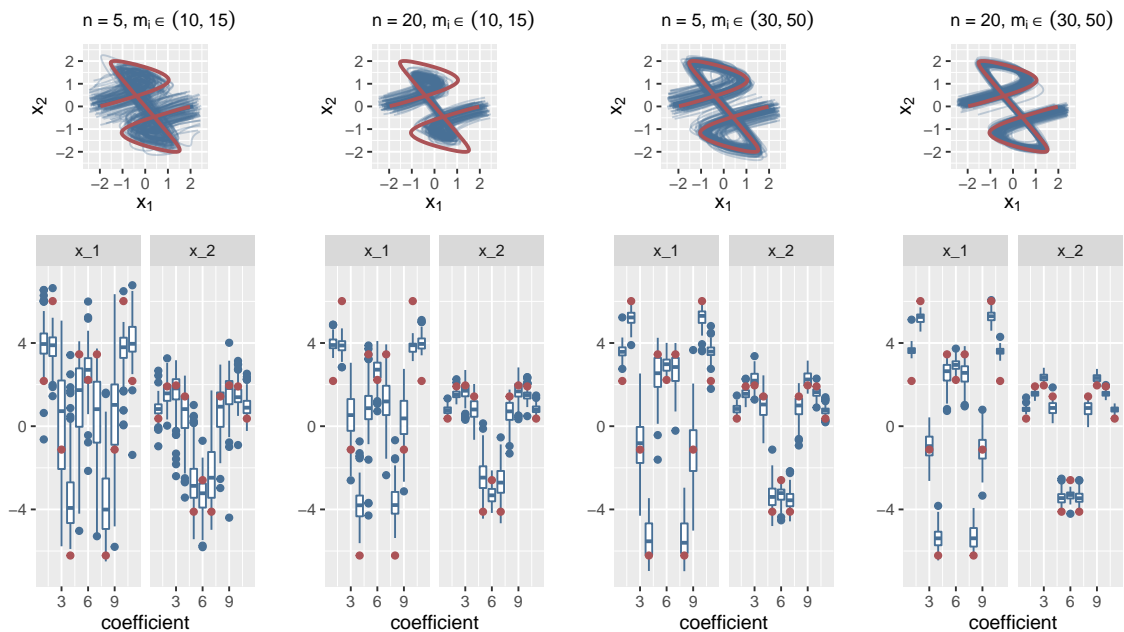


Figure 8: Top: Smooth means (in blue) computed for a set of n curves drawn from the open template curve (in red) via sampling its B-spline coefficients from a normal distribution with standard deviation $\sigma = 0.4$ and $m_i, i = 1, \dots, n$ points observed per curve. The means are computed using linear SRV splines and the same knot set as the template (nine equally spaced inner knots)
 Bottom: Corresponding distribution of spline mean coefficients (in blue) and template coefficients (in red).

D.3 Simulation: Misspecified spline model

The last simulation elaborates on the convergence of the spline means in case of model misspecification. Figure 9 shows means with varying knots using linear SRV splines (smooth means in blue) or constant SRV splines (polygonal means in red). All means are computed for the same set of $n = 20$ heart-shaped curves, which have been sampled as described above from the third template with $m_i \in \{30, \dots, 50\}$ points per curve. For a sufficient number of knots, both the smooth and the polygonal means reproduce the original heart shape well. If we consider the number of coefficients n_{coeffs} as a measure for model complexity, we observe that the smooth means are closer to the template than the polygonal ones, given the same number of coefficients, with a local minimum at the correctly specified model. This shows that one can obtain more parsimonious models for smooth means using linear SRV-curves. Even though the distance to the template for a polygonal mean can be reduced by using more knots, it does not seem to become as low as for the linear SRV mean.

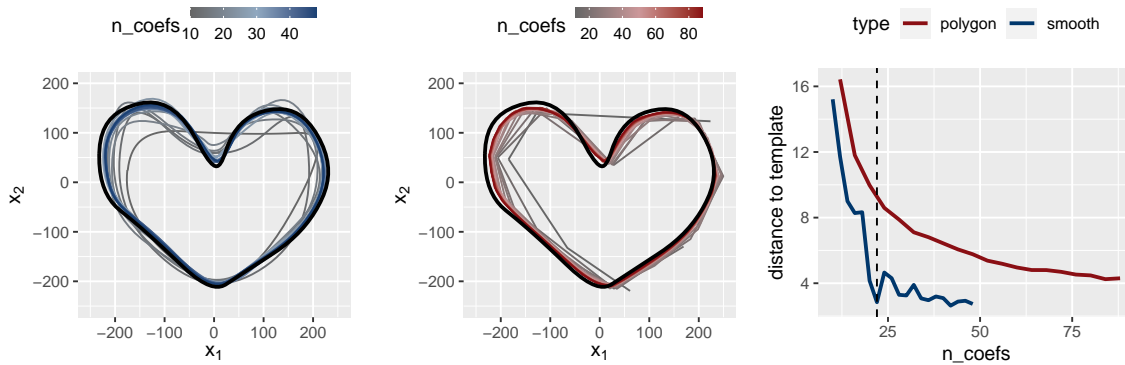


Figure 9: Left: Smooth mean based on linear splines on SRV level with varying number of knots and therefore coefficients computed on a sample of 20 curves with $m_i \in \{30, 50\}$ points per curve. The template is displayed in black.

Middle: Polygonal mean with varying number of coefficients computed for the same sample of curves (from the same template in black) as the smooth means on the left.

Right: Elastic Distance of the mean curves to the template curve given the number of coefficients in the mean model. The vertical dashed line indicates the true linear spline model with 22 coefficients.

This indicates that using linear SRV splines for modeling a smooth “true” mean might reduce the bias due to under-sampling the curves. While we see a local minimum for the n_{coeffs} used to generate the data, close n_{coeffs} give similar results and in particular values larger than the true one give similarly good results, with the distance generally decreasing in n_{coeffs} . This indicates that results are not very sensitive to n_{coeffs} given it is sufficiently large.

Web Appendix E Classifying spiral curve drawings for detecting Parkinson’s disease

The Archimedes spiral-drawing test is a common, non-invasive tool for diagnosing patients with Parkinson’s disease. Usually, the drawing task is performed on paper and analyzed by medical experts to identify deviations of the shape to the spiral template (Alty et al., 2017). Recently, there have been approaches using digitizing tablets to obtain more detailed data, not only on the image of the spiral curve but also on the position of the pen at each time point (Saunders-Pullman et al., 2008; Isenkul et al., 2014).

In addition to this *static* spiral test, Isenkul et al. (2014) proposed a modified, *dynamic* spiral test, where the template spiral curve appears and disappears (“blinks”) in certain time intervals. Fig. 10 shows the spiral curves drawn by 25 Parkinson’s patients and 15 controls in both tests. It is visually notable that the controls

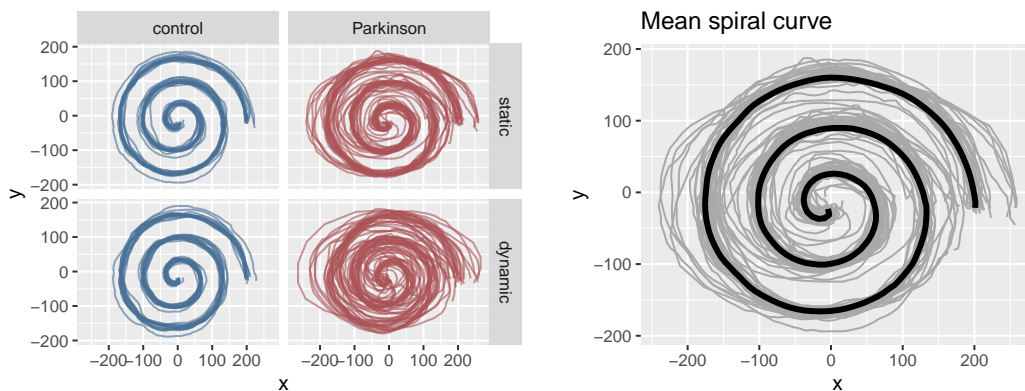


Figure 10: Left: Spiral curves drawn by either a healthy control group or by patients with Parkinson’s disease in two different settings. Right: The mean curve (black) of all static curves (gray) computed with respect to the elastic distance.

follow the template more closely than the Parkinson’s patients. This difference seems to be more severe for the dynamic spiral test.

E.1 Classification based on the elastic distance to a template

While the authors of the original study based their analysis on differences in speed distributions of both tasks, Kurt et al. (2019) pre-aligned the spiral curves using a heuristic dynamic time warping algorithm. We follow up on this, but instead we use the elastic distance in (1) as a proper distance. Moreover, we only consider highly interpretable decision rules that classify an individual as being at high risk of having Parkinson’s disease if the distance of the drawn curve to the template exceeds a threshold. This mimics the decision made by medical experts based on the spiral drawing and allows us to assess whether the additional information on time or speed provided by a tablet is actually necessary for good classification.

We only use 10% of the values per curve, which results in irregularly sampled curves with 55 to 269 points each. Visual inspection indicates good agreement between the images of the down-sampled and the original curves. In Subsection E.3 we discuss why down-sampling is necessary and how it influences the accuracy of the classification, also comparing to “fdasrvf” in subsection E.4.

We want to base our classification on deviations from the template the participants had to follow. Since the original parametrized template curve is not available, we compute the elastic mean (see Subsection 2.5) of all curves from the static spiral test using piecewise constant splines with 201 knots on SRV level. We then use the resulting polygonal mean (displayed in black in Fig. 10, top) as a template curve.

Fig. 11 shows the elastic distances of the drawn curves to the template curve. As expected, it is generally larger for Parkinson’s patients than for controls in both settings. The scatter plot on the right indicates a strong positive correlation between the distances in the static and in the dynamic test for healthy subjects, which is not present for Parkinson’s patients.

We propose intuitive decision rules of the form: Classify as “Parkinson” if the distance of the drawn curve to the mean curve exceeds a threshold. The gray areas in Fig. 11 (left and middle) indicate the corresponding decision rule for curves in the static or dynamic test. Alternatively, we classify as “Parkinson” if either of the two distances exceeds a respective threshold (Fig. 11, right). To estimate the thresholds, we optimize the zero-one loss (misclassification loss), which is feasible for our small dataset and set of decision functions.

Leave-one-out cross-validation indicates 72.5% accuracy for the static, 90.0% accuracy for the dynamic setting and 92.5% accuracy for the classifier based on both. Since for the latter we observe only one misclassi-

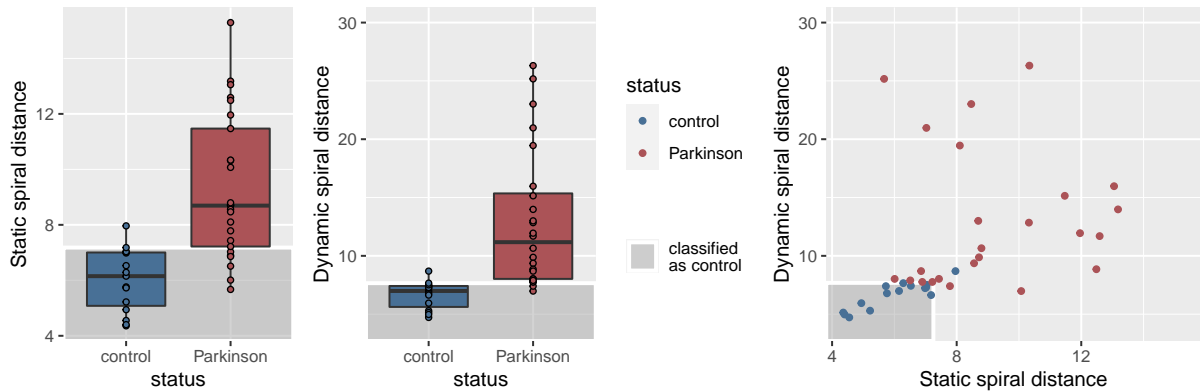


Figure 11: Left: Distance of the curves drawn by the participants to the mean spiral curve for both settings. Right: Distance of the curve in the static setting compared to the distance of the curve in the dynamic setting. The gray areas indicate the decision rule based on the zero-one loss. Note that one observation for a Parkinson’s patient with an extreme distance greater than 40 in the dynamic setting is not displayed.

fied observation in-sample, including additional features like the difference or the ratio of the two distances is not advisable, but could further improve classification if more data were available. Moreover, it seems that the spiral drawings show more variability in x-coordinate direction than in y-coordinate direction. To address this, it could also be beneficial to explore different weights on the coordinate axes in the elastic distance computation.

To see that an elastic analysis is necessary, we compare our elastic analysis to a simpler classifier based on the usual L_2 -distance. For this, we re-parametrize the curves according to relative arc length to account for different speed patterns but do no further alignment. We obtain accuracies of 55.0% for the static setting, 80.0% for the dynamic setting and 77.5% for the classifier based on both, indicating clearly better performance of an elastic analysis.

In conclusion, the elastic distance of the drawn curve to a template is an intuitive measure of performance for both the static and the dynamic spiral drawing test. Using this feature, we mimic and objectify a doctor’s medical diagnosis process, and obtain highly accurate classification. If more data were available, maybe even from patients with related neurological conditions like essential tremor, it might also be beneficial to analyze the whole aligned curves instead of their distances to the template, or to additionally analyze the temporal information provided by the warping functions, which our approach allows to separate from the images but which provide no additional information in this study.

E.2 Warping functions of misclassified subjects

Elastic alignment of the observed curves to a template allows us to separate phase and amplitude variation. Our classifiers depend on the elastic distance, which means we rely only on the amplitude variation. To see if the phase, that is the temporal pattern, yields additional information compared to only the image, we look in Fig. 12 at the warping functions separated according to the classification result. This comparison of real-time parametrization to the parametrization after alignment to the mean curve shows whether the speed patterns of patients with Parkinson’s disease are dissimilar to those of healthy individuals.

Looking at the general pattern of the warping functions in both settings, we observe more deviation from a smooth speed pattern in the group of Parkinson’s patients than in the control group. To decide whether this yields additional information to the elastic distance of the curve to the template, we further inspect the warping curves which belong to misclassified subjects. There are two Parkinson’s patients with conspicuous speed patterns we misclassify as “control” in the static setting. Their speed pattern shows starting and stopping

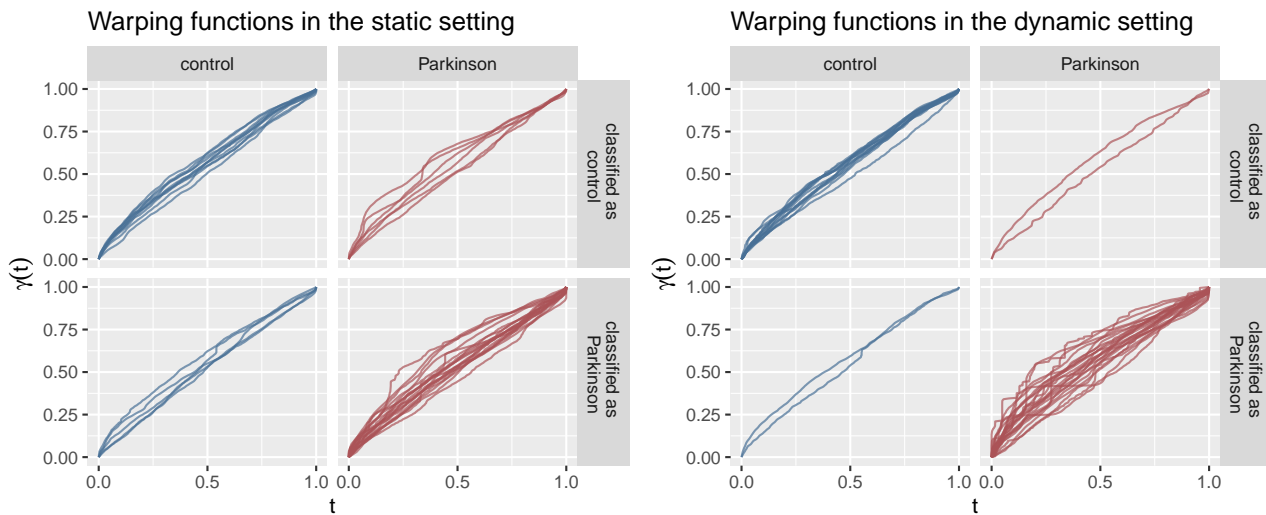


Figure 12: Optimal warping in both settings separated by the actual status and the predicted status using the classifiers based on only the corresponding distance each and leave-one-out cross-validation.

motions, which is not present in the curves of any of the healthy control subjects. Contrarily, we do not observe any noticeably different speed pattern for the misclassified individuals in the dynamic setting. Here the image of the curve seems to capture all available information on the status of the participant.

E.3 Influence of down-sampling on the accuracy

Adhoc analysis of the full data is not recommended as the spirals are observed densely but with (small) errors. This is problematic because in the SRV framework the analysis is performed at the level of the derivative and small errors on function level can cause large errors on the level of the derivative and therefore on SRV level.

Using only a fraction of the values per curve effectively serves as a smoothing method, which reduces variability in the observed SRV-curves. We used 10% of the values, as this gives 55 to 269 points per curve, which seems reasonable to represent the spiral shape of the data. To investigate the influence of the coarsening on the accuracy, we conducted our analyses again with a varying fraction of points per curve and observed the accuracy in the dynamic spiral drawing setting (Tab. 1).

fraction of points used	0.02	0.05	0.07	0.08	0.10	0.11	0.12	0.15
accuracy	0.875	0.900	0.875	0.925	0.925	0.925	0.875	0.825
run-time mean computation	11.110	13.950	16.060	18.130	23.190	22.820	24.360	30.830

Table 1: Classification accuracy in the dynamic setting with a varying fraction of points per curve. The last line gives the system run-times for the mean computation in seconds.

As expected, excessive thinning of the data (2%, 5% or 7%) leads to poorer classification results. However, it seems that the classifier is not extremely sensitive to the degree of coarsening, as 8% and 11% give the same accuracy as 10% of the data. If even more points per curve are used, the accuracy decreases again. Since the run-times (Tab. 2) for the mean computation do not increase substantially, coarsening the curves to a “reasonable” number of sample points seems to be more a matter of accuracy than of computation times. Providing rigorous guidelines for identifying optimal sample frequencies in the SRV framework is, although an interesting topic in itself, beyond the scope of this paper, which focuses on sparsely sampled curves.

E.4 Comparison with the package “fdasrvf”

In the following, we compare the performance of our methods with the implementation in the “fdasrvf” package for this application. Since the implementation in “fdasrvf” only allows an equal number of points per curve, we select points on a regular time grid. Analogously to the analysis with our package “elasdics”, we first calculate an elastic mean of the observed curves in the static setting (with the function `curve_karcher_mean` and option `rotated = "F"` to exclude rotation alignment) and then calculate the distance of each curve in the dynamic setting to this mean using their function `calc_shape_dist`.

number of points	50	100	200	400
elasdics	0.850	0.875	0.900	0.850
fdasrvf	0.800	0.850	0.825	0.850

Table 2: Comparison of the classification accuracy in the dynamic setting with a varying number of points per curve.

Tab. 2 shows that our method shows better classification results than if the implementation in the package “fdasrvf” is used for this application, regardless of the number of selected points per curve. Tab. 3 shows that the computations times for the elastic mean are always lower using “elasdics” than “fdasrvf”.

number of points	50	100	200	400
elasdics	11.090	14.520	27.530	62.550
fdasrvf	25.340	110.560	491.470	2591.210

Table 3: Run-times for the mean computation of the spiral data in seconds.

The lower classification accuracy for the “fdasrvf” package is due to a worse mean computation (Fig. 13, left). We have already seen in other scenarios that their implementation is sensitive to the grid on which the curves are observed (cf. Fig. 5, Supporting Information). To demonstrate that this is also the case here, we carry out a further simulation to evaluate the influence of the observation grid on the mean calculation (Fig. 13, middle) and the accuracy (Fig. 13, right) in the dynamic setting for both packages.

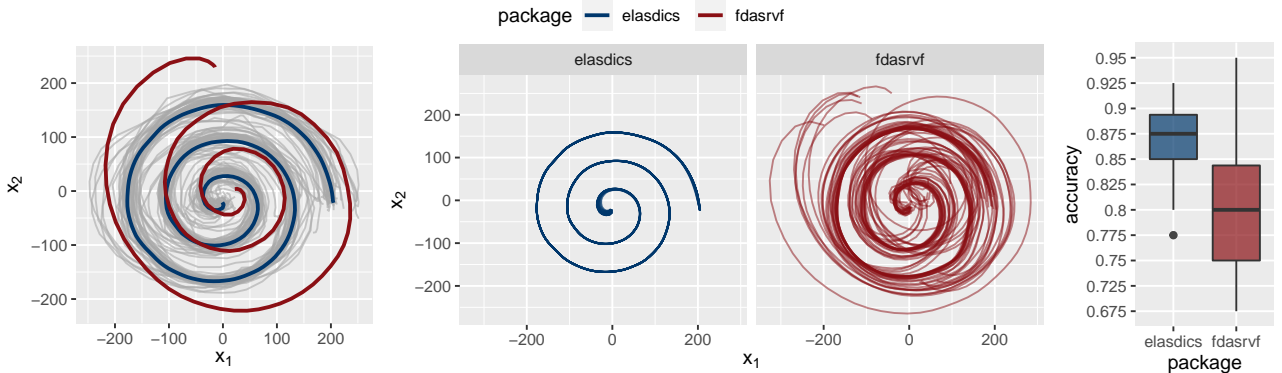


Figure 13: Left: Comparison of means for the spirals in the static setting with 100 observations per curve. Middle: Elastic means after re-sampling the observation grid. Right: Accuracy in the dynamic setting after re-sampling the observation grid.

Here, we first re-parametrize the curves with respect to arc length and then sample an observation grid for each curve via $F_{a,b}(\{0, 0.01, 0.02, \dots, 0.99, 1\})$ where $F_{a,b}$ is the cumulative distribution function of the Beta distribution with parameters $a, b \sim U_{[0.7, 1]}$. Our mean is little affected by these randomly chosen observation grids, while the “fdasrvf” mean is more scattered. Also, the accuracy in the dynamic setting varies more and is on average lower compared to an analysis with our package.

References

- Alty, J., Cosgrove, J., Thorpe, D., and Kempster, P. (2017). How to use pen and paper tasks to aid tremor diagnosis in the clinic. *Practical Neurology* **17**, 456–463.
- Bruveris, M. (2016). Optimal reparametrizations in the square root velocity framework. *SIAM Journal on Mathematical Analysis* **48**, 4335–4354.
- Isenkul, M., Sakar, B., Kursun, O., et al. (2014). Improved spiral test using digitized graphics tablet for monitoring parkinson’s disease. In *The 2nd international conference on e-health and telemedicine (ICEHTM-2014)*, volume 5, pages 171–175.
- Kurt, İ., Ulukaya, S., and Erdem, O. (2019). Classification of Parkinson’s disease using dynamic time warping. In *27th Telecommunications Forum (TELFOR)*, pages 1–4. IEEE.
- Lahiri, S., Robinson, D., and Klassen, E. (2015). Precise matching of PL curves in R^N in the square root velocity framework. *Geometry, Imaging and Computing* **2**, 133–186.
- Saunders-Pullman, R., Derby, C., Stanley, K., Floyd, A., Bressman, S., Lipton, R. B., Deligtisch, A., Severt, L., Yu, Q., Kurtis, M., et al. (2008). Validity of spiral analysis in early parkinson’s disease. *Movement disorders: official journal of the Movement Disorder Society* **23**, 531–537.
- Srivastava, A., Klassen, E., Joshi, S. H., and Jermyn, I. H. (2010). Shape analysis of elastic curves in euclidean spaces. *IEEE transactions on pattern analysis and machine intelligence* **33**, 1415–1428.
- Steyer, L. (2021). *elasdics: Elastic Analysis of Sparse, Dense and Irregular Curves*. R package version 0.2.0.
- Sun, W. and Yuan, Y. (2006). *Optimization Theory and Methods: Nonlinear Programming*. Springer Optimization and Its Applications. Springer US.
- Tucker, J. D. (2020). *fdasrvf: Elastic Functional Data Analysis*. R package version 1.9.7.

Eidesstattliche Versicherung

gemäß Promotionsordnung vom 12. Juli 2011, §8 Abs. 2 Pkt. 5

Ich versichere hiermit, dass die Dissertation von mir eigenständig und ohne unerlaubte Hilfsmittel angefertigt wurde.

Weiterhin versichere ich, dass ich keine anderen als die von mir angegebenen Quellen verwendet habe und die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Berlin, 13.12.2013

Unterschrift Lisa Maike Steyer

