Flexible Regression for Functional Object Data: Curves, Shapes and Densities

Jan Almond Stöcker

München 2022

Flexible Regression for Functional Object Data: Curves, Shapes and Densities

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

handed in by
Jan Almond Stöcker

Munich, July 27th 2022

First Referée: Prof. Dr. Sonja Greven
Second Referée: Prof. Dr. Helmut Küchenhoff
Third Referée: Prof. Anuj Srivastava, PhD

Defense of thesis on September 16th 2022

Flexible Regression für funktionale Objektdaten: Kurven, Formen und Dichten

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

eingereicht von
Jan Almond Stöcker

München, den 27. Juli 2022

Erstgutachterin: Prof. Dr. Sonja Greven
Zweitgutachter: Prof. Dr. Helmut Küchenhoff
Drittgutachter: Prof. Anuj Srivastava, PhD

Disputation am 16. September 2022

Acknowledgments

Cooperation is a lot about trust, trust that I was able to place in the people around me and trust that other people placed in me and my opinion, for which I am deeply grateful. The respectful together I enjoyed over the past years has constantly fueled my enthusiasm for statistics, while having plenty of fun at the same time.

In this sense, I first want to thank Sonja Greven for her guidance, for always being able to rely on her, for her confidence in my ideas, and for all that I could learn from her in all we have been working on together in the last years, from years-long projects to last-minute abstracts.

Similarly, I want to thank my colleagues and friends Alexander Volkmann, Lisa Steyer, Eva-Maria Maier, Sarah Brockhaus and especially also David Rügamer, with all of whom I enjoyed closely working together and share many awesome experiences, as well as Amanda Fernández-Fontelo, Fabian Scheipl, Meike Köhler, Jona Cederbaum, Matthias Eckardt, Elena Ivanova, Clara Happ, Sigbert Klinke and Karen Fuchs from my working group, with whom I had a great time, always enjoying a nice atmosphere.

Many thanks also to all my other collaborators. Here, I would like to especially mention Honey Alas, Christoph Berninger and also Manuel Pfeuffer, with whom I had the pleasure of working very intensively together, as well as Nikolaus Umlauf, Nadja Pöllath, Sebastian Düsing and, lately but also with great fun, Alessandra Menafoglio.

Warm thanks to Thomas Augustin and a long list of members of the department in Munich who made it such a friendly place to work and study at. Thank you Benjamin Sischka, Christoph Jansen, Henry Port, Shuai Shao, Elke Höfner, Brigitte Maxa, my former office mates Moritz Berger and Moritz Herrmann, and many more!

Thank you also Leslie Udvarhelyi, it was a pleasure, and Lucas Kock, Paul Bach and the other members of Nadja Klein's group in Berlin!

Thanks to Manuel Kroiss, Can Gürer, Ludwig Bothmann, Minh-Anh Le and Xudong Sun for inspiring me with their commitment.

I want to thank my friends, several of whom I know since childhood, Fritz Francisco, Merit Enghofer, Paula Giesler, Timon Enghofer, Marius Heiß, Max Pöhlmann, Tore Erdmann, Henning Bumann and recently Simon Bittmann, with whom I had great fun discussing data problems in their fields, and all my other friends who support me.

I am immensely grateful to my family, the basis of my confidence, and the dear family of Sophia. Besides for being also a great collaborator and friend, I want to thank Sophia Schaffer, who is always by my side, for her advice and support in so many regards, for the beautiful time working remotely together at home, her patience when finishing urgent work never seemed to end, and much more.

Moreover, I would like to sincerely thank Helmut Küchenhoff and Anuj Srivastava for reviewing my dissertation. It is a particular honor and pleasure, as they represent the two sides of the thesis so excellently: it was in Helmut Küchenhoff's descriptive statistics lecture where I had my first contact with regression models, and Anuj Srivastava's work on functional shape analysis plays such a vital role for the data objects we address.

Summary

The interplay of geometric and probabilistic approaches in statistics is already evident in the example of linear regression. However, it becomes particularly explicit when it comes to extending statistical methods to object data with a non-Euclidean structure. Functional data represent such a data type, where a sample of functions, such as growth curves or motion trajectories, is considered and analyzed as such.

The starting point of this work are models for functional data with flexible tensor product spline effects based on (generalized) additive regression. In this context, we discuss different scenarios with functional target variables, which require modeling beyond usual point-wise mean curves: I. distributional regression models, where, for example, also the variance function is modeled in dependence on covariates; II. models for probability densities as a functional response; and III. models for multidimensional curves and their shapes. Methodological extensions with respect to these three aspects are proposed in a total of seven subprojects, each presented in a contribution chapter, and applied to various, mostly biometric but also econometric problems:

I. Already the example of growth curves, an archetype of functional data, shows that the (implicit) assumption of a point-wise normally distributed response variable can prove to be problematic (purely positive, often skewed distribution) and exclusive modeling a mean curve in dependence on covariates sometimes proves to be restrictive (e.g. when modifications of the medium of bacterial cultures influence their growth process). To address such challenges, we extend functional additive models (FAMs) to distributional regression (GAMLSS). For model estimation based on gradient boosting, we illustrate how suitable regularization helps to handle high autocorrelation of functional responses (penalization at each step and early stopping of the algorithm based on curve-wise cross-validation). The flexible approach allows us to address experimental specifics when modeling an interaction scenario of two bacterial strains and to identify different phases of bacterial competition.

II. In contrast to the extension of (point-wise) distribution models to functional responses in I., distributions themselves can be considered as objects of functional data analysis. Accordingly, we extend FAMs to probability densities as response variables. Due to their specific properties, densities are modeled in a Bayes Hilbert space. Again, we use a boosting algorithm for estimation, with the goal of minimizing expected quadratic distances in the Bayes space. In an analysis of gender-based income inequality based on the Socio-Economic Panel (SOEP), we model the distribution of income shares of the woman in the total income of couples as a continuous density with point masses at 0 and 1 in dependence on various influencing variables.

III. Modeling multidimensional curves primarily corresponds to a multivariate ex-

tension of real-valued functional data (Contribution a)), but also poses further challenges: if, for example, the outline curve of an object, e.g. a bone, is considered, the spatial orientation (b), d), e)) or parameterization of the curve (c), d), e)) often do not play a role, which should be taken into account in the model. Although the contributions are partly strongly based on each other, they always set individual accents:

a) In the first contribution, we propose a multivariate functional mixed model in which the covariance structure of the response functions is estimated in addition to the expected value structure. The multidimensional covariance surface is estimated by covariance smoothing, which makes the model particularly suitable for irregularly/sparsely observed functions. The use of (nested/crossed/curve-specific) functional random intercepts allows for modeling longitudinal/hierarchical study designs, such as in our analysis of movement trajectories of billiard players who execute a given shot several times and on several days. For this purpose, the joint covariance is decomposed into its independent variation components. Using multivariate functional principal component analysis, the covariance structure can be taken into account in the model fit (including cross-correlations between the dimensions) and the individual modes of variation can be interpreted.

b) In many data scenarios, the coordinate system in which each multidimensional curve is recorded is arbitrary and not of interest. Thus, the actual object of analysis is the shape of a curve, i.e. its equivalence class under translation, rotation and scaling. We extend FAMs to shapes of plane curves as response object taking the Riemannian manifold structure of the shape space into account: the mean shape is modeled by a geodetic response function, and residuals and distances are determined by the shape geometry. For model estimation, we propose a Riemannian L^2-Boosting algorithm and establish a new visualization for FAMs based on suitable tensor product model factorization, which allows to systematically interpret estimated model effects graphically even in the multidimensional-functional case.

c) Complementary to b), curves with fixed orientation and size are considered in this contribution – but as equivalence classes with respect to reparameterization ("warping"). Based on the Square-Root-Velocity (SRV) framework, we develop methods to model corresponding Fréchet means of irregularly/sparsely observed curves using splines and show identifiability statements for individual spline representations. Underlying "elastic" distances involve optimal parameterization alignment of one curve to another ("registration"). Furthermore, we illustrate the use of elastic distances for classification and clustering on datasets of irregularly observed curves.

d) Starting from c), we propose elastic full Procrustes analysis of shapes of

curves. The distance between curves underlying the notion of mean shape here is obtained by optimal rotation and scaling alignment in addition to parameterization alignment as in c). While c) deals with $m \geq 2$ dimensional curves in general, we restrict ourselves here as in b) to plane curves which can be understood as (equivalence classes of) complex-valued functions. Besides relying on c), this allows us to base analysis of irregularly/sparsely observed curves on Hermitian covariance smoothing, which we propose as a generalization of symmetric covariance smoothing (e.g. as in a)).

e) In the dissertation, we also present a model extension of the FAMs from b), in which curves are considered as invariant under re-parameterization on the basis of c) in addition to the previous shape invariances. In contrast to d), curve shapes are modeled here in dependence on covariates, and not based on the full Procrustes distance but based on the Riemannian distance in the shape space.

The aim of this work is to make the variety of possibilities offered by statistical modeling for scalar data also available for the analysis of object data. It is characteristic for object data analysis that generalizations to functions, densities, forms or other data types always also promote mathematical abstraction, show methodical similarities and give rise to new developments beyond the concrete data structure.

Zusammenfassung

Das Zusammenspiel geometrischer und probabilistischer Anschauungsweisen in der Statistik wird schon am Beispiel linearer Regression deutlich. Besonders explizit wird es aber, wenn es darum geht, statistische Methoden auf Objektdaten mit nicht-euklidischer Struktur zu erweitern. Funktionale Daten stellen einen solchen Datentyp dar, bei dem eine Stichprobe von Funktionen, wie bspw. Wachstumskurven oder Bewegungstrajektorien, als solche aufgefasst und analysiert wird.

Ausgangspunkt dieser Arbeit bilden auf Basis (generalisierter) additive Regression etablierte Modelle für funktionale Daten mit flexiblen Tensorprodukt-Spline-Effekten. In diesem Zusammenhang gehen wir auf verschiedene Szenarien mit funktionalen Zielgrößen ein, die eine Modellierung jenseits üblicher punktweiser Erwartungswertskurven erforderlich machen: I. Verteilungs-regressionsmodelle, bei denen bspw. auch die Varianzfunktion in Abhängigkeit von Kovariablen modelliert wird; II. Modelle für Wahrscheinlichkeitsdichten als funktionale Zielgröße; und III. Modelle für mehrdimensionale Kurven oder deren Formen. Methodische Erweiterungen in Bezug auf diese drei Aspekte werden in insgesamt sieben Teilprojekten vorgeschlagen und auf verschiedene meist biometrische aber auch ökonometrische Fragestellungen angewandt:

I. Schon am Beispiel von Wachstumskurven, einem Archetypen funktionaler Daten, wird klar, dass sich die (implizite) Annahme einer punktweise normalverteilten Zielgröße als problematisch erweisen kann (rein positive, oft schiefe Verteilung) und sich die ausschließliche Modellierung der Erwartungswertkurve in Abhängigkeit von Kovariablen mitunter als restriktiv darstellt (bspw. wenn Modifikationen am Medium von Bakterienkulturen deren Wachstumsprozess beeinflussen). Um solchen Herausforderungen zu begegnen, erweitern wir funktionale additive Modelle (FAM) auf Verteilungsregression (GAMLSS). Für die Modellschätzung auf Basis von Gradient-Boosting illustrieren wir dabei, welchen entscheidenden Beitrag geeignete Regularisierung im Umgang mit Autokorrelation funktionaler Zielgrößen leistet (Penalisierung in jedem Schritt und frühzeitiges Stoppen des Algorithmus auf Basis kurvenweiser Kreuzvalidierung). Durch den flexiblen Ansatz können wir bei der Modellierung eines Interaktionsszenario zweier Bakterienstämme auf experimentelle Spezifika eingehen und verschiedene Phasen des bakteriellen Wettstreits herauszustellen.

II. Im Gegensatz zur Erweiterung von (punktweisen) Verteilungsmodellen auf funktionale Zielgrößen in I. lassen sich umgekehrt auch Verteilungen selbst als Objekt funktionaler Datenanalyse auffassen. Entsprechend erweitern wir FAM auf Wahrscheinlichkeitsdichten als Zielgrößen. Aufgrund ihrer spezifischen Eigenschaften werden Dichten dabei in einem Bayes-Hilbert-Raum modelliert. Auch hier verwenden wir einen Boosting-Algorithmus zu Schätzung, mit dem Ziel erwartete quadratische Abstände im Bayes-Raum zu minimieren. In einer Analyse Gender-basierter Einkommensunterschiede auf Basis des Sozio-Ökonomischen

Panels (SOEP) modellieren wir damit die Verteilung der Einkommensanteile der Frau am Gesamteinkommen von Paaren als stetige Dichte mit Punktmassen bei 0 und 1 in Abhängigkeit verschiedener Einflussgrößen.

III. Die Modellierung mehrdimensionaler Kurven entspricht zunächst einer multivariaten Erweiterung reellwertiger funktionaler Daten (Teilprojekt a)), birgt aber auch weitere Herausforderungen: betrachtet man z.B. die Umrisskurve eines Objekts, bspw. eines Knochens, so spielen oftmals die räumliche Ausrichtung (b), d), e)) oder Parameterisierung der Kurve (c), d), e)) keine Rolle, was in der Modellierung berücksichtigt werden sollte. Obwohl die Teilprojekte teils stark aufeinander aufbauen, setzen sie stets auch individuelle Akzente:

   a) Im ersten Teilprojekt schlagen wir ein multivariates funktionales gemischtes Modell vor, in dem neben der Erwartungswertstruktur auch die Kovarianzstruktur der Zielfunktionen geschätzt wird. Die multidimensionale Kovarianzoberfläche wird dabei über Kovarianzglättung geschätzt, wodurch sich das Modell insbesondere auch für irregulär/spärlich beobachtete Funktionen eignet. Der Einsatz (genesteter/ gekreuzter/ kurvenspezifischer) funktionaler zufälliger Intercepts erlaubt die Modellierung longitudinaler/ hierarchischer Studiendesigns, wie bspw. in unserer Analyse von Bewegungstrajektorien von Billardspielern, die einen vorgegebenen Stoß mehrmals und an mehreren Tagen ausführen. Dazu wird die gemeinsame Kovarianz bei der Schätzung in deren unabhängige Variationskomponenten zerlegt. Mithilfe multivariater funktionaler Hauptkomponentenanalyse lässt sich die Kovarianzstruktur so in der Modellanpassung berücksichtigen (inklusive Kreuzkorrelationen zwischen den Dimensionen) und die einzelnen Variationskomponenten anschaulich interpretieren.

   b) In vielen Datenszenarien ist das Koordinatensystem, in dem jede einzelne mehrdimensionale Kurven aufgezeichnet wird, arbiträr und nicht von Interesse. Das eigentliche Objekt der Analyse bildet damit die Form einer Kurve, d.h. deren Äquivalenzklasse unter Translation, Rotation und Skalierung. Wir erweitern FAM auf Formen planarer Kurven als Zielgröße. Dabei wird die Riemannsche Mannigfaltigkeitsstruktur des Formraumes berücksichtigt: die erwartete Form wird über eine geodätische Response-Funktion modelliert und Residuen und Abstände werden entsprechend der Geometrie definiert. Zur Modellschätzung schlagen wir einen Riemannschen L^2-Boosting-Algorithmus vor und etablieren eine neue Visualisierung für FAM auf Basis geeigneter Tensorprodukt-Modell-Faktorisierung, die es auch im multidimensional-funktionalen Fall erlaubt, geschätzte Modelleffekte systematisch graphisch zu interpretieren.

   c) Komplementär zu b) werden in diesem Teilprojekt Kurven mit fester Ausrichtung und Größe – aber als Äquivalenzklassen bezüglich Umparameter-

isierung ("Warping") betrachtet. Auf Basis des Square-Root-Velocity (SRV) Framework entwickeln wir Methoden, um entsprechende Fréchet-Mittel irregulär/spärlich beobachteter Kurven mithilfe von Splines zu modellieren und zeigen Identifizierbarkeitsaussagen für einzelne Spline-Repräsentationen. Zugrundeliegende "elastische" Distanzen beinhalten die optimale Anpassung der Parameterisierung einer Kurve an eine andere ("Registrierung"). Darüberhinaus illustrieren wir auch den Einsatz elastischer Distanzen für Klassifikation und Clustering auf Datensätzen irregulär beobachteter Kurven.

d) Ausgehend von c) schlagen wir eine elastische Voll-Prokrustes-Analyse der Formen von Kurven vor. Die Distanz zwischen Kurven, auf deren Basis dabei eine mittlere Form bestimmt wird, ergibt sich neben Anpassung hinsichtlich Umparameteridierung wie in c) nun auch durch bestmögliche Rotation und Umskalierung. Während c) allgemein $m \geq 2$ dimensionale Kurven behandelt, beschränken wir uns hier wie in b) auf planare Kurven, die als (Äquivalenzklassen von) komplexwertigen Funktionen aufgefasst werden können. Das ermöglicht uns für die Analyse spärlich/irregulär beobachteter Kurven neben c) auch auf hermitesche Kovarianzglättung zurückzugreifen, die wir als Verallgemeinerung symmetrischer Kovarianzglättung (wie bspw. in a)) vorschlagen.

e) Im Rahmen der Dissertation stellen wir auch eine Modellerweiterung des FAM aus b) vor, in der auf Basis von c) Kurven neben den bisherigen Form-Invarianzen auch als invariant unter Umparameterisierung betrachtet werden. Im Unterschied zu d) werden die Kurvenformen hier in Abhängigkeit von Kovariablen modelliert, und nicht auf Basis der Voll-Prokrustes-Distanz sondern der Riemannschen Distanz im Formraum.

Ziel der Arbeit ist es, die Fülle an Möglichkeiten, die statistische Modellierung für skalare Daten bietet, ein weiteres Stück mehr auch für die Analyse von Objektdaten bereitzustellen. Dabei ist bezeichnend für Objektdatenanalyse, dass die Erweiterung auf Funktionen, Dichten, Formen oder andere hier nicht behandelte Datentypen, immer auch die mathematische Abstraktion befördert, methodische Gemeinsamkeiten aufzeigt und über die konkrete Datenstruktur hinaus Anstoß zu neuen Entwicklungen gibt.

Contents

1. Introduction

1.1. Overview

Classical functional response regression models mean functions $\mu_i : \mathcal{T} \to \mathbb{R}$ underlying a sample of response functions $y_i : \mathcal{T} \to \mathbb{R}$, $i = 1, \ldots, n$, given respective covariates $\mathbf{x}_i$, in such a way that they can be interpreted as point-wise conditional means

$$\mathbb{E}(Y(t) \mid \mathbf{x}_i) = \mu_i(t) = h(\mathbf{x}_i)(t) \qquad (t \in \mathcal{T})$$

of an underlying process Y in dependence on a functional predictor $h(\mathbf{x}_i)$ [compare, e.g., 133]. This thesis addresses data problems that require going beyond such point-wise functional mean regression in different ways, while at the same time aiming at the modeling flexibility we are used to from semi-parametric modeling. The thesis cumulatively comprises seven self-standing contributions that can be subdivided into two parts:
In Part I, the contribution in Chapter 2 discusses distributional regression for functional responses, where besides μ_i also, say, variance functions $\sigma_i^2 : \mathcal{T} \to \mathbb{R}$ can be modeled in dependence on covariates, allowing also for point-wise response distributions beyond the Gaussian case or exponential families. The contribution of Chapter 3, by contrast, considers functional response regression, where each response function y_i presents a density function describing a probability distribution.
In Part II, multidimensional curves y_i and their shape are considered. This comprises mixed models for multivariate functional data in nested and crossed sampling designs (Contribution Chapter 4), functional response models for shape and form manifolds of plane curves under invariance with respect to translation, rotation, and potentially scaling (Contribution Chapter 5), elastic splines models for mean curve estimation under re-parameterization ("warping") invariance (Contribution Chapter 6), as well as elastic full Procrustes mean estimation (Contribution Chapter 7) and elastic additive regression (Contribution Chapter 8) for shapes of plane curves under both aforementioned types of invariances.
All contributions are the result of joint work with other authors. The personal role of the author of this thesis within the respective project is declared at the beginning of each Chapter. Here, also information concerning the original publication of respective articles is provided.

In the remainder of the introductory Chapter 1, relevant context is provided in preparation of the contributions. This includes brief literature overviews and discussion of basic concepts in object data analysis (Section 1.2), concerning functional data in general (in 1.2.1), density and compositional data (in 1.2.2), time-warping of functional data (in 1.2.3), and shape data (in 1.2.4), as well as bibliographic references to main semi-parametric regression frameworks for scalar data in Section 1.3, illustrating some of their modeling flexibility by an example.

1.2. Modeling object data: from $\mathbb{L}^2$ functions to probability densities to shapes of curves

Analysis of data in a non-Euclidean space $\mathcal{Y}$, beyond categorical variables or metric variables commonly addressed in scalar or multivariate statistical analysis, has been summarized as "object oriented data analysis" [207, 123] in a similar spirit as "geometric deep learning" [22, 29] in the context of neural networks and the earlier notion of "abstract inference" [70]. The terminology highlights similar perspectives adopted in different branches of statistics analyzing samples $y_1, \dots, y_n \in \mathcal{Y}$ of such *object data*, including analysis of functional data [154], directional and spherical data [121], compositional data [143] and, more generally, probability distributions as data objects [148], shape data [49] or object deformations [58, 193, 132], data on manifolds of symmetric positive-definite (SPD) matrices [9, 115] and of graph Laplacian matrices [171] as well as Grassmannian manifolds [79], or tree spaces [78, 207, 12] and graph spaces [28] which carry no manifold structure. While more and more established in various fields, many developments in the analysis of object data have been accompanied by biomedical research [175, 127] aiming to describe complex "real world patterns" [71], such as functional data analysis of growth curves [133], anatomical shape analysis and morphometrics [105, 159], or statistical analysis of SPD matrices in diffusion-tensor-imaging [59] and in functional connectivity studies [219]. We may identify two main directions of generalization from the Euclidean case of finite-dimensional vector spaces $\mathbb{R}^k$: *functional* generalizations developed in functional data analysis consider object data as elements of infinite-dimensional Hilbert (or Banach) spaces [86] and arise when observations are naturally understood as a sample of functions (although practically recorded at discrete evaluations); *geometric* generalizations analyze object data in non-linear spaces, which are often endowed with a Riemannian manifold geometry [145] but also include other metric spaces that may still admit some local vector space approximation such as tree spaces [19] and Wasserstein spaces [32] facilitating geometric understanding and transfer of statistical tools from Euclidean spaces. Geometric approaches are often required due to non-linear constraints to the data objects, such as in probability density functions or SPD matrices, or because they present elements of a quotient space, such as in Kendall's shape spaces [98] or graph spaces. Conversely, object space geometries are often motivated from representations in Euclidean spaces. Different scenarios demand for combining functional and geometric object data analysis [178, 114, 200, 148, 32, 58], as is the case in most data scenarios discussed as part of this thesis.

Once the mathematical structure of the object data space $\mathcal{Y}$ is given, instruments of data analysis can be established: most fundamentally, this includes notions of a mean object $\hat{\mu} \in \mathcal{Y}$ of a sample of objects $y_1, \dots, y_n$ and tools for quantifying and visualizing the variation structure in the data. Here, Fréchet means [61, 226], generalizing the method of least-squares to a (semi-)metric space $(\mathcal{Y}, d)$ by setting $\hat{\mu} \in \arg\min_{\mu \in \mathcal{Y}} \sum_{i=1}^n d^2(\mu, y_i)$ (which can, however, not always be expected to be

unique), play a prominent role. Generalizations of principal component analysis (PCA) [172, 60, 207, 32, 194] present a key tool to determine modes of variation in often high-dimensional object data. Distance and mean computation as well as employing Euclidean data representations obtained from PCA for object data analysis already facilitate transfer of various methods of data analysis from multivariate analysis [77]. For instance, the author also participated in projects on classification of movement trajectories as functional data [112] and time-series modeling of yield curves [17] aside of the presented contributions, while this thesis focuses on regression for functional object data as response. While related to the scope of object data analysis described above, manifold learning [120] and metric learning [211, 106, 100] approach the data from the opposite direction. They do not explicitly define the geometry of the space of the object but estimate it from the data instead.

The following sections provide an overview over the object data types relevant for this thesis from a modeling perspective, pointing out their various interconnections.

1.2.1. Functional data analysis between point-wise and object-oriented perspectives

After pioneer works by Karhunen [96], Loève [117], Grenander [69], and Rao [156] in the middle of the 20$^{\text{th}}$ century and establishing as statistical discipline in the 80s and early 90s [70, 152, 153] at the latest with the first edition of Ramsay and Silverman [154] in 1997, functional data analysis (FDA) has become a very active field of statistical research relying on a rich body of theoretical and applied literature, as outlined in different FDA reviews [41, 208, 8]. Textbooks providing introductions [154, 104] and discussing theoretical foundations [86] and inference [80] in FDA document parts of its developments. References to other textbooks and surveys as well as overviews over the different directions of FDA can be found in the review papers given above. Here, some very basic ideas are described to prepare what follows.

A basic motive in the analysis of a sample of functions, say $y_i : [0,1] \to \mathbb{R}$ for $i = 1, \ldots, n$, is the ambivalence between *point-wise* and *object-oriented* perspectives: point-wisely, y_i is determined by evaluations $y_i(t)$ and typically observed as a vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{ik_i})^\top = (y_i(t_{i1}), \ldots, y_{ik_i}(t_{ik_i}))^\top \in \mathbb{R}^{k_i}$ of evaluations at points $\mathbf{t}_i = (t_{i1}, \ldots, t_{ik_i})^\top \in [0,1]^{k_i}$ which are often referred to as "time-points" as in the classic FDA example of growth curves. For simplicity, $\mathbf{y}_\imath$ and $\mathbf{y}_i$ for $\imath \neq i$ can always be assumed independent throughout the introduction. In particular regarding *regularly* sampled functional data, with time-grids $\mathbf{t}_1 = \cdots = \mathbf{t}_n$ equal for all observations, authors outside of FDA take this perspective when analyzing datasets that could be considered functional data by referring solely to vectors $\mathbf{y}_1, \ldots, \mathbf{y}_n$ of multivariate data. In *irregularly* and *sparsely* sampled functional data, with time-grids $\mathbf{t}_i$ varying over i and potentially small numbers k_i of sampling points per curve, the demand for explicitly considering variation over $t \in [0,1]$ is especially evident. Still focusing on a

point-wise perspective in such data, longitudinal data analysis [46] is highly related and interwoven with FDA. Mathematical consideration of y_i, $i = 1, \dots, n$ as realized paths of a stochastic process Y over the domain $[0, 1]$ yields the transition of a point-wise perspective, understanding it as a family of random variables $\{Y_t\}_{t\in[0,1]}$, to an object-oriented perspective, considering random paths $t \mapsto Y_t$ as data objects, which are often denoted by $Y(t) = Y_t$ in FDA for simplicity. Finally and in contrast to other approaches to stochastic processes [170], the object-oriented perspective is solidified by assuming a specific structure on the space $\mathcal{Y}$ of its paths. Here, a classic assumption is that y is square-integrable with respect to the Lebesque measure ν plus some further regularity assumtions to make Y a random element in the Hilbert space $\mathbb{L}^2([0, 1])$ with inner product $\langle y_1, y_2\rangle = \int y_1\, y_2\, d\nu$ (for details see [86, Chapter 7]). Conversely, a purely object-oriented perspective typically starts from viewing Y as a random element of a separable Hilbert space $\mathcal{Y}$. For various reasons, different authors focus more on an object-oriented or a point-wise approach. Perhaps it is, however, fair to say that bridging the two perspectives is an essential part of the identity of FDA. In the following, the interplay of the different perspectives is illustrated in more detail considering the mean and variance structure of Y. For theoretical aspects, we follow excerpts of Hsing and Eubank [86, Chapter 7] to which we refer for details.

The mean and variance structure of Y can be described via their point-wise mean function $\mu(t) = \mathbb{E}(Y(t))$ and covariance function $C(s, t) = \mathbb{E}((Y(s)-\mu(s))(Y(t)-\mu(t)))$ for $s, t \in [0, 1]$ – or, assuming Y a random element in the Hilbert space $\mathbb{L}^2([0, 1])$, via the mean element $m \in \mathbb{L}^2([0, 1])$ and covariance operator $\Sigma : \mathbb{L}^2([0, 1]) \to \mathbb{L}^2([0, 1])$ in an object-oriented approach. m and Σ are defined to fulfill $\langle m, y\rangle = \mathbb{E}(\langle m, Y\rangle)$ and $\langle\Sigma(y_1), y_2\rangle = \mathbb{E}(\langle Y - m, y_1\rangle\langle y_2, Y - m\rangle)$ for all $y, y_1, y_2 \in \mathbb{L}^2([0, 1])$ assuming that $\mathbb{E}(\|Y\|^2) < \infty$. Now, if μ and C are both continuous functions, μ coincides with m and $\Sigma(y)(t) = \int C(s, t)y(s)\, d\nu(s)$ coincides with the integral operator associated with C [86, Theorem 7.4.3] which we assume in the following. This also holds, for instance, when Y is restricted to a finite-dimensional subspace. Hence, the point-wise and object-oriented perspectives align.
When it comes to practical data analysis, point-wise data representation as evaluation vectors $\mathbf{y}_i$ is complemented by basis representations. To get from observed evaluation vectors $\mathbf{y}_i$ to functional representations $\hat{y}_i : [0, 1] \to \mathbb{R}$ that approximate underlying functions y_i for evaluation at arbitrary $t \in [0, 1]$, a basis representation approach nicely fits the geometry of a separable Hilbert space $\mathcal{Y}$, since in this case $\mathcal{Y}$ admits complete orthonormal systems $\{f_l\}_{l=1}^{\infty} \subset \mathbb{L}^2([0, 1])$ to represent each $y_i = \sum_{l=1}^{\infty}\langle f_l, y_i\rangle f_l$. Given finite data and resources, we can, however, only rely on a finite basis $\mathbf{f} = (f_l, \dots, f_L)^\top$ in practice, which has to be capable of reflecting the essential variation in Y. Typically employed bases include smoothing splines, B-splines and other function bases used for smoothing [154, 212] as well as wavelets [134] especially for spiky functional data. These bases are not necessarily orthonormal, which we assume, however, without loss of generality in the following to simplify expressions.

Estimating functional means in practice: Representing each functional observation by $\hat{y}_i(t) = \sum_{l=1}^{L} \check{y}_{il} f_l(t) = \check{\mathbf{y}}_i^\top \mathbf{f}(t)$ expanded in the basis $\mathbf{f}$ with coefficients $\check{\mathbf{y}}_i = (\check{y}_{i1}, \dots, \check{y}_{iL})^\top \in \mathbb{R}^L$ obtained by fitting evaluations $\mathbf{y}_i$ point-wisely with respect to (penalized) least squares [154], the mean μ can be estimated simply as $\hat{m} = \check{\boldsymbol{m}}^\top \mathbf{f}$ with the mean coefficient vector $\check{\boldsymbol{m}} = \frac{1}{n} \sum_{i=1}^{n} \check{\mathbf{y}}_i$. This is suitable when the data curves y_i are sampled densely enough to achieve good approximations $\hat{y}_i \approx y_i$. Alternatively, the μ might be estimated as $\hat{\mu}(t) = \check{\boldsymbol{\mu}}^\top \mathbf{f}(t)$ on a point-wise basis by jointly minimizing a (penalized and weighted) least squares criterion $\check{\boldsymbol{\mu}} = \arg\min_{\boldsymbol{\theta}} (\mathbf{y} - \mathbf{F}\boldsymbol{\theta})^\top \mathbf{W} (\mathbf{y} - \mathbf{F}\boldsymbol{\theta}) + \text{pen}(\boldsymbol{\theta})$ with $\mathbf{y} = (\mathbf{y}_1^\top, \dots, \mathbf{y}_n^\top)^\top$, a design matrix $\mathbf{F} = (\mathbf{f}^\top(t_{11}), \dots, \mathbf{f}^\top(t_{1L}), \dots, \mathbf{f}^\top(t_{nL}))^\top$ and, potentially, a suitable weight matrix $\mathbf{W}$ and penalty term $\text{pen}(\boldsymbol{\theta})$, reducing it to a scalar semi-parametric regression problem fitted on all evaluations. While both approaches are obviously highly related and $\hat{m} \approx \hat{\mu}$ should be close to identical in many cases, they still show some differences: if $L < k_i$, i.e. the basis dimension can be chosen smaller than the number of sampling points per curve, $\hat{m}$ can be expected to be computationally more efficient, especially if $\text{pen}(\boldsymbol{\theta})$ involves hyper-parameter tuning. In sparse scenarios with $L > k_i$ by contrast, estimation of individual basis expansions $\hat{y}_i$ can become unstable/biased and $\hat{\mu}$ can be expected to be more statistically efficient when irregular time grids $\mathbf{t}_i$ jointly achieve a better coverage of the domain. Note that non-parametric smoothing [55] can also be employed for point-wise mean estimation. However, in particular smoothing spline and kernel estimators are known to be highly related with semi-parametric approaches in this basic scenario [164, 111, 38, 97, 54, 213]. Hence, we focus on semi-parametric approaches which underly all contributions of the thesis, assuming a fixed basis – which is yet in practice selected with respect to the data problem at hand.

Estimating the covariance function: Analogously to the mean, estimation of the covariance function can be approached via the coefficients, employing the estimated covariance matrix $\check{\boldsymbol{\Sigma}} = \frac{1}{n-1} (\sum_{i=1}^{n} \check{\boldsymbol{y}}_i \check{\boldsymbol{y}}_i^\top) - \check{\boldsymbol{m}} \check{\boldsymbol{m}}^\top$ of their centered coefficients as coefficient matrix of an estimator $\hat{C}_{\text{cf}}(s,t) = \mathbf{f}^\top(s)\, \check{\boldsymbol{\Sigma}}_{\text{cf}}\, \mathbf{f}^\top(t)$ of $C(s,t)$ (where "cf" indicates estimation on coefficient level). Alternatively, $C(s,t)$ can be estimated point-wisely interpreting $\mathbb{E}((\tilde{Y}(s) - \mu(s))(\tilde{Y}(t) - \mu(s))) = C(s,t)$ as nonlinear regression problem with responses given by all products of evaluations $(y_i(s) - \hat{\mu}(s))(y_i(t) - \hat{\mu}(t))$, $s, t \in \mathbf{t}_i$, within all observations $i = 1, \dots, n$ and with s, t as covariates. In FDA, this *covariance smoothing* approach was proposed by Yao et al. [216] as part of their "principal component analysis through conditional expectation (PACE)" algorithm and became a key tool for irregular/sparse functional data. Covariance smoothing has widely been approached via least-squares estimation employing various smoothing techniques [216, 33, 110, 158, 167]. A semi-parametric approach using the tensor-product basis $\mathbf{f}(s) \otimes \mathbf{f}(t)$ as in [33] leads to an estimator $\hat{C}_{\text{pw}}(s,t) = \mathbf{f}^\top(s)\, \check{\boldsymbol{\Sigma}}_{\text{pw}}\, \mathbf{f}^\top(t)$ for $C(s,t)$ of the same form as $\hat{C}_{\text{cf}}(s,t)$ above, only that $\check{\boldsymbol{\Sigma}}_{\text{pw}}$ arises as estimated $L \times L$ basis coefficient matrix in point-wise (penalized) least squares (here "pw" indicates point-wise estima-

tion). Arguments in favor of each estimator are the same as for mean estimation, only much more drastic: For densely sampled functions, $\hat{C}_{\text{pw}}$ can take a substantial amount of time and memory due to quadratic increase of design matrix dimensions in L and k_i. For irregularly/sparsely sampled functions, we experienced serious issues in $\hat{C}_{\text{cf}}$ (even when regularizing with additional penalties) when $\hat{C}_{\text{pw}}$ still yields convincing results. In the contributions of Chapters 4 and 7, different generalizations of the point-wise estimator $\hat{C}_{\text{pw}}$ will facilitate functional principal component analysis also in multivariate and sparsely sampled data scenarios.

Functional principal component analysis: Point-wise inspection of $C(s,t)$ yields limited intuitive understanding of the variation structure. However, the importance of eigen decomposition of its associated covariance operator Σ based on Mercer's Theorem / the Karhunen-Loève Theorem (for details, please again refer to [86]) for FDA is hard to overestimate: A complete orthonormal system $\{e_l\}_{l=1}^{\infty}$ of $\mathbb{L}^2([0,1])$ (or another Hilbert space) such that $\lambda_l e_l = \Sigma(e_l)$ is called eigenbasis of Σ with eigenvalues $\lambda_l \geq \lambda_2 \geq \cdots \geq 0$ and always exists if $\mathbb{E}(\|Y\|^2) < \infty$. It allows to write

$$C(s,t) = \sum_{l=1}^{\infty} \lambda_l e_l(s) e_l(t)$$

with the sum converging absolutely and uniformly [86, Theorem 7.2.6/p. 187]. Moreover, the eigenbasis also reflects the variation structure of Y as

$$Y = \sum_{l=1}^{\infty} \langle e_l, Y \rangle \, e_l$$

with probability one, where the $Z_l = \langle e_l, Y \rangle$ are uncorrelated random variables with mean $\mathbb{E}(Z_l) = \langle e_l, \mu \rangle$ and variance λ_l [86, Theorem 7.2.7]. The eigenbasis would also yield optimal finite-dimensional representation of residuals $\varepsilon = Y - \mu$ in the sense that $\mathbb{E}(\|\varepsilon - \sum_{l=1}^{L} \langle f_l, \varepsilon \rangle f_l\|^2) \leq \mathbb{E}(\|\varepsilon - \sum_{l=1}^{L} \langle e_l, \varepsilon \rangle e_l\|^2)$ for all L and bases $f_1, \ldots, f_L$ [86, Theorem 7.2.8]. However, the eigenbasis is unknown in practice. Given the orthonormal basis $\mathbf{f}$ assumed above, an estimate of its first components can be obtained from covariance estimators of the form $\hat{C}(s,t) = \mathbf{f}^\top(s) \check{\mathbf{\Sigma}} \mathbf{f}^\top(s)$ directly as $\hat{e}_l(t) = \check{\mathbf{e}}_l^\top \mathbf{f}(t)$ with eigenvalues $\hat{\lambda}_l$, $l = 1, \ldots, \max\{L, n\}$, from eigen decomposition $\check{\mathbf{\Sigma}} = \check{\mathbf{E}} \hat{\mathbf{\Lambda}} \check{\mathbf{E}}^\top$ of the coefficient matrix with $\check{\mathbf{e}}_l$ the lth column of the orthonormal matrix $\check{\mathbf{E}}$ and $\hat{\lambda}_l$ the lth entry of the diagonal matrix $\hat{\mathbf{\Lambda}}$.
Estimated eigenfunctions $\hat{e}_l$ can then, inter alia, be used for further dimension reduction or as a visualization tool for illustrating and interpreting principal modes of variation in the data by plotting $\mu(t) \pm \delta \, e_l(t)$ for some $\delta > 0$ [154]. Moreover, eigen decomposition is used as second building block in PACE-type algorithms to predict scores $\langle e_l, y_i \rangle$ of sparsely sampled curves y_i based on a working normality assumption with the estimated covariance structure of the data [216]. Allowing also improved predictions $\hat{y}_i$ of the latent curves y_i, similar ideas were also applied to obtain smooth reconstructions

of partially observed functional data [45, 103]. A survey on developments in functional principal component analysis until 2014 is provided by Shang [172].

These basic building blocks and the interplay between point-wise and object-oriented perspectives will re-appear throughout the presented work.

1.2.2. Geometry of probability distributions and compositional data

From early on, statisticians have addressed the question of the geometry of the space of probability distributions. In his 1945 seminal paper introducing what would later be known as Cramer-Rao lower bound and laying ground for Rao-Blackwellization [155], Rao also proposes the *Fisher-Rao metric* as Riemannian metric on "the population space". Here, he refers back to Bhattacharrya [18] who already "defines the distance between population as the angular distance between two points representing the population on a unit sphere" by square-root transformation and points out the direct correspondence to the Fisher-Rao metric. Hotelling and Fisher arrived at similar geometrical ideas in the context of a paper on "Spaces of statistical parameters" of Hotelling, of which only an abstract and a summary are still existing while the manuscript was lost [183]. Besides its role in information geometry [6, 7], this yields a possible geometry for modeling probability distributions as object data [176].
In the following, we consider functional observations y_i, $i = 1, \dots, n$, as probability density functions of random variables T_i taking values in $[0, 1]$ (again with respect to the Lebesque measure ν for simplicity) for a comparative introduction to the Fisher-Rao metric and Bayes Hilbert spaces [200] in the light of object data analysis. Providing an overview of object data analysis of probability distributions, Petersen et al. [148] discuss the aforementioned approaches, Wasserstein space / quantile function based approaches [140, 32, 220, 27, 35, 63, 10, 221, 173, 51], other alternative transformations to Hilbert spaces [147] (in particular log-hazard and log-quantile density transforms) and approaches directly considering densities in $\mathbb{L}^2([0, 1])$ [141, 15, 44]. Another Hilbert space geometry of probability distributions of finite entropy was proposed by Newton [138].

The Fisher-Rao metric: Based on Bhattacharrya's representation, an analysis of y_i, $i = 1, \dots, n$, in the Fisher-Rao metric corresponds to interpreting the square-root densities $q_i : t \to \sqrt{y_i(t)}$ as points in the Hilbert sphere $\mathbb{S} = \{q \in \mathbb{L}^2([0, 1]) : \|q\| = 1\}$ borrowing its manifold geometry. This implements the integral-one constraint of a probability density as $\int y_i \, d\nu = \|q_i\|^2 = 1$ and yields distances

$$d_{FR}(y_1, y_2) = \cos^{-1}(\langle q_1, q_2 \rangle) = \cos^{-1}\Big(\int \sqrt{y_1}\sqrt{y_2} \, d\nu\Big)$$

of densities measuring arc-lengths on the unit sphere.
One remarkable property of d_{FR} is that, for densities $\tilde{y}_i$ of random variables $\tilde{T}_i = \gamma(T_i)$

jointly transformed with a diffeomorphism $\gamma : [0,1] \to [0,1]$, distances

$$d_{FR}(\tilde{y}_1, \tilde{y}_2) = \cos^{-1}\Big(\int \sqrt{\frac{y_1(\gamma^{-1}(t))}{|\dot{\gamma}(t)|}} \sqrt{\frac{y_2(\gamma^{-1}(t))}{|\dot{\gamma}(t)|}}\, d\nu(t)\Big) = d_{FR}(y_1, y_2)$$

are preserved. Hence, the geometry is, for instance, compatible with switching between a strictly positive random variable T and $\log(T)$, as often done in statistical modeling. Moreover, this property plays a key role in re-parameterization invariant "elastic" analysis of functional data and shapes [178], which will be outlined in more detail later. The Fisher-Rao metric has been used by different authors for object data analysis of probability densities [176, 64, 187, 166, 42]. Without pointing out the connection, a square-root representation was employed for compositional data [181, 182, 210, 168], which is directly related as outlined in the following.

Compositional data analysis considers data problems where one observation vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{ik})^\top$ reflects the composition of a whole by its shares $y_{ir} \geq 0$, $r = 1, \ldots, k$, (or strictly $y_{ir} > 0$) whereby proportional compositions $\mathbf{y_i} \approx \lambda\, \mathbf{y_i}$, with $\lambda > 0$, are considered equivalent or of a fixed "size" by design, say $S_1(\mathbf{y}_i) = \sum_{r=1}^k y_{ir} = 1$. While geosciences present a classical field of application [157, 25], analyzing for instance the composition of soil samples (e.g. in sand, clay, silt shares [57]), other applications also include analysis of time-use data [182], chemometric [203] or morphometric data [161] or, more recently, microbiome data [39], as well as other data types [142]. Usually normalized to $S_1(\mathbf{y}_i) = 1$ without loss of generality, a compositional data sample $\mathbf{y}_i$ mathematically corresponds to a vector of discrete probabilities of events $\mathbf{t} = (t_1, \ldots, t_k)^\top$. In contrast to the spherical geometry for probability distributions above, today's compositional data analysis literature, however, typically considers $\mathbf{y}_i$ in the flat Aitchison geometry [4] on the open simplex $\Delta^k = \{\mathbf{y} = (y_1, \ldots, y_k)^\top : y_1 > 0, \ldots, y_k > 0, \sum_{r=1}^k y_r = 1\}$ of strictly positive probabilities, as outlined in different textbooks on the topic [57, 143, 201, 142]. This corresponds to considering *centered-log-ratio* transforms $\mathrm{clr}(\mathbf{y}_1), \ldots, \mathrm{clr}(\mathbf{y}_n)$ of the original data, given by

$$\mathrm{clr}(\mathbf{y}_i) = \Big(\log(y_{i1}) - \frac{1}{k}\sum_{r=1}^k \log(y_{ir}), \ldots, \log(y_{ik}) - \frac{1}{k}\sum_{r=1}^k \log(y_{ir})\Big)^\top \tag{1.1}$$

in an Euclidean space, mapping Δ^k to $\mathbb{R}_0^k = \{\mathbf{y} \in \mathbb{R}^k : \mathbf{y}^\top \mathbf{1}_k = 0\}$ orthogonal to the constant $\mathbf{1}_k = (1, \ldots, 1)^\top \in \mathbb{R}^k$ vector. Instead of borrowing the geometry of a sphere after square-root velocity transformation above, an analysis in the Aitchison geometry can be effectively carried out on the linear subspace $\mathbb{R}_0^k \subset \mathbb{R}^k$ with common tools of multivariate analysis. An overview over different subspace coordinates facilitating different interpretations can be found in [57, Chapter 3.3].
The Aitchison geometry reflects relative differences in the composition proportions with distances given by

$$d_A(\mathbf{y}_1, \mathbf{y}_2) = \sqrt{\frac{1}{2k}\sum_{r=1}\sum_{l=1} \log\Big(\frac{y_{1r}/y_{1l}}{y_{2r}/y_{2l}}\Big)^2}$$

depending on "odds ratios" between the components of $\mathbf{y}_1$ and $\mathbf{y}_2$. Moreover, an analysis of compositional data based on the Aitchison geometry is in accordance with the *principle of subcompositional coherence*, claiming that results of analyses of a subset of the compositions $\mathbf{y}_i' = (y_{ir_1}, \ldots, y_{ir_{k'}})^\top$ with $\{r_1, \ldots, r_{k'}\} \subset \{1, \ldots, k\}$ should not contradict conclusions from the entire set $\mathbf{y}_i$ of compositions (e.g. [68] or [57, Chapter 1.3]). This is desirable in two opposite directions: firstly, it allows to reduce large sets of components to smaller subsets for interpretation; and secondly, in many cases, the set of measured components is not fully exhaustive or can be viewed as part of a larger whole. Conversely, the restriction to components $y_{ir} > 0$ is sometimes considered a limitation of analysis in the Aitchison geometry and has motivated work on handling zeros in this framework [125, 124, 202, 204]. Following Mosimann's 1970 work [135] comparing size-independent statistical analysis with different size variables including $S_1(\mathbf{y}_i)$, $S_2(\mathbf{y}_i) = \sqrt{\sum_{r=1}^k y_{ir}^2}$ and $S_{geom}(\mathbf{y}_i) = (\prod_{r=1}^k y_{ir})^{1/k}$, a "log-shape-ratio (LSR)" approach has found independent application in the morphometrics literature [149] effectively working with clr-transforms.

Bayes Hilbert spaces , first proposed by Egozcue et al. [53], generalize the ideas of the Aitchison geometry for discrete $\mathbf{y}_i$ to probability measures with density functions y_i [199, 200]. In analogy, the Bayes Hilbert space geometry may be implemented by representing densities $y_1, \ldots, y_n$ as their centered-log-ratio transforms

$$\operatorname{clr}(y_i)(t) = \log(y_i(t)) - \frac{1}{\nu([0,1])} \int \log(y_i(t))\, d\nu(t) \tag{1.2}$$

borrowing the subspace geometry of $\mathbb{L}_0^2([0,1]) = \{q \in \mathbb{L}^2([0,1]) : \int q\, d\nu = 0\}$,. To this end, it is assumed that $y_i(t) > 0$ for (ν-almost) all $t \in [0,1]$ and $\int \log(y_i(t))^2\, d\nu(t) < \infty$. Thus, implementations for FDA in Bayes Hilbert spaces can heavily rely on methods developed for functional data in $\mathbb{L}^2([0,1])$. Although $\nu([0,1]) = 1$ in our case, we keep the factor in (1.2) to illustrate the general expression. For instance, the discrete clr defined in (1.1) above is included as special case with ν the counting measure.
Interpreting the "subcomposition" $y_i\big|_{\mathcal{U}} : \mathcal{U} \to \mathbb{R}$ as kernel of a conditional probability density on $\mathcal{U} \subset [0,1]$, which is equivalent to the conditional density in the Bayes Hilbert space geometry, suggests restating the principle of subcompositional coherence in terms of conditional probabilities: the results of an analysis of the distributions of random variables $T_1, \ldots, T_n$ conditional on $T_i \in \mathcal{U}$, for some measurable set $\mathcal{U}$ and all $i = 1, \ldots, n$, should not contradict conclusions on their unconditional distributions. Although the formulation of the principle itself and density analysis in this context certainly require further investigation, it can be assumed that, due to their relative nature, Bayes Hilbert spaces inherit a probabilistic notion of subcompositional coherence from the finite-dimensional case of the Aitchison geometry. In the contribution presented in Chapter 3, which addresses Bayes Hilbert space responses in regression, we present corresponding results that also facilitate different means of interpretation.

Literature on object data analysis in Bayes Hilbert spaces include work on principal component analysis [85] potentially employing alternative reference measures ν for weighting [190], regression with densities as covariates [188, 52] and responses [218, 189], as well as geostatistics [130, 131] and considerations of bivariate densities [84].

Just like different distribution assumptions in different data scenarios, different geometries for probability densities may be preferable in different scenarios and data problems as a modeling decision depending on their characteristic properties. Subcompositional coherence might motivate an analysis of densities in a Bayes Hilbert space. To make the notion more precise, *subcompositional dominance* [57, Chapter 3.3] is commonly considered as necessary condition for subcompositional coherence and can be formulated as follows for densities: Let $y_i' : \mathcal{U} \to \mathbb{R}$, $t \mapsto y_i(t) / \int_{\mathcal{U}} y_i \, d\nu$ denote a density conditional on a subset $\mathcal{U} \subset [0,1]$ of measure $\nu(\mathcal{U}) > 0$. Then, $d'(y_1', y_2') \leq d(y_1, y_2)$, i.e. the corresponding distance of the conditional densities does not exceed the distance of the entire distributions. While this holds for Bayes Hilbert spaces, it does not hold for the Fisher-Rao metric: a counter example with $d_{FR}(y_1, y_2) < d'_{FR}(y_1', y_2')$ can be easily constructed by choosing $y_1(t) = y_2(t)$ large enough for $t \in [0,1] \setminus \mathcal{U}$ to achieve

$$\int \sqrt{y}_1 \sqrt{y}_2 \, d\nu = \sqrt{\int_{\mathcal{U}} y_1 \, d\nu \int_{\mathcal{U}} y_2 \, d\nu} \int_{\mathcal{U}} \sqrt{y'}_1 \sqrt{y'}_2 \, d\nu + \int_{[0,1]\setminus\mathcal{U}} \sqrt{y}_1 \sqrt{y}_2 \, d\nu > \int \sqrt{y'}_1 \sqrt{y'}_2 \, d\nu'$$

with ν' the measure ν restricted to $\mathcal{U}$, such that, since $\cos^{-1}$ is strictly decreasing, subcompositional dominance is violated. Vice versa, the Bayes Hilbert space geometry is not invariant under variable transformation $\tilde{T}_i = \gamma(T_i)$ discussed above for the Fisher-Rao metric. Instead,

$$\| \operatorname{clr}(\tilde{y}_1) - \operatorname{clr}(\tilde{y}_2) \| = \| \operatorname{clr}(y_1) \circ \gamma^{-1} - \operatorname{clr}(y_2) \circ \gamma^{-1} \| = \| (\operatorname{clr}(y_1) - \operatorname{clr}(y_2)) \, |\dot{\gamma}| \|$$

which is similar to a change to reference measure $d\tilde{\nu} = |\dot{\gamma}| d\nu$ on the original densities (yet not identical, since $\int \operatorname{clr}_i(y_i) \, d\tilde{\nu} \neq 0$ in general). Hence, if compatibility with variable transformation is of primary interest, the Fisher-Rao metric presents a suitable choice for density data modeling. Moreover, the transformation $\tilde{T}_i = \gamma(T_i)$ translates to *time-warping* $\tilde{F}_i(t) = F_i(\gamma(t))$ of the respective cumulative distribution function F_i of y_i. This will be of great importance later and is discussed for general functional data in the next section.

1.2.3. Time-warping and registration of functional data

A Hilbert space perspective on functional data suggests an additive model of variation in the random variables $Y(t_{ir})$ underlying recorded functional data evaluations $y_i(t_{ir}) = y_{ir}$, $r = 1, \ldots, k_i$, $i = 1, \ldots, n$. This holds especially when working with a finite, truncated basis in practice. Besides this variability "in y-direction", various data problems, however, suggest also considering variability in t, such that temporal *registration* is required for point-wise comparison of two curves y_1 and y_2. An introduction

to the registration problem in FDA is given by Ramsay and Silverman [154, Chapter 7] and different approaches are systematically compared in Marron et al. [122].
Approaching from a point-wise perspective, variability in the time-points t_{ir} assigned to each curve evaluation y_{ir} might arise for different reasons: the t_{ir} might be subject to measurement error, for instance because of rounding or reporting issues, or present only a surrogate for true underlying original time-points $\tilde{t}_{ir} \in [0, 1]$. In this sense, temporal variability is related to measurement errors in the covariate of a scalar nonlinear model [31, Chapter 13]. From a functional perspective, the problem can be rephrased, assuming that functions $\tilde{y}_i$ of interest are randomly "warped" before observing

$$y_i(t) = \tilde{y}_i(\gamma_i(t)) \qquad (i = 1, \dots, n) \tag{1.3}$$

with latent *warping functions* $\gamma_i : [0, 1] \to [0, 1]$. While different classes of warping functions are considered in literature (affine [205], piecewise-linear and parametric [154], semi-parametric [214] also for incompletely observed functions [13], or non-parametric including all diffeomorphisms [179]), they are usually assumed monotonously increasing, such that the order of the time-points is preserved. From a temporal measurement error perspective, registration of functional data is predicting $\hat{\gamma}_i$ for γ_i, $i = 1, \dots, n$, to impute $\hat{t}_{ir} = \hat{\gamma}_i(t_{ir})$ for true unknown time-points $\tilde{t}_{i1}, \dots, \tilde{t}_{ik_i}$ of $y_{i1}, \dots, y_{ik_i}$. Beyond that, model (1.3) also implies a decomposition $y_i \mapsto (\tilde{y}_i, \gamma_i)$ into two functional features, which is practically implemented by registration, i.e. by predicting tuples $(\hat{y}_i, \hat{\gamma}_i) \approx (\tilde{y}_i, \gamma_i)$. Variability in $\tilde{y}_i$ is then often referred to as "amplitude variability", whereas variability in γ_i is referred to as "phase variability" [154] and may be of independent interest. This gives rise to considering warping functions $\gamma_1, \dots, \gamma_n$ as additional objects of data analysis.

While, in general, registration does not impose constraints on $\tilde{y}_i$, common constraints on warping functions open up a connection to data analysis of probability distributions: as monotonically increasing functions with $\gamma_i(0) = 0$ and $\gamma_i(1) = 1$, the γ_i have the same mathematical properties as a cumulative distribution function and their derivatives $\dot{\gamma}_i$ may be interpreted as density functions. Accordingly, they have been modeled based on the 2-Wasserstein metric [139, 30] and based on the Fisher-Rao metric [194] which Happ et al. [74] also compared to a Bayes Hilbert space approach and approaches of density analysis.

To allow full separation of amplitude and phase variability, Srivastava et al. [179, 177] propose a geometric approach using a warping-invariant, *elastic* metric such that distances $d(y_1 \circ \gamma, y_2 \circ \gamma) = d(y_1, y_2)$ are preserved when two curves are warped with the same γ. They discover the Fisher-Rao metric for this purpose and generalize Battacharrya's square-root transformation to potentially negative- and vector-valued curves. This yields the square-root-velocity (SRV) transform

$$q_i(t) = \frac{\dot{y}_i(t)}{\sqrt{|\dot{y}_i(t)|}} \quad \text{if well-defined and } q_i(t) = 0 \text{ otherwise}$$

where in the vector-valued case $|\cdot|$ generalizes to the length of a vector. In this representation, the $\mathbb{L}^2$ inner product $\langle \tilde{q}_1, \tilde{q}_2 \rangle = \langle q_1, q_2 \rangle$ is preserved when considering SRV-transforms $\tilde{q}_i = q_i \circ \gamma \sqrt{\dot{\gamma}}$ of jointly warped curves $\tilde{y}_i = y_i \circ \gamma$. The invariance allows to define a semi-metric on the space of curves modulo warping, i.e. on the equivalence classes $[y_i]_{\mathrm{w}} = \{y_i \circ \gamma : \gamma \in \Gamma\}$ with Γ the set of warping functions, by optimal warping-alignment as

$$d_e([y_1]_{\mathrm{w}}, [y_2]_{\mathrm{w}}) = \inf_{\gamma \in \Gamma} \|q_1 - q_2 \circ \gamma \sqrt{\dot{\gamma}}\| \tag{1.4}$$

with SRV-transforms q_i representing curves $[y_i]_{\mathrm{w}}$ up to a constant. Here, Γ is essentially the set of diffeomorphisms $[0,1] \to [0,1]$ (for details see [23]). Estimation of a Fréchet mean with respect to d_e implicitly also yields curve registration: to obtain a mean estimator $[\hat{\mu}]_{\mathrm{w}}$, registered curves $\tilde{y}_1, \ldots, \tilde{y}_n$ to a mean curve representative $\hat{\mu}$ are computed, as well as corresponding optimal warping functions $\hat{\gamma}_1, \ldots, \hat{\gamma}_n$ (approximation based on finite sampling grids and by numerical optimization, e.g., via dynamic programming [178, 16, 108, 75, 209]). While the variation of $\tilde{y}_1, \ldots, \tilde{y}_n$ around $\hat{\mu}$ reflects amplitude variability, Tucker et al. [195] use $\hat{\gamma}_1, \ldots, \hat{\gamma}_n$ to investigate also phase variability based on the Fisher-Rao metric, yielding in particular their empirical Fréchet mean $\bar{\gamma}$. Using elastic metrics for amplitude and phase, the geometry is fully compatible with joint warping γ of the representatives $\hat{\mu} \circ \gamma$, $\tilde{y}_i \circ \gamma$ and $\hat{\gamma}_i \circ \gamma$, $\bar{\gamma} \circ \gamma$. For visualization and principal component analysis, one may choose $\gamma = \bar{\gamma}^{-1}$ such that the new Fréchet mean of the warping functions is the identity.
Different authors propose Bayesian approaches in the SRV framework [36, 107, 118], also for noisy and fragmented curves [126]. Recently, also an approach for curve domains without fixed boundaries was proposed [24]. Xie et al. [215] propose functional data visualization based on amplitude and phase decomposition in the SRV framework. The SRV framework is used for elastic analysis of shapes of curves introduced below, and takes a fundamental role in Contributions 6, 7, and 8. An alternative approach to warping-invariant analysis of functional data was recently proposed by Pegoraro and Secchi [144] defining distances on tree representations.

1.2.4. Statistical shape analysis

In the spirit of D'Arcy Thompson's "On Growth and Form" [192], a milestone of mathematical biology with its first edition dating back to 1917 [89], the field of *morphometrics* quantitatively investigates variation in shape of biological organisms with statistical methods [159, 2]. While this endeavor addresses a multitude of statistical disciplines, such as for instance also compositional data analysis [161, 159], modern geometric morphometrics make extensive use of statistical shape analysis [1, 2] and morphometric applications are ubiquitous in shape analysis [49], which renders the two fields closely intertwined. To give a short introduction to basic concepts of *statistical shape analysis* [49, 178], we consider a constitutive example of a dataset of k *landmarks* $y_{i1}, \ldots, y_{ik}$ identifying points on $i = 1, \ldots, n$ objects of interest. These may, for instance, be

marked on images showing a particular perspective of an animal bone over different individuals in two dimensions [198, 149] or tracking sensors for human motion recognition in three dimensions [62, 81]. Here, we focus on the two dimensional, planar case allowing to consider $y_{ir} \in \mathbb{C} \cong \mathbb{R}^2$, $r = 1, \ldots, k$, which will simplify expressions and is focused on later. Hence, the shape of the ith object is represented by a vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{ik})^\top \in \mathbb{C}^k$ reflecting the landmark configuration. Since they are typically recorded in individual coordinate systems, analysis of the $\mathbf{y}_i$ should not depend on rotation and translation. Moreover, the shape of an object may be considered independent of its size for different reasons: when comparing the shape of body parts across patients/animals, body size might present a nuisance variable; effects of size and shape might be of seperate interest; or association of the shape of an object with its size is explicitly of interest in an "allometric" study [135] comparing, for instance, again individuals of different body size or age. Along the lines of Kendall [98], Ziezold [226] and Bookstein [20], statistical shape analysis investigates the *shape* of $\mathbf{y}_i$ as its equivalence class $[\mathbf{y}_i]_\mathrm{s} = \{\lambda \exp(\sqrt{-1}\omega)\mathbf{y}_i + z\,\mathbf{1}_k : \lambda > 0, \omega \in (-\pi, \pi], z \in \mathbb{C}\}$ invariant under translation by z, rotation by ω and scaling by λ. We refer to the quotient space of all such shapes as $\mathcal{Y}^*_{/\,\mathrm{Trl} \times \mathrm{Rot} \times \mathrm{Scl}}$. *Full Procrustes analysis* [65, 99] of planar shapes and *Kendall's shape space* geometry are briefly outlined in the following as cornerstones of statistical shape analysis, referring to Dryden and Mardia [49] for further details and shapes in more than two dimensions. This will prepare shape analysis of outline curves (as in the contribution in Chapter 5) and, finally, combine with the SRV framework to elastic shape analysis (underlying Chapter 7 and 8).

Full Procrustes analysis: Statistical shape analysis motivates the geometry on the quotient space of shapes $[\mathbf{y}_i]_\mathrm{s}$ from the Euclidean geometry on their representatives $\mathbf{y}_i$, which are centered and normalized to "pre-shapes" with centroid $\sum_{r=1}^k y_{ir} = 0$ and size $S_2(\mathbf{y}_i) = \|\mathbf{y}_i\| = \sqrt{\sum_{r=1}^k |y_{ir}|^2} = 1$ to eliminate translation and scale (excluding the $\mathbf{0}$ vector as degenerate special case). For notational simplicity, we assume the $\mathbf{y}_i$ already present such pre-shapes in the following. The *full Procrustes distance* on the shape space $\mathcal{Y}^*_{/\,\mathrm{Trl} \times \mathrm{Rot} \times \mathrm{Scl}}$, is defined by superimposition of two shape representatives $\mathbf{y}_1$ and $\mathbf{y}_2$ as

$$d_F([\mathbf{y}_1]_\mathrm{s}, [\mathbf{y}_2]_\mathrm{s}) = \min_{\lambda>0, \omega \in (-\pi,\pi], z \in \mathbb{C}} \frac{\|\mathbf{y}_1 - \lambda(\exp(\sqrt{-1}\omega)\mathbf{y}_2 + z\mathbf{1}_k)\|}{\|\mathbf{y}_1\|}$$

optimizing over all shape invariances. Here, normalization of $\mathbf{y}_1$ is important to obtain a proper metric. With $\mathbf{y}_1$ and $\mathbf{y}_2$ pre-shapes, this reduces to

$$d_F^2([\mathbf{y}_1]_\mathrm{s}, [\mathbf{y}_2]_\mathrm{s}) = 1 - |\mathbf{y}_1^\dagger \mathbf{y}_2|^2$$

where $\mathbf{y}^\dagger$ denotes the conjugate transpose of a complex vector $\mathbf{y}$ (compare, e.g., [49, Chapter 8]). Minimizing parameters are given by $\lambda^* = |\mathbf{y}_1^\dagger \mathbf{y}_2|$, by $\exp(\sqrt{-1}\omega^*) = \mathbf{y}_2^\dagger \mathbf{y}_1 / |\mathbf{y}_2^\dagger \mathbf{y}_1|$, and by $z = 0$ yielding the "full Procrustes fit". An explicit solution for the

full Procrustes mean shape $[\hat{\boldsymbol{\mu}}_F]_\mathrm{s}$, the Fréchet mean with respect to d_F, is given by

$$\hat{\boldsymbol{\mu}}_F = \operatorname*{arg\,min}_{\boldsymbol{\mu}:\|\boldsymbol{\mu}\|=1} n - \sum_{i=1}^{n} |\boldsymbol{\mu}^\dagger \mathbf{y}_i|^2 = \operatorname*{arg\,max}_{\boldsymbol{\mu}:\|\boldsymbol{\mu}\|=1} \sum_{i=1}^{n} \boldsymbol{\mu}^\dagger \mathbf{y}_i \mathbf{y}_i^\dagger \boldsymbol{\mu}^\dagger = \hat{\mathbf{e}}_1$$

with $\hat{\mathbf{e}}_1$ the leading eigenvector of the complex covariance matrix $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{k} \mathbf{y}_i \mathbf{y}_i^\dagger$ if its eigenvalue $\hat{\lambda}_1$ is of multiplicity 1 (otherwise all eigenvectors of $\hat{\lambda}_1$ define means). In more than two dimensions no explicit solution is available for $[\hat{\boldsymbol{\mu}}_F]_\mathrm{s}$. Assuming a *complex Bingham distibution* [99] with density

$$f(\mathbf{y}) \propto \exp(\mathbf{y}^\dagger \mathbf{A}\, \mathbf{y})$$

on the pre-shapes with $\mathbf{A}$ a Hermitian $k \times k$ matrix, $\hat{\boldsymbol{\mu}}_F$ yields a maximum likelihood estimate of the mode of the distribution [49, Chapter 10]. The complex Bingham distribution arises from a complex normal distribution [66] with zero mean by conditioning on norm one and presents a classic shape distribution [49].
Planar full Procrustes analysis will present the starting point for the contribution presented in Chapter 7, presenting an alternative to previous work considering intrinsic shape means in this context (and modeled also in Chapter 8).

Kendall's shape space and intrinsic shape means: A different notion of shape mean can be motivated from the Riemannian manifold geometry of the shape space $\mathcal{Y}^*_{/\,\mathrm{Trl}\,\times\,\mathrm{Rot}\,\times\,\mathrm{Scl}}$. Similar to previously discussed object data types, the shape geometry can be approached by referring to manifolds already known from other contexts. Due to normalization, pre-shapes $\mathbf{y}_i$ may be equipped with a spherical geometry. Including also rotation invariance, the shapes then correspond to points $[\mathbf{y}_i]_\mathrm{s} \in \mathbb{C}P^k$ in the projective space of complex "lines" in $\mathbb{C}^k$, i.e. the abstract manifold of all $\{z\,\mathbf{y} : z \in \mathbb{C}\}$ with $\mathbf{y} \in \mathbb{C}^k \setminus \{\mathbf{0}\}$, which is a standard example in differential geometry textbooks [102, 109]. A convenient property of $\mathbb{C}P^k$ is that, effectively, it allows to compute required geometric maps (such as geodesics and parallel transports) via suitable analogues on the complex sphere $\mathbb{C}\mathbb{S}^k = \{\mathbf{y} \in \mathbb{C}^k : \|\mathbf{y}\| = 1\}$, which offer well-known closed form expressions (compare, e.g., the supplement of [87]). Accordingly, also the *intrinsic distance* on the shape manifold (also referred to as "geodesic distance" or simply as "Procrustes distance" for shapes) is given by

$$d_R^2([\mathbf{y}_1]_\mathrm{s}, [\mathbf{y}_2]_\mathrm{s}) = \cos^{-1}(|\mathbf{y}_1^\dagger \mathbf{y_2}|) \tag{1.5}$$

reflecting the arc length between the pre-shape $\mathbf{y}_1$ and the pre-shape $\tilde{\mathbf{y}}_2 = \frac{\mathbf{y}_2^\dagger \mathbf{y}_1}{|\mathbf{y}_2^\dagger \mathbf{y}_1|}\mathbf{y_2}$ rotation aligned to $\mathbf{y}_1$. As Fréchet mean with respect to the intrinsic distance, the corresponding shape mean $[\hat{\boldsymbol{\mu}}_R]_\mathrm{s}$ is referred to as *intrinsic mean* or "Riemannian center of mass" [95, 3]. In the shape context it is also simply called "Procrustes mean". However, no explicit form is available also in the two dimensional case.

Dryden et al. [48] and Huckemann [88] compare full Procrustes means, intrinsic means, and other related notions of shape mean with respect to their estimation properties. In some data problems, it might be also desirable to preserve the size and analyze the *form* or "size-and-shape" $[\boldsymbol{y}_i]_{\mathrm{f}} = \{\exp(\sqrt{-1}\omega)\mathbf{y}_i + z\,\mathbf{1}_k : \omega \in (-\pi, \pi], z \in \mathbb{C}\}$ of $\mathbf{y}_i$, $i = 1, \ldots, n$, only modulo translation and rotation instead [226, 227]; or also the reflected landmark configuration $(\mathbf{y}_i^{\dagger})^{\top}$ might be considered equivalent to $\mathbf{y}_i$, motivating an analysis of reflection-shapes [49] or -forms [11, 150].

Shape analysis of outline curves: In various data problems, the ith shape of interest is not represented by landmarks, but recorded at k_i sampling points $\mathbf{y}_i = (y_{i1}, \ldots, y_{ik_i})^{\top}$ describing the outline of a two-dimensional section of an object, with k_i often varying over $i = 1, \ldots, n$. Considering the outlines as images of parameterized curves $y_i : [0, 1] \to \mathbb{C}$ and $y_{i1} = y_i(t_{ir})$ at t_{ir}, $r = 1, \ldots, k_i$, the "time" t indicates point correspondences between points $y_1(s)$ and $y_2(t)$ with $s = t$ of two different curves. This opens up a direct connection to FDA. Suitable indices $\mathbf{t}_i = (t_{i1}, \ldots, t_{ik_i})^{\top}$ might be directly obtained if the sampling design allows it or can be reasonably "imputed", say, by constant-speed parameterization. Uncertainty in t refers to the registration problem in FDA and will be revisited below in elastic shape analysis.
In analogy to FDA, shape analysis of (parameterized) outline curves can be subdivided into approaches on a point-wise basis and approaches expanding curves in bases $y_i(t) \approx \hat{y}_i(t) = \check{\mathbf{y}}_i^{\top}\mathbf{f}(t)$ with $\mathbf{f}(t) = (f_1(t), \ldots, f_L(t))^{\top}$ as described in Section 1.2.1. Although in practice $\mathbf{f}$ is often composed of L basis functions for each dimension with $2L$ real coefficients, here we equivalently assume $f : [0, 1] \to \mathbb{R}$ with complex coefficients $\check{\mathbf{y}}_i \in \mathbb{C}^L$ for illustration: as translation, rotation and scaling present affine transformations, the shape of a curve $[\hat{y}_i]_{\mathrm{s}} = \{\lambda\, \exp(\sqrt{-1}\omega)y_i + z\,\mathbf{1}_k : \lambda > 0, \omega \in (-\pi, \pi], z \in \mathbb{C}\} = \{\check{\mathbf{y}}^{\top}\mathbf{f}(t) : \check{\mathbf{y}} \in [\check{\mathbf{y}}_i]\}$ corresponds to the shape $[\check{\mathbf{y}}_i]$ of its coefficients. Hereby, we either assume that the constant function is spanned by $\mathbf{f}$ or consider $[\check{\mathbf{y}}_i]$ only modulo rotation and scale but not modulo translation. For an orthonormal basis $\mathbf{f}$, $\check{\mathbf{y}}_1^{\dagger}\check{\mathbf{y}}_2 = \langle \hat{y}_1, \hat{y}_2 \rangle$ reflects the $\mathbb{L}^2$ inner product and shape analysis of the outlines effectively reduces to shape analysis of their coefficients as "landmarks". Typical basis choices in the morphometrics literature are Fourier transforms for closed curves [162] and discrete cosine transforms [47] for open curves, but other bases used in FDA may equally be employed. Alternatively, regularly observed sampling vectors $\mathbf{y}_i$, with $k_i = k$ for all $i = 1, \ldots, n$, may directly be treated like landmarks [21]. While such an approach might seemingly dispense with parameterization, it implicitly assumes that $t_{ir} = t_r$, $r = 1, \ldots, k$, are equal for all i. As this assumption is often problematic, semi-landmark analysis [2] allows sampling points to slide tangentially along the outline curves to increase their comparability. However, it remains difficult to formalize the missing point correspondence when treating sampling points along the outline as landmarks without considering underlying parameterized curves $y_i(t)$.

In general, parameterization plays an ambiguous role in the shape analysis of curves:

on the one hand side, it implements a notion of point correspondence across curves allowing to base statistical analysis on local structures of the shapes (irrespective of whether they are parameterized explicitly or implicitly by defining matching point pairs); on the other side, the parameterization is often arbitrary and we often consider a curve as an object independent of the particular parameterization, which demands for basing analysis on a re-parameterization invariant geometry. Elastic shape analysis in the SRV framework [178] resolves this conflict, offering a rigorous platform for statistical analysis of parameterized curves under translation, rotation, re-scaling and re-parameterization invariance and providing the basis for Chapter 7 and 8.

Elastic shape analysis After initial work in the direction [94], Srivastava et al. [177] propose using the SRV framework (introduced in Section 1.2.3) for statistical analysis of shapes of curves. Statistical shape analysis and elastic analysis of curves are directly compatible since scaling and rotation $\lambda \exp(\sqrt{-1}\omega)\, y_i$ of a parameterized curve y_i by some $\lambda > 0$ and $\omega \in [-\pi, \pi)$ correspond to scaling and rotation $\sqrt{\lambda} \exp(\sqrt{-1}\omega)\, q_i$ of its SRV-transform $q_i = \dot{y}/\sqrt{|\dot{y}|}$ and translation simply vanishes in q_i. Now, a pre-shape is given by a normed SRV-transform q_i with $\|q_i\| = 1$. This corresponds to normalizing the underlying parameterized curve $\mathrm{L}(y_i) = \int |y_i(t)|\, d\nu = \|q_i\|^2 = 1$ with respect to its length, a natural size measure for curves invariant under re-parameterization. Combining the ideas behind the intrinsic shape distance d_R in (1.5) and the elastic distance d_e in (1.4) yields the elastic shape distance

$$d_{eR}([y_1],[y_2]) = \inf_{\gamma\in\Gamma,\omega\in\mathbb{R}} \cos^{-1}(\langle\, q_1, \exp(\sqrt{-1}\omega)\, q_2 \circ\gamma \sqrt{\dot{\gamma}}\,\rangle) \overset{\text{plane}}{=} \inf_{\gamma\in\Gamma} \cos^{-1}(|\langle\, q_1, q_2\circ\gamma \sqrt{\dot{\gamma}}\,\rangle|)$$

between two curve shapes $[y_i] = \{\lambda \exp(\sqrt{-1}\omega)\, y_i \circ \gamma + z : \lambda > 0, \omega \in [-\pi, \pi), z \in \mathbb{C}, \gamma \in \Gamma\}$ modulo all involved invariances, $i = 1, 2$, represented by SRV-transforms q_i. Based on the distance, Fréchet mean computation and other statistical analysis can be performed, as outlined by Srivastava and Klassen [178] in their introductory book on elastic shape analysis (including references to various applications and extensions). Assuming curves to be closed, i.e. $y_i(0) = y_i(1)$, presents a non-linear constraint $\int q_i(t)|q_i(t)|\, d\nu(t) = 0$ on SRV level, which Srivastava et al. [177] and following authors typically approach by basing d_{eR} on the intrinsic metric of the submanifold of pre-shapes of closed curves instead of d_R, for which however no closed-form solution is available. [14, 184] consider elastic shape analysis for landmark constrained curves. Building on ideas of the SRV framework, different generalizations to more complex geometrical structures have been introduced: for analysis of surfaces – in particular in the square-root-normal (SRN) approach – we refer to Jermyn et al. [93] on this topic; SRV approaches have been discussed for curves taking values on $\mathbb{S}^2$ [222] or more generally in a homogeneous space [186]; other recent developments include elastic analysis of tree structures [50, 206] and brain arterial networks as elastic graphs [73] composed of curve shapes.

1.3. It's all in the GAMe: a universe of semi-parametric regression built on the linear model

This thesis addresses generalization of semi-parametric modeling to object data. While the previous section introduced the different types of object data that will be considered, we now focus on the other side of the coin and illustrate the flexibility provided by semi-parametric regression in scalar data. For discussion of existing regression approaches for object data, including generalizations of additive models to functional data [72, 134] and generalized linear model (GLM) type regression on manifolds [40, 224, 174], we refer to the literature overviews provided in the single contributing articles.

Since *semi-parametric* generalized linear models [67] and *generalized additive models* (GAM) [76] have been proposed, a multitude of different model extensions have followed modeling response observations $y_1, \ldots, y_n$ with the generic GAM model structure

$$g(\vartheta_i) = h(\mathbf{x}_j) = \sum_{j=1}^{J} h_j(\mathbf{x}_i) \qquad (i = 1, \ldots, n) \tag{1.6}$$

through a characteristic ϑ_i of the conditional distribution of the response variable Y given the ith vector of covariates $\mathbf{x}_i$ via a *link function* g or, vice versa, via a *response function* g^{-1}, and with an additive predictor h composed of covariate effects h_j. While especially semi-parametric approaches have also been summarized as "structured additive regression" (STAR) models [56] highlighting the variety of extensions subsumed in the term, model extensions as well as semi- and non-parametric approaches are also commonly referred to as GAM [212]. Here, the broad notion of GAM is adopted while focusing on semi-parametric modeling, which will play a prominent role in all contributions of this thesis. The semi-parametric approach reduces estimation of general non-linear covariate effects effectively to (penalized) estimation of multiple linear effects, which has allowed to embed them into a variety of different model extensions of (generalized) linear models, providing a "tool-box" for applied data analysis. Different extensions, from various non-linear effect types over mixed models to generalized distribution assumptions, are often modularly combinable, and implemented based on different estimation and inference paradigms. Parts of these general developments are summarized in different text books [56, 212, 180]. They are often linked to different GAM frameworks with software implementations based on penalized maximum likelihood [212, 160, 217], Bayesian approaches [196, 197], or approaches based on gradient boosting [191, 82] or neural network architectures [165]. Each of the frameworks has successively grown involving series of publications. Software implementations or front-ends for their implementations are provided in R [151] for all examples listed above and often use similar syntax, which facilitates switching between different frameworks for practical modeling purposes. Moreover, there exist also connections to conditional transformation models [83] or tree-based boosting [34]. To give an idea of how this "universe" of semi-parametric modeling approaches allows adapting to various exper-

imental conditions in a data problem, we briefly discuss some of the available tools by the example of an analysis of pollutant particle concentrations in urban areas in the following. The author of this thesis participated in the project aside of the thesis contributions.

Alas et al. [5] analyze effects of different urban environmental factors on black carbon concentration in the cities of Rome and Leipzig based on mobile measurement data. The data was recorded by repeatedly walking fixed routes carrying mobile measurement devices in a backpack over several weeks. They model the black carbon concentration Y in dependence on different covariates (such as street type, wind, daytime, traffic) while accounting for various challenges implied by the spatio-temporal experimental setup.

The response distribution: As typical for inherently positive variables, recorded concentrations $y_1, \ldots, y_n$ show a distinctly right-skewed distribution, which suggests a corresponding conditional distribution assumption for Y. Assuming Y conditionally normal distributed as in a linear model instead would be problematic, as can be observed in heavily skewed residuals in this case. Generalized linear models (GLMs) [137] offer distribution families (such as, e.g., gamma distributions) for modeling skewed positive response distributions. They assume a distribution family that, when fixing a nuisance parameter, presents an exponential family in the parameter of interest. The original semi-parametric models / GAMs referred to above, generalize GLMs to non-linear predictors. In our example, a log-normal distribution is chosen for Y, which corresponds to modeling $Z = \log(Y)$ as conditionally normal distributed. However, Y is subject to additive instrumental noise common for the employed black carbon measurement devises such that, in fact, only $\tilde{Y} = Y + \epsilon = \exp(Z) + \epsilon$ is observed, where ϵ might be reasonably assumed to present an independent and zero mean Gaussian error (as indicated by lab experiments). In the data, this becomes eminent by encountering a non-negligible share of negative particle concentration measurements despite their theoretically positive range. To avoid serious bias, ϵ is explicitly included into the model, assuming now a *log-normal-normal convolution* (logNNC) as a non-standard distribution for $\tilde{Y}$ underlying $y_1, \ldots, y_n$. A *generalized additive model for location, scale and shape* (GAMLSS) framework for more general response distributions beyond GLMs was proposed by Rigby and Stasinopoulos [160]. Subsequently, different such *distributional regression* approaches where offered by other GAM frameworks. After a pilot implementation based on gradient boosting, we implemented the logNNC approach in the Bayesian framework of Umlauf et al. [197] extending the range of available distributions. Figure 1.1 illustrates estimated conditional distributions of Y in front of empirical distributions at an example segment of the Leipzig route in winter and summer. The depicted posterior distributions reflect mixtures over other covariates in the model predictor.

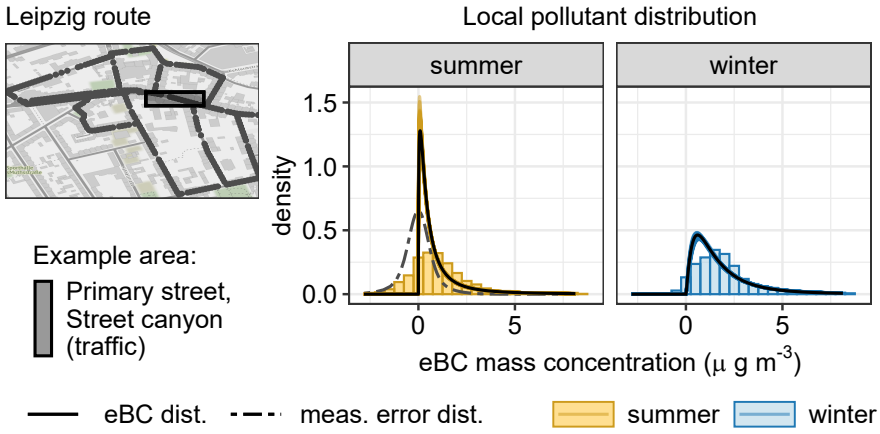

Figure 1.1.: Illustration of the logNNC model results on a distributional level: for an exemplary segment of the Leipzig route, seasonal differences in the empirical eBC distribution are depicted (shaded histograms). Corresponding densities resulting from the model are deconvoluted into eBC distribution (black solid lines) and measurement error distribution (gray dashed line). Colored fading shades around model eBC densities reflect estimation uncertainties (presenting 500 posterior samples). *Figure and caption are borrowed from Alas et al. [5] which is published under a Creative Commons Attribution 4.0 International License.*

Simultaneous modeling of multiple distribution parameters: In contrast to GLMs modeling only one response parameter ϑ_i (typically the conditional mean), multiple parameters are simultaneously modeled in GAMLSS / distribution regression including, for instance, also the response variance. For each of the parameters, a separate GAM model formula as in (1.6) is specified. In the given particle concentration example, the logNNC model assumed for $\tilde{Y}$ involves three distribution parameters: the mean μ_i and standard deviation σ_i of the original concentration Y on log-scale, i.e. $Z \sim N(\mu_i, \sigma_i^2)$ given covariates $\mathbf{x}_i$, and the standard deviation τ_i of the error process $\epsilon \sim N(0, \tau_i^2)$ which was found to sightly depend on environmental conditions in test experiments. As typical choices, $\mu_i = h^\mu(\mathbf{x}_i)$ is modeled with an identity link, i.e. $g(h) = h$, and standard deviations with log-links as $\log(\sigma_i) = h^\sigma(\mathbf{x}_i)$ and $\log(\tau_i) = h^\tau(\mathbf{x}_i)$. Here, h^μ, h^σ, h^τ denote individual additive predictors as in (1.6) for the different parameters that might, in general, depend on the same or different sets of covariates. While h^μ and h^σ are chosen to depend on the same set of covariates of interest, h^τ includes only an intercept and a temporally varying effect (substituting unknown micro-effects), since other covariates are assumed not to influence the instrument noise.

Semi-parametric modeling of non-linear covariate effects: Having the model targets specified, we may consider the composition of the predictor(s) $h(\mathbf{x}_i)$ in more detail. In a GLM, the predictor is the one of a linear model comprising linear effects of metric covariates, such as in our example the current wind speed, and categorical covariate effects, e.g., of street-type (primary, secondary, tertiary, residential, park), as well as potential interactions thereof. GAMs extend this to an additive predictor including also non-linear covariate effects in addition to specified linear model effects. In our example,

a non-linear effect is specified for the time of day when the measurement was conducted, with the effect on black carbon concentration periodically and "smoothly" varying over the day. In the semi-parametric approach, effect functions $h_j(\mathbf{x}) = \sum_{l=1}^{L_j} b_{j1}(\mathbf{x})\theta_{jl}$ are expanded in a function basis $b_{j1}, \ldots, b_{jL_j}$ with basis coefficients $\theta_{j1}, \ldots, \theta_{jL_j} \in \mathbb{R}$, such that model equation (1.6), comprising both linear and non-linear effects, can be rewritten in the parametric form of a GLM as

$$g(\vartheta_i) = \sum_{j=1}^{J} \sum_{l=1}^{L_j} b_{jl}(\mathbf{x}_i)\theta_{jl} = \mathbf{b}_i^\top \boldsymbol{\theta}$$

where the $b_{jl}(\mathbf{x}_i)$ are fixed, known quantities. Evaluations $\mathbf{b}_i = (b_{11}(\mathbf{x}_i), \ldots, b_{JL_J}(\mathbf{x}_i))^\top$ over all basis functions $l = 1, \ldots, L_j$ for all covariate effects $j = 1, \ldots, J$ become pseudo-covariates for estimation of a linear coefficient vector $\boldsymbol{\theta} = (\theta_{11}, \ldots, \theta_{JL_J})^\top$ with all basis coefficients in the additive model predictor. While also linear effects are covered as simple special cases by this basis approach, different polynomial spline bases, such as B-splines, are often used for implementing non-linear effects. Another popular choice are thin-plate splines [212]. With potentially a considerable number L_j of coefficients for the jth covariate effect, penalized estimation strategies – also used for high-dimensional data in general – are typically employed in semi-parametric modeling. Their penalties are mostly chosen in such a way that estimation of covariate effects h_j is regularized towards smoothness.

Mixed models and spatio-temporal modeling: In many data scenarios, the dependency structure behind response observations $y_1, \ldots, y_n$ cannot be fully explained by available covariates. Instead, associations can be expected within grouped measurements, between subsequent measurements in time, or between measurements recorded close to each other in space. While there are also approaches directly modeling correlations between response observations in a GLM [225], mixed models [185] approach this problem by incorporating latent random effects into the model predictor $h(\mathbf{x}_i)$. In our example, measurements are recorded within single "runs" where a person walks the given route equipped with instruments. These runs are conducted once or twice a day over several weeks, in Leipzig also with some runs during summer and some during winter time. Adding a run identifier $\texttt{runID}_i$ to the set of covariates, for $i = 1, \ldots, n$, a run-specific random intercept $h_j(\mathbf{x}_i) = \xi_{\texttt{runID}_i}$ can explain common deviations between runs. For each run, the $\xi_{\texttt{runID}_i}$ are assumed independent zero-mean Gaussian random variables. Due to this randomness assumption which regularizes $\xi_{\texttt{runID}_i}$ towards zero, the run-specific random effect is not designed to explain all of the variation between runs that could be explained by the effect, but only parts that cannot be attributed to other effects. By this means, a random effect should account for unexplained homogeneity within groups without covering effects of other covariates.
However, the ith particle concentration measurement is not only part of a run but recorded at a particular time t_i and at a particular location s_i on the map. This sug-

gests a *geoadditive* extension of the model with the predictor

$$h(\mathbf{x}_i) = \cdots + \xi_{time}(t_i) + \xi_{space}(s_i)$$

containing special temporal and spatial (random) components. To implement the assumption that measurements close in time are dependent, a latent zero-mean Gaussian process $\xi_{time}(t_i)$ is added to the predictor. Instead of assuming $\xi_{time}(t_\imath)$ and $\xi_{time}(t_i)$ independent for different $t_\imath \neq t_i$ as for the random intercepts described above, their correlation is now typically specified via a correlation function $\varrho(|t_\imath - t_i|) = \text{Corr}(\xi_{time}(t_\imath), \xi_{time}(t_i))$. Proceeding similarly with locations on the map would yield a generalized Kriging analogue approach. However, in our example mobile measurement routes go along roads where buildings on the sides likely influence spatial distribution of particles and spatial correlation is expected to be stronger along the roads than across building blocks. To account for this more complex spatial structure, routes are discretized into small segments in this case to utilize a Gauss Markov random field $\xi_{space}(s_i)$ with s_i an identifier for the route segment. This allows borrowing the neighborhood structure of the discrete segments s_i.

The mobile measurement example indicates the variety of tools readily available in semi-parametric GAMs to flexibly model challenging data scenarios, when integrating novel extensions such as in our case the logNNC model. Although model assumptions do not always fit perfectly or cannot always be conclusively validated, this provides the basis for multi-faceted and understandable analysis of real-life data problems. In our example, for instance, effect estimates match expert knowledge and model diagnostics are improved by far comparing them to simpler models, more evidence on the dependence of black carbon concentration on local, temporary sources can be provided (in contrast to general ultra-fine particle measurements), and even estimates of the instrumental noise variance are in line with results from lab experiments.

1.4. Discussion and outlook

Facing complex object data on the one hand, we cannot expect experimental designs to be less demanding than for scalar data on the other hand. Or, to put it more concretely, in regression for object data responses, we cannot expect required covariate and dependency structures to be simpler than in scalar modeling scenarios. The purpose of this work is to carry on recent developments in functional regression [72] to extend the range of the semi-parametric GAM "tool box" to models for further, geometric object data. Building on well-established regression frameworks allows developments to modularly draw on available tools and facilitates practical use in real-world data problems. Consequently, even though some approaches to modeling nonlinear effects of multiple covariates beyond point-wise functional mean regression have recently been proposed with additive models for Hilbert space responses [91, 92, 90] and Lie groups

[116], and Fréchet regression [146] for very general response objects, I am not aware of any other regression framework to date that offers a comparable infrastructure for flexible and detailed modeling of corresponding object data problems as we are able to provide ready to use in the contributions of this thesis.

Beyond prediction of response objects, which is often of secondary interest in research problems, model interpretation and visualization are crucial for an insightful analysis, but become challenging when modeling already multidimensional object data with non-linear covariate effects. An analytic and systematic investigation requires decomposing potentially complex predictors into simpler, understandable parts. Throughout the different contributions, thorough analysis, thus, involves careful decomposition of model predictors into meaningful orthogonal parts (Chapter 3 and 5), decomposition of multivariate effects into marginal parts and interactions (Chapter 2 and 5), decomposition of the functional covariance structure into different independent components (Chapter 4), decomposition of smooth effects into linear and non-linear parts (Chapters 2, 5, 8), and decomposition of non-linear effects into main effect directions by tensor product factorization proposed in Chapter 5 and used also in Chapter 8, which allow for graphical visualization and produce tangible model results.

While contributions in Chapter 6 and Chapter 7 consider unconditional mean estimation providing fundamentals for analysis of (shapes of) irregularly/sparsely observed curves and the multivariate functional mixed model in Chapter 4 is fitted based on penalized maximum likelihood, regression models in several of the contributions are fitted based on the model-based boosting framework of Hothorn et al. [82] (Chapters 2, 3, 5 and 8) which comes with some important advantages but also with limitations: the component-wise fitting strategy involves double regularization, including quadratic penalties in each fitting iteration and a global regularization by stopping the algorithm early based on cross-validation, which not only offers automated model selection but also allows to fit models with high dimensional predictors and responses despite typically high autocorrelation in functional data. Moreover, it is very flexible also in the sense that it reduces fitting arbitrary loss functions to re-fitting of pseudo-residuals. Therefore, boosting proves suitable for the challenging model scenarios discussed in this thesis, where we build on the GAMLSS extension of Thomas et al. [191] in Chapter 2 and generalize L_2-Boosting [26] to Bayes Hilbert spaces in Chapter 3, to Riemannian manifolds in 5 and to elastic Riemannian L_2-Boosting in Chapter 8. However, despite model-based boosting being well-established [128, 129] and especially tree-based gradient boosting being known for its good performance [34], limited theoretical results are available concerning asymptotic properties [e.g. 119] such that, employing generalized boosting approaches, we rely on simulation studies for model evaluation. Also inference, in particular after model selection, is only available for special cases, yet [163]. Bootstrapping can be considered in this context, but is problematic due to shrinkage bias induced by the regularization. Still, automated model selection and boosting-based variable importance measures give alternative indication of meaningful covariate effects.

More generally, besides e.g. results in manifold regression by Cornea et al. [40], inference in complex object data structures, such as shapes of curves in the SRV-framework, presents a challenging task leaving many open questions for future research. Moreover, the contributions of this thesis focus on semi-parametric modeling of functional data, densities, curves and shapes – mostly in dependence on scalar covariates. Although infrastructure is available for specify also functional covariates in the model (see Chapter 2), more work will be required for including covariate effects of such object data as covariates into the framework. Similarly, more work will be needed for incorporating tools for modeling longitudinal/hierarchical experimental designs with geometric functional responses, which has been considered by several authors [169, 101, 136, 223, 43, 113, 37] in different related contexts. Finally, further extending flexible regression models to other, potentially more complexly composed object data structures (some of which referred to in Section 1.2) will present a rich and exciting field of future research, facing the challenges of real-world data problems – both in terms of the objects of analysis and the experimental demands and questions.

In a world composed of complex structures, we are experiencing an era of rapid developments in the analysis of object data and, at the same time, rich discoveries of new object data structures for analysis. We are, thus, able to rely on a multitude of preceding developments, but not less importantly, a multitude of data problems lies ahead in what Wang et al. [208] call "next generation" functional data analysis, demanding statisticians to bridge between advanced geometric object data structures, statistical models and communication with applied researchers.

Bibliography

[1] Adams, D., Rohlf, F., and Slice, D. (2004). Geometric morphometrics: ten years of progress following the 'revolution'. *Italian Journal of Zoology*, 71. 12

[2] 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(1):7–14. 12, 15

[3] Afsari, B. (2011). Riemannian L^p center of mass: existence, uniqueness, and convexity. *Proceedings of the American Mathematical Society*, 139(2):655–673. 14

[4] Aitchison, J. (1986). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160. 8

[5] Alas, H. D., Stöcker, A., Umlauf, N., Senaweera, O., Pfeifer, S., Greven, S., and Wiedensohler, A. (2021). Pedestrian exposure to black carbon and PM2.5 emissions in urban hot spots: New findings using mobile measurement techniques and flexible Bayesian regression models. *Journal of Exposure Science & Environmental Epidemiology*. 18, 19

[6] Amari, S. (1990). *Differential-geometrical methods in statistics*. Springer Science & Business Media, 2 edition. 7

[7] Amari, S. (2016). *Information geometry and its applications*, volume 194. Springer. 7

[8] Aneiros, G., Cao, R., Fraiman, R., Genest, C., and Vieu, P. (2019). Recent advances in

functional data analysis and high-dimensional statistics. *Journal of Multivariate Analysis*, 170:3–9. 3

[9] Arsigny, V., Fillard, P., Pennec, X., and Ayache, N. (2007). Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 29(1):328–347. 2

[10] Bachoc, F., Suvorikova, A., Ginsbourger, D., Loubes, J.-M., and Spokoiny, V. (2020). Gaussian processes with multidimensional distribution inputs via optimal transport and Hilbertian embedding. *Electronic Journal of Statistics*, 14(2):2742–2772. 7

[11] Bandulasiri, A., Bhattacharya, R. N., and Patrangenaru, V. (2009). Nonparametric inference for extrinsic means on size-and-(reflection)-shape manifolds with applications in medical imaging. *Journal of Multivariate Analysis*, 100(9):1867–1882. 15

[12] Barden, D., Le, H., and Owen, M. (2018). Limiting behaviour of fréchet means in the space of phylogenetic trees. *Annals of the Institute of Statistical Mathematics*, 70(1):99–129. 2

[13] Bauer, A., Scheipl, F., Küchenhoff, H., and Gabriel, A.-A. (2021). Registration for incomplete non-Gaussian functional data. *arXiv preprint arXiv:2108.05634*. 11

[14] Bauer, M., Eslitzbichler, M., and Grasmair, M. (2015). Landmark-guided elastic shape analysis of human character motions. *arXiv preprint arXiv:1502.07666*. 16

[15] Benatia, D., Carrasco, M., and Florens, J.-P. (2017). Functional linear regression with functional response. *Journal of Econometrics*, 201(2):269–291. 7

[16] Bernal, J., Dogan, G., and Hagwood, C. R. (2016). Fast dynamic programming for elastic registration of curves. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 12

[17] Berninger, C., Stöcker, A., and Rügamer, D. (2022). A Bayesian time-varying autoregressive model for improved short-term and long-term prediction. *Journal of Forecasting*, 41(1):181–200. 3

[18] Bhattacharrya, A. (1943). On discrimination and divergence. *Proc. 29th Indian Sci. Cong*, part III(3). 7

[19] Billera, L. J., Holmes, S. P., and Vogtmann, K. (2001). Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, 27(4):733–767. 2

[20] Bookstein, F. L. (1978). The measurement of biological shape and shape change. 13

[21] Bookstein, F. L. (1997). Landmark methods for forms without landmarks: morphometrics of group differences in outline shape. *Medical Image Analysis*, 1(3):225–243. 15

[22] Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P. (2017). Geometric deep learning: going beyond Euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42. 2

[23] Bruveris, M. (2016). Optimal reparametrizations in the square root velocity framework. *SIAM Journal on Mathematical Analysis*, 48(6):4335–4354. 12

[24] Bryner, D. and Srivastava, A. (2021). Shape analysis of functional data with elastic partial matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 12

[25] Buccianti, A., Mateu-Figueras, G., and Pawlowsky-Glahn, V. (2006). Compositional data analysis in the geosciences: from theory to practice. Geological Society of London. 8

[26] Bühlmann, P. and Yu, B. (2003). Boosting with the L2 loss: regression and classification. *Journal of the American Statistical Association*, 98(462):324–339. 22

[27] Cabrera, B. L. and Schulz, F. (2017). Forecasting generalized quantiles of electricity demand: A functional data approach. *Journal of the American Statistical Association*, 112(517):127–136. 7

[28] Calissano, A., Feragen, A., and Vantini, S. (2022). Graph-valued regression: Prediction of unlabelled networks in a non-euclidean graph space. *Journal of Multivariate Analysis*, 190:104950. 2

[29] Cao, W., Yan, Z., He, Z., and He, Z. (2020). A comprehensive survey on geometric deep learning. *IEEE Access*, 8:35929–35949. 2

[30] Carroll, C. and Müller, H.-G. (2021). Latent transport models for multivariate functional data. *arXiv preprint arXiv:2107.05730*. 11

[31] Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC. 11

[32] Cazelles, E., Seguy, V., Bigot, J., Cuturi, M., and Papadakis, N. (2018). Geodesic PCA versus log-PCA of histograms in the Wasserstein space. *SIAM Journal on Scientific Computing*, 40(2):B429–B456. 2, 3, 7

[33] Cederbaum, J., Scheipl, F., and Greven, S. (2018). Fast symmetric additive covariance smoothing. *Computational Statistics & Data Analysis*, 120:25–41. 5

[34] Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. 17, 22

[35] Chen, Y., Lin, Z., and Müller, H.-G. (2021). Wasserstein regression. *Journal of the American Statistical Association*, 0(0):1–14. 7

[36] Cheng, W., Dryden, I. L., and Huang, X. (2016). Bayesian registration of functions and curves. *Bayesian Analysis*, 11(2):447–475. 12

[37] Chevallier, J., Debavelaere, V., and Allassonnière, S. (2021). A coherent framework for learning spatiotemporal piecewise-geodesic trajectories from longitudinal manifold-valued data. *SIAM Journal on Imaging Sciences*, 14(1):349–388. 23

[38] Claeskens, G., Krivobokova, T., and Opsomer, J. D. (2009). Asymptotic properties of penalized spline estimators. *Biometrika*, 96(3):529–544. 5

[39] Combettes, P. L. and Müller, C. L. (2021). Regression models for compositional data: General log-contrast formulations, proximal optimization, and microbiome data applications. *Statistics in Biosciences*, 13(2):217–242. 8

[40] 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, 23

[41] Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, 147:1–23. 3

[42] Dai, X. (2022). Statistical inference on the Hilbert sphere with application to random densities. *Electronic Journal of Statistics*, 16(1):700–736. 8

[43] Debavelaere, V., Durrleman, S., and Allassonnière, S. (2020). Learning the clustering

of longitudinal shape data sets into a mixture of independent or branching trajectories. *International Journal of Computer Vision*, 128(12):2794–2809. 23

[44] Delicado, P. (2011). Dimensionality reduction when data are density functions. *Computational Statistics & Data Analysis*, 55(1):401–420. 7

[45] Descary, M.-H. and Panaretos, V. M. (2019). Recovering covariance from functional fragments. *Biometrika*, 106(1):145–160. 7

[46] Diggle, P. J., Heagerty, P., Liang, K.-Y., and Zeger, S. (2002). *Analysis of longitudinal data*. Oxford university press. 4

[47] Dommergues, C. H., Dommergues, J.-L., and Verrecchia, E. P. (2007). The discrete cosine transform, a Fourier-related method for morphometric analysis of open contours. *Mathematical Geology*, 39(8):749–763. 15

[48] 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. 15

[49] Dryden, I. L. and Mardia, K. V. (2016). *Statistical Shape Analysis: With Applications in R*. John Wiley & Sons. 2, 12, 13, 14, 15

[50] Duncan, A., Klassen, E., and Srivastava, A. (2018). Statistical shape analysis of simplified neuronal trees. *The Annals of Applied Statistics*, 12(3):1385–1421. 16

[51] Ebert, J., Spokoiny, V., and Suvorikova, A. (2019). Elements of statistical inference in 2-Wasserstein space. In *Topics in Applied Analysis and Optimisation*, pages 139–158. Springer. 7

[52] Eckardt, M., Mateu, J., and Greven, S. (2022). Generalised functional additive mixed models with compositional covariates for areal Covid-19 incidence curves. *arXiv preprint arXiv:2201.08362*. 10

[53] 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(4):1175–1182. 9

[54] Eilers, P. H., Marx, B. D., and Durbán, M. (2015). Twenty years of P-splines. *Sort*, 39(2):149–186. 5

[55] Eubank, R. L. (1999). *Nonparametric regression and spline smoothing*. CRC press. 5

[56] Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). Regression models. In *Regression*, pages 21–72. Springer. 17

[57] Filzmoser, P., Hron, K., and Templ, M. (2018). *Applied compositional data analysis*. Springer. 8, 9, 10

[58] Fishbaugh, J., Prastawa, M., Gerig, G., and Durrleman, S. (2013). *Geodesic Shape Regression in the Framework of Currents*, pages 718–729. Springer Berlin Heidelberg, Berlin, Heidelberg. 2

[59] Fletcher, P. T. and Joshi, S. (2007). Riemannian geometry for the statistical analysis of diffusion tensor data. *Signal Processing*, 87(2):250–262. 2

[60] Fletcher, P. T., Lu, C., and Joshi, S. (2003). Statistics of shape via principal component analysis on Lie group. In *In Proceedings of CVPR*. Citeseer. 3

[61] Fréchet, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distan-

cié. In *Annales de l'institut Henri Poincaré*, volume 10, pages 215–310. 2

[62] Friji, R., Drira, H., Chaieb, F., Kurtek, S., and Kchok, H. (2020). KShapeNet: Riemannian network on Kendall shape space for skeleton based action recognition. *arXiv preprint arXiv:2011.12004*. 13

[63] Ghodrati, L. and Panaretos, V. M. (2022). Distribution-on-distribution regression via optimal transport maps. *Biometrika*. asac005. 7

[64] Goh, A. and Vidal, R. (2008). Unsupervised Riemannian clustering of probability density functions. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 377–392. Springer. 8

[65] Goodall, C. (1991). Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society: Series B*, 53(2):285–321. 13

[66] Goodman, N. R. (1963). Statistical analysis based on a certain multivariate complex Gaussian distribution (an introduction). *The Annals of Mathematical Statistics*, 34(1):152–177. 14

[67] Green, P. J. and Yandell, B. S. (1985). Semi-parametric generalized linear models. In *Generalized linear models*, pages 44–55. Springer. 17

[68] Greenacre, M. and Lewi, P. (2009). Distributional equivalence and subcompositional coherence in the analysis of compositional data, contingency tables and ratio-scale measurements. *Journal of Classification*, 26(1):29–54. 9

[69] Grenander, U. (1950). Stochastic processes and statistical inference. *Arkiv för matematik*, 1(3):195–277. 3

[70] Grenander, U. (1981). *Abstract Inference*. John Wiley & Sons, Inc. 2, 3

[71] Grenander, U. and Miller, M. I. (2006). *Pattern theory: from representation to inference*. OUP Oxford. 2

[72] 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. 17, 21

[73] Guo, X., Bal, A. B., Needham, T., and Srivastava, A. (2020). Statistical shape analysis of brain arterial networks (ban). *arXiv preprint arXiv:2007.04793*. 16

[74] Happ, C., Scheipl, F., Gabriel, A.-A., and Greven, S. (2019). A general framework for multivariate functional principal component analysis of amplitude and phase variation. *Stat*, 8(1):e220. 11

[75] Hartman, E., Sukurdeep, Y., Charon, N., Klassen, E., and Bauer, M. (2021). Supervised deep learning of elastic SRV distances on the shape space of curves. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4425–4433. 12

[76] Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, pages 297–310. 17

[77] Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer. 3

[78] Holmes, S. (2003). Statistics for phylogenetic trees. *Theoretical population biology*, 63(1):17–32. 2

[79] Hong, Y., Yang, X., Kwitt, R., Styner, M., and Niethammer, M. (2017). Regression

uncertainty on the Grassmannian. In *Artificial Intelligence and Statistics*, pages 785–793. PMLR. 2

[80] Horváth, L. and Kokoszka, P. (2012). *Inference for functional data with applications*, volume 200. Springer Science & Business Media. 3

[81] Hosni, N. and Ben Amor, B. (2020). A geometric ConvNet on 3D shape manifold for gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 852–853. 13

[82] 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. 17, 22

[83] Hothorn, T., Kneib, T., and Bühlmann, P. (2014). Conditional transformation models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):3–27. 17

[84] Hron, K., Machalová, J., and Menafoglio, A. (2020). Bivariate densities in Bayes spaces: orthogonal decomposition and spline representation. *arXiv preprint arXiv:2012.12948*. 10

[85] Hron, K., Menafoglio, A., Templ, M., Hruzová, K., and Filzmoser, P. (2016). Simplicial principal component analysis for density functions in Bayes spaces. *Computational Statistics & Data Analysis*, 94:330–350. 10

[86] Hsing, T. and Eubank, R. (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*. John Wiley & Sons. 2, 3, 4, 6

[87] 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(4):593–603. 14

[88] 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(6):1227–1259. 15

[89] Iurato, G. and Igamberdiev, A. U. (2020). D'Arcy W. Thompson's on growth and form: A landmark for the mathematical foundations of epigenetics. *Biosystems*, 198:104279. 12

[90] Jeon, J. M., Lee, Y. K., Mammen, E., and Park, B. U. (2022). Locally polynomial Hilbertian additive regression. *Bernoulli*, 28(3):2034–2066. 21

[91] Jeon, J. M. and Park, B. U. (2020). Additive regression with Hilbertian responses. *The Annals of Statistics*, 48(5):2671–2697. 21

[92] Jeon, J. M., Park, B. U., and Van Keilegom, I. (2021). Additive regression for non-Euclidean responses and predictors. *The Annals of Statistics*, 49(5):2611–2641. 21

[93] Jermyn, I. H., Kurtek, S., Laga, H., and Srivastava, A. (2017). Elastic shape analysis of three-dimensional objects. *Synthesis Lectures on Computer Vision*, 12(1):1–185. 16

[94] Joshi, S. H., Klassen, E., Srivastava, A., and Jermyn, I. (2007). An efficient representation for computing geodesics between n-dimensional elastic shapes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 16

[95] Karcher, H. (1977). Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics*, 30(5):509–541. 14

[96] Karhunen, K. (1946). Zur Spektraltheorie stochasticher Prozesse. In *Annales Academiae Scientiarum Fennicae Series A*, volume 1, page 34. 3

[97] Kauermann, G., Krivobokova, T., and Fahrmeir, L. (2009). Some asymptotic results on

generalized penalized spline smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):487–503. 5

[98] Kendall, D. G. (1977). The diffusion of shape. *Advances in applied probability*, 9(3):428–430. 2, 13

[99] Kent, J. T. (1994). The complex Bingham distribution and shape analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(2):285–299. 13, 14

[100] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. (2020). Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*. 3

[101] 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)*, pages 5777–5786. 23

[102] Klingenberg, W. (1995). *Riemannian geometry*. de Gruyter. 14

[103] Kneip, A. and Liebl, D. (2020). On the optimal reconstruction of partially observed functional data. *The Annals of Statistics*, 48(3):1692 – 1717. 7

[104] Kokoszka, P. and Reimherr, M. (2017). *Introduction to functional data analysis*. Chapman and Hall/CRC. 3

[105] Krim, H. and Yezzi, A. J. (2006). *Statistics and analysis of shapes*. Springer. 2

[106] Ktena, S. I., Parisot, S., Ferrante, E., Rajchl, M., Lee, M., Glocker, B., and Rueckert, D. (2018). Metric learning with spectral graph convolutions on brain connectivity networks. *NeuroImage*, 169:431–442. 3

[107] Kurtek, S. (2017). A geometric approach to pairwise bayesian alignment of functional data using importance sampling. *Electronic Journal of Statistics*, 11(1):502–531. 12

[108] Lahiri, S., Robinson, D., and Klassen, E. (2015). Precise matching of PL curves in $\mathbb{R}^n$ in the square root velocity framework. *Geometry, Imaging and Computing*, 2(3):133–186. 12

[109] Lee, J. M. (2018). *Introduction to Riemannian manifolds*, volume 176. Springer. 14

[110] Li, C., Xiao, L., and Luo, S. (2020). Fast covariance estimation for multivariate sparse functional data. *Stat*, 9(1):e245. 5

[111] Li, Y. and Ruppert, D. (2008). On the asymptotics of penalized splines. *Biometrika*, 95(2):415–436. 5

[112] Liew, B. X., Rügamer, D., Stöcker, A., and De Nunzio, A. M. (2020). Classifying neck pain status using scalar and functional biomechanical variables–development of a method using functional data boosting. *Gait & posture*, 76:146–150. 3

[113] Lila, E. and Aston, J. A. (2020). Functional random effects modeling of brain shape and connectivity. *arXiv preprint arXiv:2009.06059*. 23

[114] Lila, E., Aston, J. A., and Sangalli, L. M. (2016). Smooth principal component analysis over two-dimensional manifolds with an application to neuroimaging. *The Annals of Applied Statistics*, 10(4):1854–1879. 2

[115] Lin, Z. (2019). Riemannian geometry of symmetric positive definite matrices via cholesky decomposition. *SIAM Journal on Matrix Analysis and Applications*, 40(4):1353–

1370. 2

[116] 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*. 22

[117] Loève, M. (1946). Fonctions aléatoires à décomposition orthogonale exponentielle. *La Revenue Scientifique*, 84:159–162. 3

[118] 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. 12

[119] Lutz, R. W. and Bühlmann, P. (2006). Boosting for high-multivariate responses in high-dimensional linear regression. *Statistica Sinica*, pages 471–494. 22

[120] Ma, Y. and Fu, Y. (2012). *Manifold learning theory and applications*, volume 434. CRC press Boca Raton. 3

[121] Mardia, K. V. and Jupp, P. E. (2000). *Directional Statistics*. John Willey and Sons, LTD. 2

[122] Marron, J., Ramsay, J. O., Sangalli, L. M., and Srivastava (Eds.), A. (2014). Statistics of time warpings and phase variations [special section]. *Electronic Journal of Statistics*, 8(2):1697–1939. 11

[123] Marron, J. S. and Alonso, A. M. (2014). Overview of object oriented data analysis. *Biometrical Journal*, 56(5):732–753. 2

[124] Martín-Fernández, J.-A., Hron, K., Templ, M., Filzmoser, P., and Palarea-Albaladejo, J. (2015). Bayesian-multiplicative treatment of count zeros in compositional data sets. *Statistical Modelling*, 15(2):134–158. 9

[125] Martın-Fernandez, J. A., Palarea-Albaladejo, J., and Olea, R. A. (2011). Dealing with zeros. *Compositional data analysis*, pages 43–58. 9

[126] Matuk, J., Bharath, K., Chkrebtii, 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*, pages 1–17. 12

[127] Matuk, J., Mohammed, S., Kurtek, S., and Bharath, K. (2020). Biomedical applications of geometric functional data analysis. In *Handbook of Variational Methods for Nonlinear Geometric Data*, pages 675–701. Springer. 2

[128] Mayr, A., Binder, H., Gefeller, O., and Schmid, M. (2014a). The evolution of boosting: From machine learning to statistical modelling. *Methods of information in medicine*, 53:419–27. 22

[129] Mayr, A., Binder, H., Gefeller, O., and Schmid, M. (2014b). Extending statistical boosting: An overview of recent methodological developments. *Methods Inf Med*, 53:428–35. 22

[130] Menafoglio, A., Guadagnini, A., and Secchi, P. (2014). A kriging approach based on Aitchison geometry for the characterization of particle-size curves in heterogeneous aquifers. *Stochastic Environmental Research and Risk Assessment*, 28(7):1835–1851. 10

[131] Menafoglio, A., Secchi, P., and Guadagnini, A. (2016). A class-kriging predictor for functional compositions with application to particle-size curves in heterogeneous aquifers. *Mathematical Geosciences*, 48(4):463–485. 10

[132] Miller, M. I., Trouvé, A., and Younes, L. (2006). Geodesic shooting for computational anatomy. *Journal of mathematical imaging and vision*, 24(2):209–228. 2

[133] Morris, J. S. (2015). Functional Regression. *Annual Review of Statistics and its Applications*, 2:321–359. 1, 2

[134] Morris, J. S. and Carroll, R. J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society, Series B*, 68(2):179–199. 4, 17

[135] Mosimann, J. E. (1970). Size allometry: size and shape variables with characterizations of the lognormal and generalized gamma distributions. *Journal of the American Statistical Association*, 65(330):930–945. 9, 13

[136] Nava-Yazdani, E., Hege, H.-C., and Tycowicz, C. v. (2019). A geodesic mixed effects model in Kendall's shape space. In *Multimodal Brain Image Analysis and Mathematical Foundations of Computational Anatomy*, pages 209–218. Springer. 23

[137] Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384. 18

[138] Newton, N. J. (2012). An infinite-dimensional statistical manifold modelled on Hilbert space. *Journal of Functional Analysis*, 263(6):1661–1681. 7

[139] Panaretos, V. M. and Zemel, Y. (2016). Amplitude and phase variation of point processes. *The Annals of Statistics*, 44(2):771 – 812. 11

[140] Panaretos, V. M. and Zemel, Y. (2020). *An invitation to statistics in Wasserstein space*. Springer Nature. 7

[141] Park, J. Y. and Qian, J. (2012). Functional regression of continuous state distributions. *Journal of Econometrics*, 167(2):397–412. 7

[142] Pawlowsky-Glahn, V. and Buccianti, A. (2011). *Compositional data analysis*. Wiley Online Library. 8

[143] Pawlowsky-Glahn, V., Egozcue, J. J., and Tolosana-Delgado, R. (2015). *Modeling and analysis of compositional data*. John Wiley & Sons. 2, 8

[144] Pegoraro, M. and Secchi, P. (2021). Functional data representation with merge trees. 12

[145] Pennec, X. (2006). Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127–154. 2

[146] Petersen, A. and Müller, H.-G. (2019). Fréchet regression for random objects with euclidean predictors. *The Annals of Statistics*, 47(2):691–719. 22

[147] Petersen, A. and Müller, H.-G. (2016). Functional data analysis for density functions by transformation to a Hilbert space. *The Annals of Statistics*, 44(1):183 – 218. 7

[148] Petersen, A., Zhang, C., and Kokoszka, P. (2022). Modeling probability density functions as data objects. *Econometrics and Statistics*, 21:159–178. 2, 7

[149] Pöllath, N., Alibert, P., Schafberg, R., and Peters, J. (2018). Striking new paths–distinguishing ancient ovis orientalis from its modern domestic descendant (Karakul breed) applying geometric and traditional morphometric approaches to the astragalus. In *Archaeozoology of the Near East XII. Proceedings of the 12th International Symposium of the ICAZ Archaeozoology of Southwest Asia and Adjacent Areas Working Group, Groningen Institute of Archaeology, June 14-15 2015, University of Groningen, the Netherlan*, pages 207–225.

9, 13

[150] Preston, S. and Wood, A. T. (2011). Bootstrap inference for mean reflection shape and size-and-shape with three-dimensional landmark data. *Biometrika*, 98(1):49–63. 15

[151] R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 17

[152] Ramsay, J. O. (1982). When the data are functions. *Psychometrika*, 47(4):379–396. 3

[153] 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. 3

[154] Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer New York. 2, 3, 4, 5, 6, 11

[155] Rao, C. R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Reson. J. Sci. Educ*, 20:78–90. 7

[156] Rao, C. R. (1958). Some statistical methods for comparison of growth curves. *Biometrics*, 14(1):1–17. 3

[157] Reimann, C., Filzmoser, P., Fabian, K., Hron, K., Birke, M., Demetriades, A., Dinelli, E., Ladenberger, A., and Team, T. G. P. (2012). The concept of compositional data analysis in practice—total major element concentrations in agricultural and grazing land soils of europe. *Science of the total environment*, 426:196–210. 8

[158] Reiss, P. T. and Xu, M. (2020). Tensor product splines and functional principal components. *Journal of Statistical Planning and Inference*, 208:1–12. 5

[159] Rholf, F. and Marcus, L. F. (1993). A revolution in morphometrics. *Trends in Ecology & Evolution*, 8(4). 2, 12

[160] Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554. 17, 18

[161] Rodrigues, L. A., Daunis-i Estadella, J., Mateu-Figueras, G., and Thio-Henestrosa, S. (2011). Flying in compositional morphospaces: evolution of limb proportions in flying vertebrates. *Compositional Data Analysis: Theory and applications. John Wiley & Sons, Ltd.* 8, 12

[162] Rohlf, F. and Archie, J. (1984). A comparison of Fourier methods for the description of wing shape in mosquitoes (diptera: Culicidae). *Systematic Biology*, 33:302–317. 15

[163] Rügamer, D. and Greven, S. (2020). Inference for L_2-Boosting. *Statistics and Computing*, 30(2):279–289. 22

[164] Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 11(4):735–757. 5

[165] Rügamer, D., Kolb, C., Fritz, C., Pfisterer, F., Kopper, P., Bischl, B., Shen, R., Bukas, C., Sousa, L. B. d. A. e., Thalmeier, D., Baumann, P., Kook, L., Klein, N., and Müller, C. L. (2021). deepregression: a flexible neural network framework for semi-structured deep distributional regression. 17

[166] Saha, A., Banerjee, S., Kurtek, S., Narang, S., Lee, J., Rao, G., Martinez, J., Bharath, K., Rao, A. U., and Baladandayuthapani, V. (2016). Demarcate: Density-based magnetic resonance image clustering for assessing tumor heterogeneity in cancer. *NeuroImage:*

Clinical, 12:132–143. 8

[167] Sarkar, S. and Panaretos, V. M. (2021). Covnet: Covariance networks for functional data on multidimensional domains. *arXiv preprint arXiv:2104.05021*. 5

[168] Scealy, J. L. and Welsh, A. H. (2011). Regression for compositional data by using distributions defined on the hypersphere. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):351–375. 8

[169] Schiratti, J.-B., Allassonniere, S., Routier, A., Colliot, O., Durrleman, S., Initiative, A. D. N., et al. (2015). A mixed-effects model with time reparametrization for longitudinal univariate manifold-valued data. In *International Conference on Information Processing in Medical Imaging*, pages 564–575. Springer. 23

[170] Schuss, Z. (2009). *Theory and applications of stochastic processes: an analytical approach*, volume 170. Springer Science & Business Media. 4

[171] Severn, K. E., Dryden, I. L., and Preston, S. P. (2022). Manifold valued data analysis of samples of networks, with applications in corpus linguistics. *The Annals of Applied Statistics*, 16(1):368–390. 2

[172] Shang, H. L. (2014). A survey of functional principal component analysis. *AStA Advances in Statistical Analysis*, 98(2):121–142. 3, 7

[173] Sharma, A. and Gerig, G. (2020). Trajectories from distribution-valued functional curves: A unified Wasserstein framework. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 343–353. Springer. 7

[174] 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*, pages 192–199. Springer. 17

[175] Sørensen, H., Goldsmith, J., and Sangalli, L. M. (2013). An introduction with medical applications to functional data analysis. *Statistics in medicine*, 32(30):5222–5240. 2

[176] 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. 7, 8

[177] Srivastava, A., Klassen, E., Joshi, S. H., and Jermyn, I. H. (2011a). Shape analysis of elastic curves in Euclidean spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(7):1415–1428. 11, 16

[178] Srivastava, A. and Klassen, E. P. (2016). *Functional and Shape Data Analysis*. Springer-Verlag. 2, 8, 12, 16

[179] Srivastava, A., Wu, W., Kurtek, S., Klassen, E., and Marron, J. S. (2011b). Registration of functional data using Fisher-Rao metric. *arXiv: Statistics Theory*. 11

[180] Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., and De Bastiani, F. (2017). *Flexible regression and smoothing: using GAMLSS in R*. CRC Press. 17

[181] Stephens, M. A. (1980). The von Mises distribution in p-dimensions with applications. Technical report, STANFORD UNIV CA DEPT OF STATISTICS. 8

[182] Stephens, M. A. (1982). Use of the von Mises distribution to analyse continuous proportions. *Biometrika*, 69(1):197–203. 8

[183] Stigler, S. M. (2007). The epic story of maximum likelihood. *Statistical Science*, pages

598–620. 7

[184] Strait, J., Kurtek, S., Bartha, E., and MacEachern, S. N. (2017). Landmark-constrained elastic shape analysis of planar curves. *Journal of the American Statistical Association*, 112(518):521–533. 16

[185] Stroup, W. W. (2012). *Generalized linear mixed models: modern concepts, methods and applications*. CRC press. 20

[186] Su, Z., Klassen, E., and Bauer, M. (2017). The square root velocity framework for curves in a homogeneous space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–10. 16

[187] Sun, J., Xie, Y., Ye, W., Ho, J., Entezari, A., Blackband, S. J., and Vemuri, B. C. (2013). Dictionary learning on the manifold of square root densities and application to reconstruction of diffusion propagator fields. In *International Conference on Information Processing in Medical Imaging*, pages 619–631. Springer. 8

[188] Talská, R., Hron, K., and Grygar, T. M. (2021). Compositional scalar-on-function regression with application to sediment particle size distributions. *Mathematical Geosciences*, 53(7):1667–1695. 10

[189] Talská, R., Menafoglio, A., Machalová, J., Hron, K., and Fišerová, E. (2018). Compositional regression with functional response. *Computational Statistics & Data Analysis*, 123:66–85. 10

[190] Talská, R., Menafoglio, A., Hron, K., Egozcue, J. J., and Palarea-Albaladejo, J. (2020). Weighting the domain of probability densities in functional data analysis. *Stat*, 9(1):e283. 10

[191] Thomas, J., Mayr, A., Bischl, B., Schmid, M., Smith, A., and Hofner, B. (2018). Gradient boosting for distributional regression: faster tuning and improved variable selection via noncyclical updates. *Statistics and Computing*, 28(3):673–687. 17, 22

[192] Thompson, D. W. (1945). *On Growth and Form*. Cambridge: Univercity Press, New York: The Macmillian Company, 2nd edition. 12

[193] Trouvé, A. and Vialard, F.-X. (2012). Shape splines and stochastic shape evolutions: A second order point of view. *Quarterly of Applied Mathematics*, pages 219–251. 2

[194] Tucker, J. D., Wu, W., and Srivastava, A. (2013a). Generative models for functional data using phase and amplitude separation. *Computational Statistics & Data Analysis*, 61:50–66. 3, 11

[195] Tucker, J. D., Wu, W., and Srivastava, A. (2013b). Generative models for functional data using phase and amplitude separation. *Computational Statistics & Data Analysis*, 61:50 – 66. 12

[196] Umlauf, N., Adler, D., Kneib, T., Lang, S., and Zeileis, A. (2015). Structured additive regression models: An R interface to BayesX. *Journal of Statistical Software*, 63(21):1–46. 17

[197] Umlauf, N., Klein, N., and Zeileis, A. (2018). BAMLSS: Bayesian additive models for location, scale and shape (and beyond). *Journal of Computational and Graphical Statistics*, 27(3):612–627. 17, 18

[198] Valenzuela, N., Adams, D. C., Bowden, R. M., and Gauger, A. C. (2004). Geometric morphometric sex estimation for hatchling turtles: a powerful alternative for detecting

subtle sexual shape dimorphism. *Copeia*, 2004(4):735–742. 13

[199] van den Boogaart, K., Egozcue, J. J., and Pawlowsky-Glahn, V. (2010). Bayes linear spaces. *SORT: Statistics and Operations Research Transactions*, 34(4):201–222. 9

[200] 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. 2, 7, 9

[201] van den Boogaart, K. G. and Tolosana-Delgado, R. (2013). *Analyzing compositional data with R*, volume 122. Springer. 8

[202] van den Boogaart, K. G., Tolosana-Delgado, R., and Templ, M. (2015). Regression with compositional response having unobserved components or below detection limit values. *Statistical Modelling*, 15(2):191–213. 9

[203] Varmuza, K., Steiner, I., Glinsner, T., and Klein, H. (2002). Chemometric evaluation of concentration profiles from compounds relevant in beer ageing. *European Food Research and Technology*, 215(3):235–239. 8

[204] Verbelen, R., Antonio, K., and Claeskens, G. (2018). Unravelling the predictive power of telematics data in car insurance pricing. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(5):1275–1304. 9

[205] 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

[206] Wang, G., Laga, H., and Srivastava, A. (2021). On the statistical analysis of complex tree-shaped 3D objects. *arXiv preprint arXiv:2110.08693*. 16

[207] Wang, H. and Marron, J. (2007). Object oriented data analysis: Sets of trees. *The Annals of Statistics*, 35(5):1849–1873. 2, 3

[208] Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (2015). Functional data analysis. *The Annual Review of Statistics and Its Application*, 1:41. 3, 23

[209] Wang, K. and Gasser, T. (1997). Alignment of curves by dynamic time warping. *The Annals of Statistics*, 25(3):1251–1276. 12

[210] Watson, D. (1988). Natural neighbor sorting on the n-dimensional sphere. *Pattern Recognition*, 21(1):63–67. 8

[211] Weinberger, K. Q. and Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(2). 3

[212] Wood, S. (2017). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2 edition. 4, 17, 20

[213] 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(516):1548–1563. 5

[214] Wrobel, J., Zipunnikov, V., Schrack, J., and Goldsmith, J. (2019). Registration for exponential family functional data. *Biometrics*, 75(1):48–57. 11

[215] Xie, W., Kurtek, S., Bharath, K., and Sun, Y. (2017). A geometric approach to visualization of variability in functional data. *Journal of the American Statistical Association*, 112(519):979–993. 12

[216] Yao, F., Müller, H., and Wang, J. (2005). Functional data analysis for sparse longitu-

dinal data. *Journal of the American Statistical Association*, 100(470):577–590. 5, 6

[217] Yee, T. W. (2010). The VGAM package for categorical data analysis. *Journal of Statistical Software*, 32:1–34. 17

[218] Yoshiyuki, A. (2017). A Functional Linear Regression Model in the Space of Probability Density Functions. Discussion papers 17015, Research Institute of Economy, Trade and Industry (RIETI). 10

[219] You, K. and Park, H.-J. (2021). Re-visiting Riemannian geometry of symmetric positive definite matrices for the analysis of functional connectivity. *NeuroImage*, 225:117464. 2

[220] Zemel, Y. and Panaretos, V. M. (2019). Fréchet means and Procrustes analysis in Wasserstein space. *Bernoulli*, 25(2):932–976. 7

[221] Zhang, C., Kokoszka, P., and Petersen, A. (2022a). Wasserstein autoregressive models for density time series. *Journal of Time Series Analysis*, 43(1):30–52. 7

[222] Zhang, Z. and Saparbayeva, B. (2021). Amplitude mean of functional data on $\mathbb{S}^2$. *arXiv preprint arXiv:2107.13721*. 16

[223] Zhang, Z., Wu, Y., Xiong, D., Ibrahim, J. G., Srivastava, A., and Zhu, H. (2022b). LESA: Longitudinal elastic shape analysis of brain subcortical structures. *Journal of the American Statistical Association*. 23

[224] 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(487):1203–1212. 17

[225] Ziegler, A. (2011). *Generalized estimating equations*, volume 204. Springer Science & Business Media. 20

[226] 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. 2, 13, 15

[227] Ziezold, H. (1994). Mean figures and mean shapes applied to biological figure and shape distributions in the plane. *Biometrical Journal*, 36(4):491–510. 15

Part I.

Distributional Regression for Functions and Functional Regression for Distributions

2. Boosting Functional Response Models for Location, Scale and Shape with an Application to Bacterial Competition

Already the example of growth curves, an archetype of functional data, illustrates that it can be problematic to (implicitly) assume target curves to be point-wise normally distributed as often done (as opposed to strictly positive, often skewed distribution) and exclusive modeling of the mean in dependence on covariates is restrictive (e.g., if external factors influence the whole growth process). Facing such challenges in an experimental setup of bacterial competition, we extend functional additive models (FAM) to generalized models for location, scale and shape (GAMLSS) in this contribution. Fitting models via gradient boosting, we illustrate the important role of the implied regularization on performance in this setting. The flexibility both in terms of response distribution and covariate effects lets us account for experimental details in the bacterial data scenario and identify different phases of bacterial interaction.

Contributing article:
Stöcker, A., Brockhaus, S., Schaffer, S., 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.
DOI: 10.1177/1471082X20917586.
Supplementary material provided in Appendix A.

Declaration on personal contributions:
Starting the project in his master's thesis, the author of the thesis has conducted major parts of this project including writing, data application and simulation studies. He also initiated the collaboration underlying the data problem. During his doctoral studies, he revised all parts prepared during the master's thesis extending the applied model and simulations.

Statistical Modelling 2021; **21(5)**: 385–404

Boosting functional response models for location, scale and shape with an application to bacterial competition

Almond Stöcker[1], Sarah Brockhaus[2], Sophia Anna Schaffer[3], Benedikt von Bronk[3], Madeleine Opitz[3] and Sonja Greven[1]

[1]School of Business and Economics, HU Berlin, Germany.
[2]Department of Statistics, LMU Munich, Germany.
[3]Department of Physics, LMU Munich, Germany.

Abstract: We extend generalized additive models for location, scale and shape (GAMLSS) to regression with functional response. This allows us to simultaneously model point-wise mean curves, variances and other distributional parameters of the response in dependence of various scalar and functional covariate effects. In addition, the scope of distributions is extended beyond exponential families. The model is fitted via gradient boosting, which offers inherent model selection and is shown to be suitable for both complex model structures and highly auto-correlated response curves. This enables us to analyse bacterial growth in *Escherichia coli* in a complex interaction scenario, fruitfully extending usual growth models.

Key words: bacterial growth, distributional regression, functional data, functional regression, GAMLSS

Received May 2019; **revised** December 2019; **accepted** March 2020

1 Introduction

In functional data analysis (Ramsay and Silverman, 2005), functional response regression aims at estimating covariate effects on response curves (Morris, 2015; Greven and Scheipl, 2017). The response curves might, for instance, be given by annual temperature curves, growth curves or spectroscopy data. We propose a flexible approach to regression with functional response allowing for simultaneously modelling multiple distributional characteristics of response curves following potentially non-Gaussian distribution families. It therefore generalizes usual functional mean regression models.

The problem of non-Gaussian functional response appears in many applications and is, accordingly, addressed in several publications. Following early works on non-Gaussian functional data by Hall et al. (2008) and van der Linde (2009),

Address for correspondence: Almond Stöcker, School of Business and Economics, Chair of Statistics, Humboldt University of Berlin, Unter den Linden 6, D–100 99 Berlin, Germany.
E-mail: almond.stoecker@hu-berlin.de

 10.1177/1471082X20917586

authors such as Goldsmith et al. (2015), Wang and Shi (2014) and Scheipl et al. (2016) proposed generalized linear mixed model (GLMM)-type regression models, which are suitable for, for example, positive, discrete or integer-valued response functions. The linear predictor for the mean function is typically composed of a covariate effect term and a latent random Gaussian error process accounting for auto-correlation, and combined with a link function. Li et al. (2014) jointly model continuous and binary-valued functional responses in a similar fashion, but without considering covariates. Other ideas to account for the auto-correlation of functional response curves include robust covariance estimation for valid inference after estimation under a working independence assumption (Gertheiss et al., 2015) and an overall regularization by early stopping a gradient boosting fitting procedure guided by curve-wise re-sampling methods as applied by Brockhaus et al. (2015, 2017) besides using curve-specific smooth errors. Moreover, this latter approach also offers quantile regression for functional data. The above approaches present important steps in generalizing functional regression models. In particular, our approach is a direct generalization of the framework of Brockhaus et al. (2015). However, the previous methods are restricted to one predictor such that none of them allows for simultaneously modelling also the response variance or other distributional parameters in a similar fashion as the mean. Staicu et al. (2012) propose a method for estimating mean, variance and other shape parameter functions nonparametrically and also for non-Gaussian point-wise distributions, while modelling auto-correlation via copulas. However, they do not allow for including covariate effects. Only the framework of Scheipl et al. (2016) now allows to perform simultaneous mean and variance regression in the Gaussian case (Greven and Scheipl, 2017) and is also the only one currently offering a comparable range of smooth/linear effects of scalar and functional covariates, which are implemented in the R package `refund` (Goldsmith et al., 2018). Thus, we will compare to their model in a simulation. However, they are so far restricted to the Gaussian special case and do not offer the flexibility to specify multiple predictors for other distributions and no more than two predictors, as we need, in particular, in the data scenario presented in this article. We overcome these limitations by introducing generalized additive models for location, scale and shape (GAMLSS) for functional responses. For the case of scalar response regression, GAMLSS were introduced by Rigby and Stasinopoulos (2005) extending usual generalized additive models (GAMs; Hastie and Tibshirani, 1990) to multiple distributional parameters. Each parameter of the assumed response distribution is modelled with a separate predictor depending on covariates, allowing, for example, for covariate effects on mean and variance. Hence, doubtful assumptions of homoscedasticity can be overcome. In addition to this extension, the range of applicable distributions of GAMs is also extended to non-exponential family distributions in the GAMLSS framework. Brockhaus et al. (2018a) discussed GAMLSS scalar-on-function regression based on the flexible gradient boosting regression framework introduced by Bühlmann and Hothorn (2007) combining a scalar GAMLSS framework developed by Mayr et al. (2012) and Thomas et al. (2018) and the flexible functional regression framework of Brockhaus et al. (2017). We further extend this to functional GAMLSS for

function-on-scalar and function-on-function regression, such that our framework now offers the full flexibility of scalar GAMLSS also for functional responses and covariates.

We apply this framework to analyse bacterial interaction: as bacterial resistances increase, producing effective antibiotics gets harder and harder. Understanding bacterial interactions might help finding alternatives. In particular, we analyse growth curves of two competing *Escherichia coli* bacteria strains (von Bronk et al., 2017)—a toxin producing 'C-strain' and a toxin sensitive 'S-strain'—to obtain insights into the underlying growth affecting bacterial interaction. Our aim is to model the S-strain growth behaviour in dependence on the toxin emitting C-strain and under different experimental conditions, and allow this dependence to affect both mean and variability of growth as well as the extinction probability. This requires a functional response regression model for several parameters of the non-Gaussian response distribution with linear and smooth effects of functional and scalar covariates.

There are various approaches applied to modelling bacterial growth curves in the literature. Gompertz and Baranyi-Roberts models are two common parametric approaches to modelling growth curves (see, e.g., López et al., 2004; Perni et al., 2005). Weber et al. (2014) implement a model particularly for analysing bacterial interaction. The models are usually fitted using least squares methods, which corresponds to assuming a Gaussian distribution of bacterial propagation. This is problematic as response values are naturally positive and very small in the beginning, starting from single cell level. Thus, also assuming a constant variance over the whole time span seems not appropriate. Moreover, they do not offer the opportunity to include covariate effects for modelling, for example, the impact of external factors. Thus, these models are not applicable here. In addition, they are often highly non-linear, which may introduce problems in parameter estimation. Gasser et al. (1984) discuss this point and some further advantages of nonparametric growth curve regression over parametric models. They propose a kernel method for this purpose, which does, however, not include covariates and lacks the flexibility needed for our purposes. Still, non- or semi-parametric functional regression models present a natural choice from a statistical perspective, also because they can approximate the above parametric growth models very well when these are appropriate (Online Appendix E.2).

Besides providing the flexibility to meet all the challenges arising from the present analysis of bacterial interaction, we can show in extensive simulation studies that the presented approach is indeed well suited for complex scenarios with highly auto-correlated response curves despite working independence assumption: early stopping the gradient boosting algorithm based on curve-wise re-sampling techniques plays a key role in avoiding over-fitting and leads to highly improved estimation quality when comparing to the approach of Greven and Scheipl (2017).

The approach is implemented in the R (R Core Team, 2018) add-on package `FDboost` (Brockhaus and Ruegamer, 2018). Brockhaus et al. (2018b) provide a tutorial article to the package. Even though the discussion and illustration of GAMLSS focuses on the scalar-on-function and not the functional response case, we recommend it as a general software introduction.

The remainder of the article is structured as follows: In Section 2, we formulate the general model and describe the fitting algorithm. In Section 3, we apply the proposed model to analysing *E. coli* bacteria growth. Section 4 provides the results of two simulation studies for Gaussian response curves as well as for the growth model. Section 5 concludes with a discussion. Further details concerning the model, application and simulation studies are provided as Online Supplement, as well as the code for the simulations and fitting of the model with the R-package `FDboost`.

2 Model formulation

Consider a data scenario with N observations of a functional response Y and respective covariates $\mathbf{X}$. Y is a stochastic process, such that its realized trajectories $y_i : \mathcal{T} \to \mathbb{R},\ t \mapsto y_i(t)$ for $i = 1, ..., N$ represent the response curves over an index set $\mathcal{T}$. For notational simplicity, we assume that response curves are observed on a common grid $\mathcal{T}_0 \subset \mathcal{T}$, where $\mathcal{T} = [0, t_{max}]$ is a real interval starting at zero and $\mathcal{T}_0$ a finite discrete set of evaluation points. However, the curves could be measured on different grids as well. As this is the case in many applications, the variable t is referred to as time variable. Scalar response is contained as special case where $\mathcal{T}$ is a one point set. Let $\mathbf{x}_i = (x_{i,1}, ..., x_{i,p})^\top$ denote the i-th observed covariate vector, that is, realization of $\mathbf{X}$, which can contain scalar and functional covariates. A functional covariate may have a different domain $\mathcal{S}$ from the response and is denoted as $x_{i,j} : \mathcal{S} \to \mathbb{R}$, $s \mapsto x_{i,j}(s)$. We suppress potential dependence of $\mathcal{S}$ on j in our notation.

We assume that for all $t \in \mathcal{T}$ the point-wise response distribution $\mathcal{F}_{Y(t)|\mathbf{X}}$ is known up to the distribution parameters $\boldsymbol{\vartheta}(t) = \left(\vartheta^{(1)}(t), ..., \vartheta^{(Q)}(t)\right)^\top$. For instance, for a Gaussian process, the parameters might represent the conditional mean and variance over time, that is, $\vartheta^{(1)}(t) = \mathbb{E}(Y(t)|\mathbf{X} = \mathbf{x})$ and $\vartheta^{(2)}(t) = \mathbb{V}\text{ar}(Y(t)|\mathbf{X} = \mathbf{x})$, suppressing the dependence on $\mathbf{x}$ in the notation. For each parameter an additive regression model is assumed. The model is specified by

$$g^{(q)}(\vartheta^{(q)}) = h^{(q)}(\mathbf{x}) = \sum_{j=1}^{J^{(q)}} h_j^{(q)}(\mathbf{x}) \qquad ,\ q = 1, ..., Q,$$

where $g^{(q)}$ is a monotonic link function for the qth distribution parameter. This model structure corresponds to the GAMLSS introduced by Rigby and Stasinopoulos (2005). However, covariates and response may now be functions. Correspondingly, $\vartheta^{(q)} := \vartheta^{(q)}(.)$ and the predictor $h^{(q)}(\mathbf{x}) := h^{(q)}(\mathbf{x}, .)$ are now functions over the domain $\mathcal{T}$ of the response, for $q = 1, ..., Q$. For $Q = 1$ parameter corresponding to the mean, the model reduces to the functional additive regression model of Scheipl et al. (2015, 2016) and Brockhaus et al. (2015).

Both covariate and time dependency are modelled within the additive predictor via the effect functions $h_j^{(q)}(\mathbf{x}, t)$. The predictor is typically composed of a functional

Table 1 Overview of possible effect types (adapted from Brockhaus et al., 2015)

Covariate(s)	Type of Effect	$h_j^{(q)}$
(none)	Smooth intercept	$\beta_0(t)$
Scalar covariate z	Linear effect	$z\,\beta(t)$
	Smooth effect	$f(z, t)$
Two scalars z_1, z_2	Linear interaction	$z_1\,z_2\,\beta(t)$
	Functional varying coefficient	$z_1\,f(z_2, t)$
	Smooth interaction	$f(z_1, z_2, t)$
Grouping variable g	Group-specific intercept	$\beta_g(t)$
Group. variable g, scalar z	Group-specific linear effect	$z\,\beta_g(t)$
	Group-specific smooth effect	$f_g(z, t)$
Group. variables g_1, g_2	Group-interaction	$\beta_{g_1,g_2}(t)$
Functional covariate $x(s)$	Functional linear effect	$\int x(s)\,\beta(s, t)\,ds$
Functional cov. $x(s)$, scalar z	Linear interaction	$z \int x(s)\,\beta(s, t)\,ds$
	Smooth interaction	$\int x(s)\,\beta(z, s, t)\,ds$
Functional cov. $x(s)$ over $\mathcal{T}$	Concurrent effect	$x(t)\beta(t)$
	Historical effect	$\int_0^t x(s)\,\beta(s, t)\,ds$
	Effect with t-specific integration limits	$\int_{l(t)}^{u(t)} x(s)\,\beta(s, t)\,ds$

intercept $h_1^{(q)}(\mathbf{x}, t) = \beta_0(t)$ and linear or smooth covariate effects $h_j^{(q)}(\mathbf{x}, t)$, of which each depends on one or more covariates. The construction of the effects follows a modular principle, which allows for flexible specification of effect types and is outlined in the next subsection. Table 1 gives an overview of different effect types available and Section 2.2 will discuss inherent selection of effects within the fitting approach.

Apart from the Gaussian, a variety of other distributions can be specified for the response. In principle, $\mathcal{F}_{Y(t)|\mathbf{X}}$ can be any distribution for which both the likelihood and its derivatives are computable. The derivatives with respect to the parameters are required for the model estimation via gradient boosting. For functional response, usually only continuous distributions are under consideration. However, this does not necessarily have to be the case (see, e.g., Scheipl et al., 2016). As we built on the approach of Mayr et al. (2012) for scalar GAMLSS, all of the distributions implemented in their R package `gamboosLSS` (Hofner et al., 2017) are directly available for the present boosting approach. Moreover, they also provide an interface to use the comprehensive list of distributions available in the R package `gamlss.dist` (Stasinopoulos and Rigby, 2019) and custom distributions can be specified.

2.1 Construction of effect functions $\mathrm{h}_j^{(q)}$

As both covariate and time dependency of the functional response are specified by the effect functions $h_j^{(q)}$, they play a key role in the framework. We briefly illustrate their modular structure and refer to Brockhaus et al. (2018b) and Greven and Scheipl (2017) for further examples and details, as their construction is not new to the functional GAMLSS. The novelty is, however, that we may now use them in multiple predictors for multiple parameter functions.

For each effect type, $h_j^{(q)}$ is represented by a linear combination of specified basis functions, such that the predictor is linear in its coefficients. Multivariate basis functions are constructed as tensor products of univariate bases providing flexible modular means of specification (cf. Scheipl et al., 2015), giving the basis representation

$$h_j^{(q)}(\mathbf{x}, t) = \left(\mathbf{b}_{Xj}^{(q)}(\mathbf{x}, t) \otimes \mathbf{b}_{Yj}^{(q)}(t)\right)^\top \boldsymbol{\theta}_j^{(q)} \qquad , t \in \mathcal{T}\,. \tag{2.1}$$

A vector $\mathbf{b}_{Yj}^{(q)}$ of $K_{Yj}^{(q)}$ basis functions for the time variable is combined with a vector $\mathbf{b}_{Xj}^{(q)}$ of $K_{Xj}^{(q)}$ basis functions for the covariate effects. The basis $\mathbf{b}_{Xj}^{(q)}(\mathbf{x}, t)$ might be time dependent, for example, for a functional historical effect. However, for many effect types, it only depends on covariates, such that we can write $\mathbf{b}_{Xj}^{(q)}(\mathbf{x})$. Applying the Kronecker product $\otimes$, a new basis is obtained. Its elements correspond to the pairwise products of elements in $\mathbf{b}_{Yj}^{(q)}$ and $\mathbf{b}_{Xj}^{(q)}$. For details, see Online Supplement A.1. The coefficient vector $\boldsymbol{\theta}_j^{(q)} \in \mathbb{R}^{K_{Yj}^{(q)} K_{Xj}^{(q)}}$ specifies the concrete form of the effect. Fitting the model corresponds to estimating $\boldsymbol{\theta}_j^{(q)}$ for all effect functions.

A typical choice for $\mathbf{b}_{Yj}^{(q)}(t)$ is a spline basis. Then, in case of time-independent covariate basis functions $\mathbf{b}_{Xj}^{(q)}$, $h_j^{(q)}(\mathbf{x}_0, t)$ describes a spline curve for a fixed value $\mathbf{x} = \mathbf{x}_0$ of the covariate. Usually, quadratic penalty terms are employed in order to control smoothness of the effect functions (see Section 2.2). A typical effect function $h_j^{(q)}(\mathbf{x}, t)$ depends on a single covariate. For example, for a linear effect $z\beta(t)$ of a scalar covariate z, this yields $\mathbf{b}_{Xj}^{(q)}(\mathbf{x}, t) = \mathbf{b}_{Xj}^{(q)}(z) = z$. In order to obtain a smooth covariate effect $f(z, t)$, a spline basis can be chosen for $\mathbf{b}_{Xj}^{(q)}(z)$ just like for the time curve, yielding a tensor product spline basis in (2.1). For a functional covariate $x : \mathcal{T} \mapsto \mathbb{R}$ a historical effect of the form $\int_0^t x(s)\,\beta(s, t)\,ds$ can be constructed using a basis of time-dependent linear functionals $\mathbf{b}_{j,k}^{(q)}(x, t) = \int_0^t x(s)\varphi_k(s)\,ds$, where $\varphi_k(s)$, $k = 1, \ldots, K_{Xj}^{(q)}$, is a spline basis and the integral is numerically approximated over the observation grid of x in

$\mathcal{T}$. Using also a spline basis for $\mathbf{b}_{Yj}^{(q)}(t)$, this corresponds to specifying a tensor product spline basis for $\beta(s, t)$.

2.2 Model fit

Component-wise gradient boosting is a gradient descend method for model fitting, where the model is iteratively updated. In each iteration, the algorithm aims at minimizing a loss function following the direction of its steepest descent. Instead of updating the full additive predictor at once, the individual effect functions $h_j^{(q)}$ are separately fit to the negative gradient in a component-wise approach. These individual effect models are called *base-learners*, as they present simple base models that jointly form the model predictor. In each iteration, only the effect function with the best fit is updated with a step length ν in the direction of its fit. The component-wise and stepwise procedure yields automated model selection and allows for fitting models with more parameters than observations.

Let $f\left(y(t)\middle|\boldsymbol{\vartheta}(t)\right) = f_h\left(y\,(t)\;\middle|\,\mathbf{h}\,(\mathbf{x}, t)\right)$ with $\mathbf{h} = (h^{(1)}, \ldots, h^{(\mathcal{Q})})^\top$ denote the conditional probability density function (PDF) of the response at $t \in \mathcal{T}$ for a given parameter setting. We define the point-wise loss function to be the negative log-likelihood

$$\varrho\left(y\,(t)\,, \mathbf{h}\,(\mathbf{x}, t)\right) = -\log f_h\left(y\,(t)\;\middle|\,\mathbf{h}\,(\mathbf{x}, t)\right)\,.$$

The functional GAMLSS loss function is then obtained as

$$\ell(y, \mathbf{h}(\mathbf{x})) = \int_{\mathcal{T}} \varrho\left(y(\,t\,), \mathbf{h}(\,\mathbf{x}, t\,)\right) dt\,,$$

the integral over the point-wise loss functions over $\mathcal{T}$. Therefore, we assume that f_h and $\mathbf{h}$ are chosen such that the integral exists, which is no restriction in practice. The aim of gradient boosting is to find the predictor

$$\mathbf{h}_{optimal} = \underset{\mathbf{h}}{\operatorname{argmin}}\; \mathbb{E}\left[\,\ell(Y, \mathbf{h}(\mathbf{X}))\,\right] = \underset{\mathbf{h}}{\operatorname{argmin}} \int_{\mathcal{T}} \mathbb{E}\left[\,\varrho\left(Y(\,t\,), \mathbf{h}(\,\mathbf{X}, t\,)\right)\,\right] dt \tag{2.2}$$

minimizing the expected loss.

Based on data $(y_i, \mathbf{x}_i)_{i=1,\ldots,N}$, this is estimated by optimizing the empirical mean loss. Hence, the estimated predictor vector $\hat{\mathbf{h}} = (\hat{h}^{(1)}, \ldots, \hat{h}^{(\mathcal{Q})})^\top$ is given by

$$\hat{\mathbf{h}} \approx \underset{\mathbf{h}}{\operatorname{argmin}}\; \frac{1}{N} \sum_{i=1}^{N} \hat{\ell}(y_i, \mathbf{h}(\mathbf{x}_i))\,, \tag{2.3}$$

where $\hat{\ell}(y_i, \mathbf{h}(\mathbf{x}_i)) = \sum_{t \in \mathcal{T}_0} \varrho\left(y_i(\,t\,), \mathbf{h}(\,\mathbf{x}_i, t\,)\right)$ is an approximation of the loss. However, to avoid over-fitting, the optimization is generally not run until convergence. Instead, a re-sampling strategy is employed to find an optimal stopping iteration.

The minimization in (2.2) can be seen to minimize the Kullback–Leibler divergence (KLD) of the model density f_h to the true underlying density. Hastie and Tibshirani (1990) formulate a similar regression aim for GAMs. However, in the functional case, we consider the point-wise KLD integrated over the domain $\mathcal{T}$.

The base-learners fitted in each boosting iteration correspond to the effects $h_j^{(q)}(\mathbf{x}, t)$ with $j = 1, ..., J^{(q)}$ and $q = 1, ..., Q$. For any given loss function, they represent single regression models, which are fitted to the gradient of the loss function via penalized least squares. The coefficients $\boldsymbol{\theta}_j^{(q)}$ of the respective $h_j^{(q)}$, as defined in equation (2.1), are subject to a quadratic penalty of the form $(\boldsymbol{\theta}_j^{(q)})^\top \mathbf{P}_j^{(q)} \boldsymbol{\theta}_j^{(q)}$, where $\mathbf{P}_j^{(q)}$ is a penalty matrix. As described for bivariate smooth terms, for example, in Wood (2006) or Brockhaus et al. (2015), the penalty matrix is constructed as $\mathbf{P}_j^{(q)} = \lambda_{Xj}^{(q)} \left(\mathbf{P}_{Xj}^{(q)} \otimes \mathbf{I}_{K_{Yj}^{(q)}} \right) + \lambda_{Yj}^{(q)} \left(\mathbf{I}_{K_{Xj}^{(q)}} \otimes \mathbf{P}_{Yj}^{(q)} \right)$ with smoothing parameters $\lambda_{Yj}^{(q)}, \lambda_{Xj}^{(q)} \geq 0$ and penalty matrices $\mathbf{P}_{Yj}^{(q)} \in \mathbb{R}^{K_{Yj}^{(q)} \times K_{Yj}^{(q)}}$ and $\mathbf{P}_{Xj}^{(q)} \in \mathbb{R}^{K_{Xj}^{(q)} \times K_{Xj}^{(q)}}$ for the time basis $\mathbf{b}_{Yj}^{(q)}(t)$ and covariate basis $\mathbf{b}_{Xj}^{(q)}(\mathbf{x}, t)$, respectively. For instance, a common choice for B-spline bases is a first, or second-order difference penalty matrix yielding P-Splines (compare Eilers and Marx, 2010). Base-learners for group effects might be regularized with a ridge penalty. If no penalization should be applied for either the response or the covariates, this can also be obtained by setting $\lambda_{Yj}^{(q)} = 0$ or $\lambda_{Xj}^{(q)} = 0$, respectively. Thomas et al. (2018) compare different gradient boosting methods for GAMLSS, which can all be analogously generalized to functional response. While the 'cyclic' method and a 'non-cyclic' method are available in the `R` package `gamboostLSS`, only the algorithm of the 'non-cyclic' method is described here in detail. Comparing it to the 'cyclic' method, it performed better in simulations (Online Appendix Table 3), is faster (Online Appendix Figure 11) and provides the advantage of unified model selection across parameters $\vartheta^{(q)}$, $q = 1, \ldots, Q$.

Algorithm: gradient boosting for functional GAMLSS

1. **To set up the model specify**

   (a) **a functional loss function** ℓ with point-wise loss ϱ corresponding to the assumed response distribution with Q distribution parameters

   (b) **the base-learners** by choosing the desired bases for the effects $h_j^{(q)}(\mathbf{x}, t) = (\mathbf{b}_{Xj}^{(q)}(\mathbf{x}, t) \otimes \mathbf{b}_{Yj}^{(q)}(t))^\top \boldsymbol{\theta}_j^{(q)}$, penalty matrices $\mathbf{P}_j^{(q)}$ for all $j = 1, ..., J^{(q)}$ and $q = 1, ..., Q$ and their respective smoothing parameters.

   (c) **gradient boosting hyper-parameters:** the step-lengths $\nu^{(q)} \in]0, 1]$ for $q = 1, ..., Q$ and the maximum number of iterations m_{stop}.

   Initialize the coefficients $\boldsymbol{\theta}_j^{(q)[0]}$ for the initial predictor $\mathbf{h}^{[0]}(\mathbf{x}_i, t)$, for example, to $\mathbf{0}$, and set $m = 0$.

2. **For** $m = 0, ..., m_{stop} - 1$ **iterate:**

   i) **Find best update for each distribution parameter.**
   For $q = 1, ..., Q$ **do:**

(a) **Evaluate negative partial gradients** for $i = 1, ..., N$ at the current predictor $\mathbf{h}^{[m]}$

$$u_i^{(q)}(t) := -\frac{\partial \varrho}{\partial h^{(q)}}\left(y_i(t), \mathbf{h}\right)\Big|_{\mathbf{h}=\mathbf{h}^{[m]}(\mathbf{x}_i,t)}$$

(b) **Fit base-learners to the gradients,** that is, for $j = 1, ..., J^{(q)}$ find $\tilde{\boldsymbol{\theta}}_j^{(q)}$ with

$$\tilde{\boldsymbol{\theta}}_j^{(q)} := \underset{\boldsymbol{\theta}_j^{(q)}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \sum_{t\in\mathcal{T}_0} \left(u_i^{(q)}(t) - \left(\mathbf{b}_{Xj}^{(q)}(\mathbf{x}_i, t) \otimes \mathbf{b}_{Yj}^{(q)}(t)\right)^\top \boldsymbol{\theta}_j^{(q)} \right)^2 + (\boldsymbol{\theta}_j^{(q)})^\top \mathbf{P}_j^{(q)} \boldsymbol{\theta}_j^{(q)} \right\}$$

- **Determine the best-fitting base-learner** with index $\tilde{\jmath}$ following the least squares criterion

$$\tilde{\jmath} := \underset{j}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t\in\mathcal{T}_0} \left(u_i^{(q)}(t) - (\mathbf{b}_{Xj}^{(q)}(\mathbf{x}_i, t) \otimes \mathbf{b}_{Yj}^{(q)}(t))^\top \tilde{\boldsymbol{\theta}}_j^{(q)} \right)^2$$

- **Determine updated predictor candidate,** that is, determine ${}^*_j\mathbf{h}$ where only the coefficients of the best-fitting base-learner are updated, such that the coefficients are given by

$${}^*_j\boldsymbol{\theta}_k^{(p)} = \begin{cases} \boldsymbol{\theta}_k^{(p)[m]} + \nu^{(p)} \tilde{\boldsymbol{\theta}}_k^{(p)} & \text{for } p = q, k = \tilde{\jmath}, \\ \boldsymbol{\theta}_k^{(p)[m]} & \text{else} \end{cases}$$

end for.

ii) **Select best update across the distributional parameters** and update the linear predictor accordingly

$$\mathbf{h}^{[m+1]} = \underset{{}^*_j\mathbf{h}}{\operatorname{argmin}} \sum_{i=1}^{N} \hat{\ell}\left(y_i, {}^*_j\mathbf{h}(\mathbf{x}_i)\right)$$

end for.

The smoothing parameters for the penalty matrices $\mathbf{P}_{jY}^{(q)}$ can be chosen indirectly specifying the base-learner degrees of freedom, as described by Hofner et al. (2011). They are typically specified such that equal degrees of freedom for all base-learners are attained to ensure a fair base-learner selection. Note that these degrees of freedom only specify the flexibility of each base-learner for one iteration, while the final effective degrees of freedom can be higher due to repeated selection of the same base-learner. $\nu = 0.1$ is a popular choice for the step-length (Bühlmann and Hothorn, 2007). It should be chosen small enough to prevent overshooting. Yet, too small values greatly increase computation time. The optimal stopping iteration m_{stop}, with respect to equation (2.2), is the main tuning parameter. It can be estimated using, for example, curve-wise cross-validation or bootstrapping. As determining $\mathbf{h}^{[m_{stop}]}$ involves computation of all earlier predictors, this can be done very efficiently (and in parallel over cross-validation folds). Early stopping induces regularization of effect functions and provides automated model selection: effect functions $h_j^{(q)}$ which were never selected drop out of the model. As each base-learner is fitted separately, models with more covariates than observations can be fit and computational effort scales linearly in the number of covariate effects. By appropriately decomposing terms

into, for example, a linear and a non-linear base-learner, we cannot only select covariates, but also distinguish linear effects from smooth effects depending on the same covariate (compare Kneib et al., 2009) and covariate interactions from additive marginal effects (see Online Appendix A.2).

3 Analysis of bacterial interaction in *E. coli*

The coexistence of various bacterial species is a key factor in environmental systems. Equilibria in this biodiversity stand or fall with the species' interaction. Certain bacteria strains produce toxins and use them to assert themselves in bacterial competition. von Bronk et al. (2017) establish an experimental set-up with two cohabiting *Escherichia coli* bacteria strains: a 'C-strain' producing the toxin ColicinE2 and a colicin sensitive 'S-strain' pipetted together on an agar surface. Single bacteria of the C-strain population sacrifice themselves in order to liberate colicin. The emitted colicin diffuses through the agar and kills numerous S-strain bacteria on contact. On the other hand, the S-strain might outgrow the C-strain and starts in a favoured position of an initial ratio S:C of about 100:1. The arising population dynamics are influenced by external stress induced with the antibiotic agent Mitomycin C (MitC). MitC slightly damages the DNA of the bacteria. While it has little effect on the S-strain, it triggers colicin production in the C-strain as an SOS-response. A higher dose of MitC increases the fraction of colicin producing C-bacteria and, thus, colicin emission (von Bronk et al., 2017).

At a total of $N = 334$ observation sites, bacteria under consideration are exposed to one of four different MitC concentrations. Bacterial growth curves $S_i(t)$ of the S-strain and $C_i(s)$ of the C-strain, $i = 1, \ldots, N$, are observed over 48 hours. Their values correspond to the propagation areas of the bacterial strains, which are obtained from the automated image segmentation procedure implemented by von Bronk et al. (2017). S- and C-strain areas can be distinguished as the bacteria are marked with red and green fluorescence, respectively. The resulting area growth curves are measured on a fixed time grid with $G = 105$ measurements per curve. The experiments are conducted in batches of about 40 bacterial spots and with two batches for each MitC concentration. In order to keep track of bacterial growth, the zoom level of the microscope was adjusted after $12\frac{1}{4}h$, $18\frac{1}{2}h$ and $33\frac{1}{2}h$. As the performance of the automatic bacterial area segmentation may depend on the zoom level, it has to be incorporated into the analysis.

3.1 Model for S-strain growth

In order to obtain insights into bacterial interaction dynamics, we model the i-th propagation area curve of the S-strain $S_i(t)$ in dependence on the C-strain growth and other covariates. While usually $S_i(t) > 0$, it might equal zero, if the S-strain is completely extinct or masked by the fluorescence of the C-strain. Therefore, we assume a conditional zero adjusted gamma (ZAGA) distribution

for $S_i(t)$, which is a mixed continuous and discrete distribution with its PDF given by $f_{ZAGA}\left(s_i \middle| \mu_i, {}^{\sigma_i}/_{\mu_i}, p_i\right) = p_i\, \delta_{s_i} + f_{GA}(s_i \middle| \mu_i, {}^{\sigma_i}/_{\mu_i})(1 - \delta_{s_i})$ with $\delta_{s_i} = 1$ if $s_i = 0$ and 0 otherwise and f_{GA} the density of a gamma distribution parametrized by its mean $\mu_i(t)$ and the coefficient of variation $\sigma_i(t)/\mu_i(t)$ with $\sigma_i(t)$ the standard deviation (Stasinopoulos and Rigby, 2019). This corresponds to some extent to the zero adjustment in a zero-inflated Poisson model. However, unlike the Poisson distribution, the gamma distribution is continuous and does not have a point mass at zero by itself. For $S_i(t) > 0$, it offers the flexibility to model both a location and a scale parameter conditional on the survival of the S-strain at time t, while in addition, we model the probability of extinction of the S-strain, $p_i(t) = P\,(S_i(t) = 0)$ over time. Each component of the resulting parameter vector $\boldsymbol{\vartheta}_i(t) = \left(\vartheta_i^{(\mu)}(t), \vartheta_i^{({}^{\sigma}/_{\mu})}(t), \vartheta_i^{(p)}(t)\right)^\top = \left(\mu_i(t), \frac{\sigma_i(t)}{\mu_i(t)}, p_i(t)\right)^\top$ is modelled as

$$g^{(q)}\left(\vartheta_i^{(q)}(t)\right) = \beta_0^{(q)}(t) + \beta_{MitC_i}^{(q)}(t) + \beta_{Batch_i}^{(q)}(t) + h_1^{(q)}(C_i\,, t) + h_2^{(q)}(C_i', t)$$

for $q \in \{\mu, {}^{\sigma}/_{\mu}, p\}$, with link-functions $g^{(\mu)} = g^{({}^{\sigma}/_{\mu})} = \log$ and $g^{(p)} = \text{logit}$, and with historical effects $h_j^{(q)}(C_i, t) = \int_0^t C_i(s)\beta_j^{(q)}(s, t)\, ds$ (compare Brockhaus et al., 2017).

For each distribution parameter, the model includes a functional intercept $\beta_0^{(q)}(t)$. As there are only four MitC concentrations employed, they are considered as categorical grouping variable and represented by group-specific intercepts $\beta_{MitC_i}^{(q)}(t)$ per MitC level centred around the functional intercept. As a functional random intercept, we include an additional group-specific intercept $\beta_{Batch_i}^{(q)}(t)$ to compensate for batch effects, which are centred around $\beta_{MitC_i}^{(q)}(t)$ in order to preserve identifiability of the functional intercept. The impact of the C-strain on S-strain growth is modelled using historical effects with coefficient functions $\beta_j^{(q)}(s, t)$. Historical effects are included both for the current C-strain propagation $C_i(s)$ and for its derivative $C_i'(s)$ reflecting the current C-strain growth. The covariate curves are centred around their empirical point-wise mean curve, such that $\frac{1}{N}\sum_{i=1}^{N} C_i(s) = 0$ for each s, and scaled with the corresponding standard deviation, such that $\text{sd}(C(s)) = 1$. For $C_i'(s)$ correspondingly. Doing so, the coefficient functions can be uniformly interpreted over the whole time span. By integrating, the historical effect includes information about the curves from time point $t = 0$ to the current time point t. For p, we include an additional step-function base-learner to capture the different zoom levels applied during the experiment at fixed known time points, which can lead to different visibility of small S populations. Corresponding effects are not expected to be necessary for μ and ${}^{\sigma}/_{\mu}$ and are thus not included, as they might even lead to spurious boundary effects in this experimental set-up. Apart from the step function, all effect functions are modelled with cubic P-splines and second-order difference penalties,

such that for the functional intercepts we penalize deviations from exponential growth when employing a log-link for the mean and the scale parameter. For the MitC and batch effects, a ridge-type penalty over factor levels is utilized to achieve the same number of effective degrees of freedom for all base-learners. A common step-length of $\nu = 0.1$ is used for μ, σ/μ and p. We fit the model with both implemented GAMLSS boosting methods and decide for the 'non-cyclic' method, described in Section 2.2, which performed better in ten fold curve-wise bootstrapping and is computationally more efficient. With a maximum of 3 000 boosting iterations, the model fit took less than 16 min on a 64-bit Windows laptop followed by 156 min of bootstrapping without parallelization. The latter can be easily accelerated by running it on several cores in parallel.

3.2 Results

3.2.1 MitC effect and effect of experimental batches

An overview of the effects of the toxin MitC can be found in Figure 1. We observe that mean S-strain growth is slightly increasing for low MitC levels compared to no MitC, but is particularly higher for $MitC_i = 0.1\,{}^{\mu g}/_{ml}$. This indicates that, if $S_i(t) \geq 0$, the S-strain even grows better under this condition.

For the standard deviation, we observe a gradual but distinct rise with the MitC level. Due to the log-link we may not only interpret effects on the shape parameter ${}^{\sigma}/_{\mu}$ but also on σ: effect functions $h_j^{(\sigma)}$ for σ are obtained as $h_j^{(\sigma)} = h_j^{(\mu)} + h_j^{({}^{\sigma}/_{\mu})}$. In this plot, we choose to depict σ instead of ${}^{\sigma}/_{\mu}$, as it is more straightforward to interpret on the response level. We observe that positive skewness increases with MitC concentration.

It is important to note, that control experiments indicate no considerable effect of MitC on S-strain growth (von Bronk et al., 2017). Thus, present covariate effects of MitC reflect effects of C-cells which cannot be explained by the observed C-strain growth curves. Showing distinct shifts at the zoom points, $p_i(t) = P(S_i(t) = 0)$ seems to depend highly on the zoom level of the microscope. This suggests, that besides full extinction of the S-strain, $S_i(t) = 0$ is also linked to limitations in area recognition. Additionally, the probability for $S_i(t) = 0$ is higher for positive MitC concentrations. Overall, the conditional mean for positive $S_i(t)$ but also the variability and probability for zero increase with the MitC concentration.

The smooth functional effects for each of the eight experimental batches are relatively small in size. For the conditional mean μ, they cause an average deviation of about 3% of the intercept growth curve (geometric mean over observed time points and batches); for the scale parameter ${}^{\sigma}/_{\mu}$, the average deviation is about 9%; and for p about 6%. While point-wise 95% bootstrap confidence interval, type uncertainty bounds (Online Appendix Section E.4) show less accuracy for the batch effects (in particular, those on $p(t)$), they indicate a high estimation precision for the MitC effects and functional intercepts. This corresponds to our findings in the simulation study in Section 4.

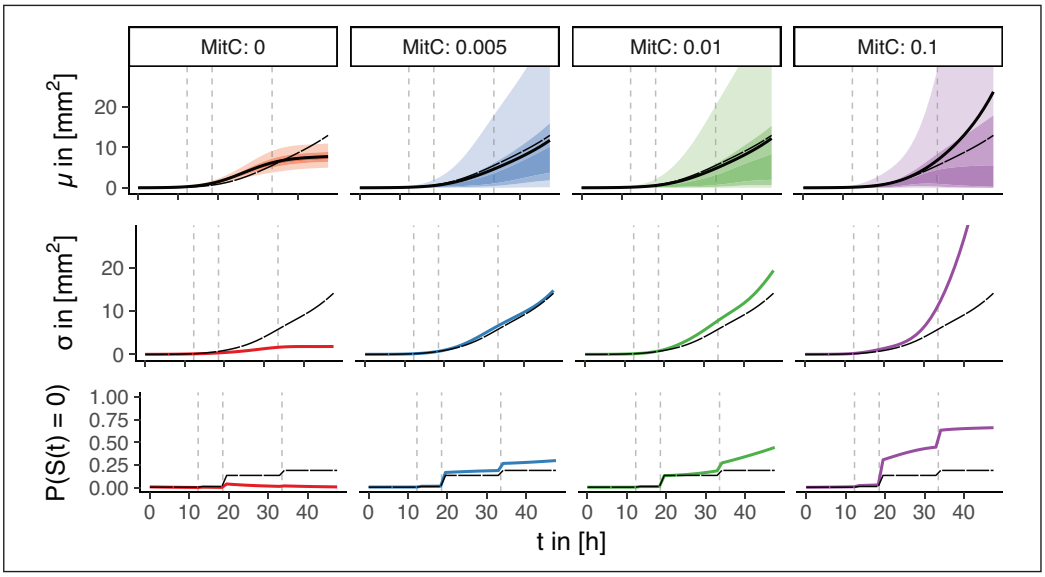

Figure 1 Estimated point-wise mean *(top)* and standard deviation *(centre)* of the S-strain growth curves $S_i(t)$ conditional on $S_i(t) > 0$ and the extinction probabilities *(bottom)* of S-strain growth curves for each MitC concentration. Long-dashed curves correspond to the functional intercept, dashed vertical lines to the zoom level change-points. Thick solid lines indicate the estimates, transparent ribbons reflect the point-wise inner 25%, 50% and 90% probability mass intervals of the estimated gamma distributions conditional on $S_i(t) > 0$ *(top)*

3.2.2 C-strain effect

The base-learner for the C-strain area propagation $C_i(s)$ effect on $\mu_i(t)$ is never selected throughout the boosting procedure and the effect on $\sigma_i(t)/\mu_i(t)$ is small (Online Appendix Figure 16). Thus, we only discuss the effect of the area increment $C_i'(s)$ here (Figure 2). Looking at the C'-μ-effect (effect of $C'(s)$ on mean S area), we can distinguish two main impact phases.

In the earlier growth phase with $s \leq 10\,h$, we observe a positive C'-μ-effect concerning almost the whole time curve of the S-strain. That means that C-strain growth above [below] the average indicates increased [decreased] S-strain propagation. Both colicin production and colicin secretion are costly to the population and slow down C-propagation. A low value of $C_i'(s)$ indicates early colicin secretion. We conclude that this first phase delineates a time window, where colicin emission is able to severely harm the S-strain population.

In the second phase for $s > 10\,h$, we observe a negative C'-μ-effect, which is maximal at short time lags and slowly fading. This likely reflects spatial competition of the S- and the C-strain (compare Online Appendix Figure 14). At this time, bacteria have grown together to coherent formations and strains obstruct expansion of each other. Even though the C'-μ-effect offers this clear interpretation, it is rather small compared to, for example, the MitC effect on $\mu(t)$. Moreover, while in simulation studies we observe a rather high estimation precision for most of the historical effects (Online

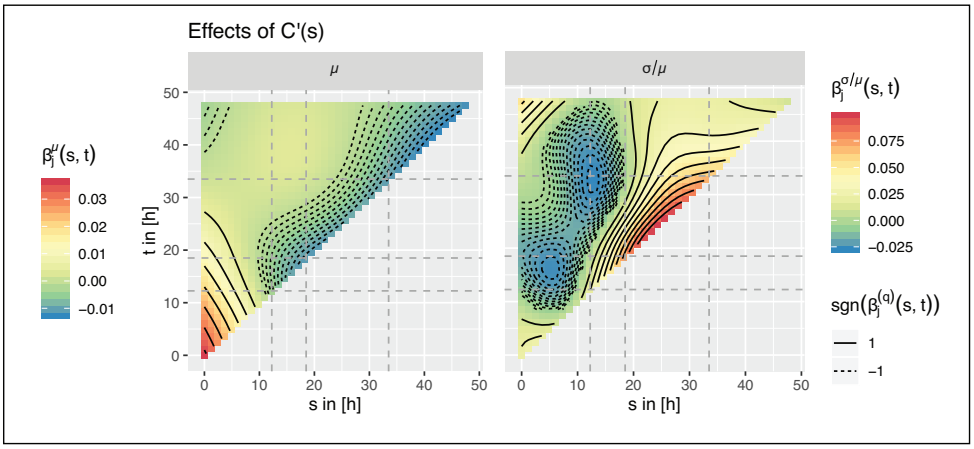

Figure 2 *Left:* Coefficient function $\beta^{(\mu)}(s, t)$ for the historical effects of $C_i'(s)$ on the mean of S-strain growth curves. *Right:* the corresponding plot for the effect of $C_i'(s)$ on the scale parameter ${}^{\sigma_i(t)}/_{\mu_i(t)}$. The *y*-axis represents the time line for the response curve, the *x*-axis represents the one for the C-strain growth curve. The change-points in zoom level are marked with dashed lines. For a fixed $s = s_0$, $\beta^{(\mu)}(s_0, t)$ and $\beta^{(\sigma/\mu)}(s_0, t)$ describe the effect of the normalized covariate at time s_0 on the S-strain growth curve over the whole remaining time interval.

Appendix Figure 12), 95% bootstrap confidence interval-type uncertainty bounds indicate distinctly less precision than for the MitC-effects (Online Appendix E.4). However, the historical C'-${}^{\sigma}/_{\mu}$-effect also corroborates the distinction into two phases of interaction: While in the first phase there is a negative effect of C-strain growth, the effect turns positive in the second phase. Thus, relative variability is increased for slow C-strain growth early in the experiment (colicin production) and for fast C-strain growth later in the experiment (areal competition).

For the probability $p_i(t) = P(S_i(t) = 0)$ both the effects of $C_i(s)$ and $C_i'(s)$ were selected. Corresponding plots can be found in the Online Appendix Figure 16. However, as already indicated by the marked cuts between the different zoom levels (Figure 1), vanishing of the S-strain is particularly sensitive to the precision of the area recognition. Hence, we are careful with interpreting the effects further in terms of the bacterial dynamics.

4 Simulation studies

4.1 Simulation set-up

Model-based gradient boosting approaches to non-functional or one-parameter special cases of the present model are well tested with respect to their fitting performance and variable selection quality (e.g., see Brockhaus et al., 2018a; Brockhaus et al., 2015; Thomas et al., 2018; Mayr et al., 2012) showing a typically slow over-fitting behaviour. However, modelling functional response variables with GAMLSS presents an important additional challenge: high auto-correlation in

response functions may lead to severe over-fitting when estimating typically complex base-learners. While this is already the case for non-GAMLSS functional response models, it gets particularly acute for GAMLSS models with multiple predictors—if it is not properly controlled for by early stopping based on curve-wise re-sampling methods. We focus on this issue in an extensive simulation study investigating the fitting performance for different levels of in-curve dependency while also comparing different sample sizes, choices of hyper parameters, the non-cyclic and cyclic fitting method, and different (curve-wise) re-sampling methods. Moreover, we consider three different models in the simulation study: one model is directly based on the bacterial interaction scenario in Chapter 3 taking the model estimated on the original data as true underlying model; and two models with a Gaussian response distribution and categorical effects or more complex smooth (interaction) effects of metric covariates, respectively. There, we randomly generate different sets of true underlying effects in order to obtain as general results as possible. In the Gaussian case, where this is possible, we also compare to the penalized likelihood approach of (Greven and Scheipl, 2017) which is implemented in the R package `refund` (Goldsmith et al., 2018). For details concerning the simulation set-up, the data generation and a more thorough discussion of the results, please refer to the corresponding sections in the Online Supplement.

4.2 Simulation results

Considering the mean $\overline{\text{KLD}}$ of the estimated to the true underlying model, we observe that for conditionally independent measurements within response curves, the optimal stopping iteration m_{stop} is typically far higher than for dependent or highly dependent measurements (Figure 3 *(left)*), that is, in the independent case a model can be fit distinctly longer without resulting in over-fitting. At the same time, we find that m_{stop} selected by curve-wise bootstrapping (performing slightly better than other curve-wise re-sampling methods) reflects these differences very well, which shows that it is desirably sensitive to in-curve dependency and prevents over-fitting. The resulting regularization improves the estimation accuracy strongly in particular for complex base-learners. The effect becomes especially visible when comparing it to the penalized likelihood (`refund`) approach (Figure 3 *(right)*), which currently lacks a corresponding regularization mechanism for GAMLSS models: When only modelling the response mean, there are typically curve-specific functional random intercepts included in order to account for in-curve dependency; however, they would interfere with modelling the marginal standard deviation in a separate predictor and are, thus, not included into the GAMLSS-type model. Measuring the fitting error in Root Mean Squared Error (RMSE), the `refund` approach shows a better performance in the independent case. However, it exceeds the RMSE of our `FDboost` approach by far in realistic scenarios with high in-curve dependency.

In the application motivated simulation scenario, we observe that most of the RMSEs for the estimated covariate effects are lower than 10% of the effect range even in the highly dependent setting (Online Appendix Figure 12). Exceptions are the functional intercept in the predictor for the extinction probability $p(t)$ being

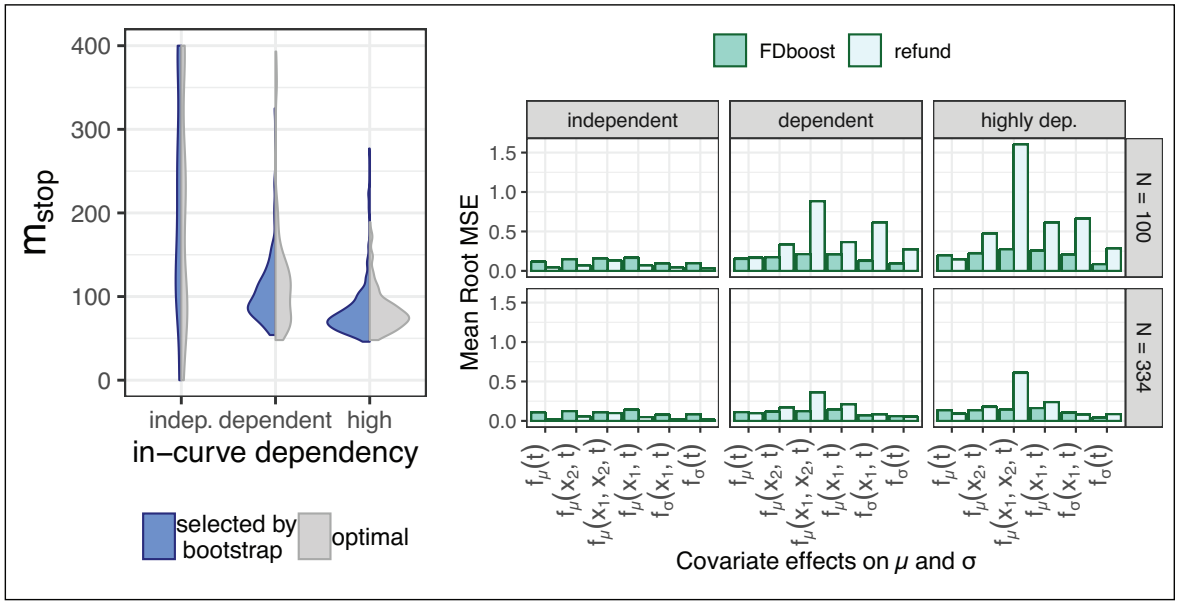

Figure 3 Plots referring to a Gaussian model scenario including smooth covariate effects $f_\vartheta(x_j, t)$ for $\vartheta \in \{\mu, \sigma\}$, the mean and standard deviation over time $t \in [0, 1]$, and for two metric covariates $j \in \{1, 2\}$, and a smooth interaction $f_\mu(x_1, x_2, t)$ effect for μ (200 model fits per combination of sample size N and in-curve dependency level). *Left:* Violin-plots reflecting the empirical density of the stopping iterations m_{stop} selected via 10-fold bootstrap (*left*) and for the $\overline{\text{KLD}}$-optimal m_{stop} (*right*) for $N = 334$ sampled curves. *Right:* Bar-plots indicating the mean RMSE of the different effects for our approach based on gradient boosting (`FDboost`, *dark*) and the approach based on penalized likelihood (`refund`, *light*). The highly dependent setting is the most realistic in many functional data scenarios and is—as far as the analogy can be drawn—the closest to the correlation structure in our application.

composed of a smooth functional intercept and a step function, and the σ/μ-effect of $C(t)$, which has a comparably large relative RMSE, due to its small effect size, while having a quite small absolute RMSE. Although we do not focus on variable selection in this article, the C-μ-effect and the smooth (non-step) functional intercept for $p(t)$, which were not selected in the original model fit in Chapter 3, serve as nuisance effects in the application motivated simulation. While the sensitivity is quite high for most of the non-zero effects (mostly 100%, minimum 70%), the nuisance C-μ-effect is still selected in rather many simulation runs (44% independent, 49% dependent, 33% highly dependent scenario), see Online Supplement Figure 13. To improve on this, stability selection as applied, for example, by Brockhaus et al. (2017) and Thomas et al. (2018) might be used. However, the mean RMSE of the effect is still extremely low indicating that even if the effect is selected it is very small in size. Overall, we observe the effects to be estimated quite well despite in-curve dependency and the high complexity of the model in both the Gaussian and the application motivated simulation studies.

5 Discussion and outlook

The functional GAMLSS regression framework we present in this article allows for very flexible modelling of functional responses. We may simultaneously model multiple parameters of functional response distributions in dependence of time and covariates, specifying a separate additive predictor for each parameter function. In addition, point-wise distributions for the response curves beyond exponential family distributions can be specified. Doing so a vast variety of new data scenarios can be modelled. These new possibilities have shown to be crucial, when applying the framework to analyse growth curves in the present bacterial interaction scenario.

The results we obtain confirm and extend previous work: Focusing on the outcome after 48 hour and on the number of C-clusters at the edge of the S-colony after 12 hour, von Bronk et al. (2017) already identify a phase of 'stochastic toxin dynamics' followed by a phase of 'deterministic dynamics' similar to the two phases of bacterial interaction we find in the historical functional effects of the C-strain growth. The functional regression model not only provides new evidence for this distinction from a completely new perspective, but now also allows to quantitatively discuss the effect of the C-strain on the S-strain over the whole time range: We now observe C-growth to have a positive effect on S-growth in the early phase and a negative effect in the later phase. The separation of these two phases appears even more distinct in the effect on the relative standard deviation, which we would not be able to recognize without GAMLSS.

Regarding the fraction of S- and C-strain area after 48 hour von Bronk et al. (2017) categorized three different states of the bacterial interaction: for no MitC, there is either dominance of the S-strain or coexistence; for a moderate MitC concentration, there occurs a splitting into two extremes—either dominance of the S-strain or extinction; and for the highest MitC concentration, the toxin strategy of the C-strain fails and the S-strain either dominates or both strains go extinct. Now, referring to the complete growth curves, our results also reflect this categorization: If MitC is added, and conditional on a positive area, the mean S-strain growth increases, whereas also the probability for zero area and the variance increase. However, these differences would not be captured by non-GAMLSS regression models for the mean only, as the mean growth curves not conditioning on the response being positive are very similar for no and moderate MitC concentration (see Online Appendix Figure 17). Apart from that, the framework provides the flexibility to account for special challenges in the experimental set-up, such as dependencies between observations in the same experimental batch and differences between zoom levels of the microscope.

In simulation studies, we confirm that by fitting our models via component-wise gradient boosting we are capable of estimating even complex covariate effects on multiple distributional parameter functions. As it prevents over-fitting, early stopping of the boosting algorithm based on curve-wise re-sampling plays a key role: It enables us to face settings with highly auto-correlated response curves without explicitly modelling the correlation structure.

Supplementary material

Supplementary material including an Online Appendix with further details and illustrations, as well as R code and data used for the analysis of bacterial interaction and simulations is available from `http://www.statmod.org/smij/archive.html`

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

Financial support from the Deutsche Forschungsgemeinschaft (DFG) through Emmy Noether grant GR 3793/1-1 (AS, SB, SG) and through grant OP252/4-2 part of the DFG Priority Program SPP1617 is gratefully acknowledged. B.v.B was supported by a DFG Fellowship through the Graduate School of Quantitative Biosciences Munich (QBM). Additional financial support by the Center for Nanoscience (CeNS) and the Nano Systems Initiative - Munich (NIM) is gratefully acknowledged.

References

Brockhaus S, Fuest A, Mayr A and Greven S (2018a) Signal regression models for location, scale and shape with an application to stock returns. *Journal of the Royal Statistical Society: Series C*, **67**, 665–86.

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**, 913–26.

Brockhaus S and Ruegamer D (2018) *FDboost: Boosting functional regression models*. R package version 0.3–1.

Brockhaus S, Rügamer D and Greven S (2018b) Boosting functional regression models with FDboost. *Journal of Statistical Software*. URL `https://arxiv.org/pdf/1705.10662.pdf` (last accessed 6 April 2020).

Brockhaus S, Scheipl F and Greven S (2015) The functional linear array model. *Statistical Modelling*, **15**, 279–300.

Bühlmann P and Hothorn T (2007) Boosting algorithms: Regularization, prediction and model fitting (with discussion). *Statistical Science*, **22**, 477–505.

Eilers PHC and Marx BD (2010) *Splines, knots, and penalties. Wiley Interdisciplinary Reviews: Computational Statistics*, **2**, 637–53.

Gasser T, Müller H-G, Kohler W, Molinari L and Prader A (1984) Nonparametric regression analysis of growth curves. *The Annals of Statistics*, **12**, 210–29.

Gertheiss J, Maier V, Hessel EF and Staicu A-M (2015) Marginal functional regression models for analysing the feeding behaviour of pigs. *Journal of Agricultural, Biological, and Environmental Statistics*, **20**, 353–70.

Goldsmith J, Scheipl F, Huang L, Wrobel J, Gellar J, Harezlak J, McLean MW, Swihart B, Xiao L, Crainiceanu C and Reiss PT (2018) *refund: Regression with Functional Data*. R

package version 0.1–17.

Goldsmith J, Zipunnikov V and Schrack J (2015) Generalized multilevel function-on-scalar regression and principal component analysis. *Biometrics*, **71**, 344–53.

Greven S and Scheipl F (2017) A general framework for functional regression modelling (with discussion). *Statistical Modelling*, **17**, 1–35.

Hall P, Mueller H-G and Yao F (2008) Modelling sparse generalized longitudinal observations with latent Gaussian processes. *Journal of the Royal Statistical Society: Series B*, **70**, 703–23.

Hastie T and Tibshirani R (1990) *Generalized Additive Models*. London: Chapman & Hall.

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**, 956–71.

Hofner B, Mayr A, Fenske N and Schmid M (2017) *gamboostLSS: Boosting methods for GAMLSS models*. R package version 2.0–0.

Kneib T, Hothorn T and Tutz G (2009) Variable selection and model choice in geoadditive regression models. *Biometrics*, **65**, 626–34.

Li H, Staudenmayer J and Carroll RJ (2014) Hierarchical functional data with mixed continuous and binary measurements. *Biometrics*, **70**, 802–11.

López S, Prieto M, Dijkstra J, Dhanoa M and France J (2004) Statistical evaluation of mathematical models for microbial growth. *International Journal of Food Microbiology*, **96**, 289–300.

Mayr A, Fenske N, Hofner B, Kneib T and Schmid M (2012) Generalized additive models for location, scale and shape for high dimensional data: A flexible approach based on boosting. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **61**, 403–27.

Morris JS (2015) Functional regression. *Annual Review of Statistics and Its Application*, **2**, 321–59.

Perni S, Andrew PW and Shama G (2005) Estimating the maximum growth rate from microbial growth curves: Definition is everything. *Food Microbiology*, **22**, 491–95.

R Core Team (2018) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. R version 3.5.1.

Ramsay J and Silverman BW (2005) *Functional Data Analysis*. New York, NY: Springer Science & Business Media.

Rigby RA and Stasinopoulos DM (2005) Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **54**, 507–54.

Scheipl F, Gertheiss J and Greven S (2016) Generalized functional additive mixed models. *Electronic Journal of Statistics*, **10**, 1455–92.

Scheipl F, Staicu A-M and Greven S (2015) Functional additive mixed models. *Journal of Computational and Graphical Statistics*, **24**, 477–501.

Staicu A-M, Crainiceanu CM, Reich DS and Ruppert D (2012) Modeling functional data with spatially heterogeneous shape characteristics. *Biometrics*, **68**, 331–43.

Stasinopoulos M and Rigby R (2019) *gamlss.dist: Distributions for generalized additive models for location scale and shape*. R package version 5.1-4.

Thomas J, Mayr A, Bischl B, Schmid M, Smith A and Hofner B (2018) Gradient boosting for distributional regression: Faster tuning and improved variable selection via noncyclical updates. *Statistics and Computing*, **28**, 673–87.

van der Linde A (2009) A Bayesian latent variable approach to functional principal components analysis with binary and count data. AStA *Advances in Statistical Analysis*, **93**, 307–33.

von Bronk B, Schaffer SA, Götz A and Opitz M (2017) Effects of stochasticity and division of labor in toxin production on two-strain bacterial competition in Escherichia coli. *PLoS Biology*, **15**, e2001457.

Wang B and Shi JQ (2014) Generalized Gaussian process regression model for non-Gaussian functional data. *Journal of the American Statistical Association*, **109**, 1123–33.

Weber MF, Poxleitner G, Hebisch E, Frey E and Opitz M (2014) Chemical warfare and survival strategies in bacterial range expansions (online supplement). *Journal of The Royal Society Interface*, **11**, 20140172.

Wood SN (2006) Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics*, **62**, 1025–36.

3. Additive Density-on-Scalar Regression in Bayes Hilbert Spaces with an Application to Gender Economics

In contrast to the extension of (point-wise) distributional assumptions in Chapter 2, distributions themselves can also be object of functional data analysis. Distribution densities naturally occur as data objects, for instance, when data privacy only allows sharing summarized data, when efficiency suggests summarizing massive data in densities, or when densities are of primary interest, say, because they are expected to be multi-modal. Here, we generalize FAMs and gradient boosting to model probability densities as response variables in Bayes Hilbert spaces. We use the approach to analyze gender-based income inequality using the German socioeconomic panel (SOEP) data.

Contributing article:
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 pre-print*. DOI: 10.48550/ARXIV.2110.11771.

Declaration on personal contributions:
This project arose from a seminar topic conducted by Eva Maier that the author of the thesis proposed, prototyped and supervised (jointly with Sonja Greven). While Eva Maier was substantially extending it (also by the application and several theoretical results) during her master's thesis and later her own doctoral studies, the author of this thesis was passively/supportingly involved thoughout the process taking a consulting role.

Additive Density-on-Scalar Regression in Bayes Hilbert Spaces with an Application to Gender Economics

Eva-Maria Maier[1], Almond Stöcker[1], Bernd Fitzenberger[2], and Sonja Greven[1]

[1] *Chair of Statistics, School of Business and Economics, Humboldt-Universität zu Berlin, Germany*
[2] *IAB (Institute for Employment Research), Nuremberg, Germany*

Abstract

Motivated by research on gender identity norms and the distribution of the woman's share in a couple's total labor income, we consider functional additive regression models for probability density functions as responses with scalar covariates. To preserve nonnegativity and integration to one under summation and scalar multiplication, we formulate the model for densities in a Bayes Hilbert space with respect to an arbitrary finite measure. This enables us to not only consider continuous densities, but also, e.g., discrete or mixed densities. Mixed densities occur in our application, as the woman's income share is a continuous variable having discrete point masses at zero and one for single-earner couples. We discuss interpretation of effect functions in our model via odds-ratios. Estimation is based on a gradient boosting algorithm, allowing for potentially numerous flexible covariate effects. We show how to handle the challenging estimation for mixed densities within our framework using an orthogonal decomposition. Applying this approach to data from the German Socio-Economic Panel Study (SOEP) shows a more symmetric distribution in East German than in West German couples after reunification and a smaller child penalty comparing couples with and without minor children. These West-East differences become smaller, but are persistent over time.

Keywords: Density Regression; Functional Additive Model; Gradient Boosting; Mixed Densities.

1 Introduction

Analyzing the distribution of the female income share for couples in the U.S., Bertrand et al. (2015) show that the fraction of couples with a share below 0.5 is much higher than the fraction of those with a share above 0.5 and that there is

a discontinuous drop in the density at 0.5. This drop is attributed to gender identity norms with men being averse to a situation with their female partners making more money than themselves. Subsequent studies, however, showed mixed results (e.g., Sprengholz et al., 2020; Kuehnle et al., 2021). Most of the literature does not consider how the share distribution changes depending on covariates, but this in itself is of great interest. Social norms change over time towards higher employment of females, with part-time employment becoming more prevalent, especially in the presence of children. And the employment and earnings of female partners show a strong childhood penalty (Kleven et al., 2019; Fitzenberger et al., 2013).
From a methodological perspective, the focus on a univariate analysis of the share distribution reflects the lack of an interpretable multivariate analysis of its determinants. Filling this gap, we introduce a regression approach for outcomes that are probability density functions with scalar covariates and we use this new approach to analyze how the female income share distribution in Germany varies by place of residence (e.g., between West and East Germany), the presence of children, and over time.
For the continuous density case, our approach could be viewed as a special case of functional regression, which is part of the vast field of functional data analysis (e.g., Ramsay and Silverman, 2005). One usually distinguishes three types of functional regression models (e.g., Brockhaus et al., 2015): *scalar-on-function*, where the response is scalar while the covariates are functions, *function-on-scalar* with functional response and scalar covariates and *function-on-function*, where both, response and covariates, are functions. Analogously, we refer to our regression setting as *density-on-scalar*. Existing function-on-scalar methods are not applicable in this case, as multiplying a density with a negative scalar or adding two densities in the classical sense immediately violates the nonnegativity and integrate-to-one constraints of densities. An appropriate alternative normed vector space structure for densities is provided by Bayes Hilbert spaces, motivated by Aitchison's work about compositional data (Aitchison, 1986). Egozcue et al. (2006) first introduced Bayes Hilbert spaces for densities with respect to the Lebesgue measure on a finite interval. This was extended by Boogaart et al. (2014) to Bayes Hilbert spaces on finite measure spaces. Talská et al. (2018) use Bayes Hilbert spaces for linear density-on-scalar regression, considering only densities defined on a finite interval and Lebesgue integrals. For estimation, the model is mapped into a subspace of the L^2 of square integrable functions applying the centered log-ratio (clr) transformation. We extend their framework to additive density-on-scalar regression models for densities on arbitrary finite measure spaces. This enables us to handle not only densities with respect to the Lebesgue measure on a finite interval (*continuous case*) or to the weighted sum of Dirac measures on a finite set (*discrete case*) but also mixtures of both (*mixed case*) in a unified framework. We introduce a gradient boosting algorithm based on the approach of Hothorn et al. (2014), enabling estimation directly in the Bayes Hilbert space. Furthermore, we develop a method to interpret the estimated effects analogously to odds ratios. In our motivating application, we analyze the distribution of the woman's share in a couple's total labor income in Germany – an example of the mixed case: The corresponding densities defined on $[0, 1]$ have positive point mass at the boundary values 0 and 1, corresponding to single-earner

couples. This leads to a mixed (Dirac/Lebesgue) reference measure.
Apart from the Bayes Hilbert space approach, different ideas for density regression have been proposed. Park and Qian (2012) discuss density-on-density regression without any positivity constraints, performing linear regression directly on the deviations from the mean density. Petersen and Müller (2019) present linear regression for densities based on the Wasserstein metric, which constitutes a popular approach to statistical analysis of distributional data (Ollivier et al., 2014). However, with this approach the densities are considered in a nonlinear space, which makes modeling and interpretation more difficult. Han et al. (2020) introduce additive functional regression models for the density-on-scalar case as well. They transform the densities to the L^2, in particular proposing the log hazard and the log quantile density transformations. Happ et al. (2019) show that for both, numerical instabilities may occur and finally prefer the clr transformation to both. In contrast to Han et al. (2020), which only considers the transformed densities for modeling, our Bayes Hilbert space approach provides an entire conceptual framework that allows to embed the densities and specify the model in a vector space structure. The clr transformation, an isometric isomorphism, allows an equivalent formulation in the L^2, which enables appealing odds-ratio-type interpretations on the original density-level.
Density regression is related to several other areas of research. For the discrete case, also known as compositional data (i.e., a multivariate vector of non-negative fractions summing to one, e.g., Pawlowsky-Glahn et al., 2015), regression has also been studied, with Boogaart et al. (2015) for instance considering Bayesian regression with compositional response. In general, there are also approaches not modeling densities but equivalent functions. E.g., Yang et al. (2018) present a Bayesian approach to model quantile functions as response in a functional linear regression by introducing their quantlet basis representation. In contrast, modeling the density function has the distinct advantage that shifts of probability masses and special characteristics of the distribution such as bimodality can be identified straightforwardly.
All methods mentioned so far share the assumption that a sample of densities (or, e.g., quantile functions) has been observed (or estimated). In contrast, there are also individual-level approaches, which model the conditional density or equivalent functions given covariates based on a sample of individual scalar data. Parametric approaches such as generalized additive models for location, scale and shape (GAMLSS, also known as distributional regression; e.g., Rigby and Stasinopoulos, 2005) require a known distribution family and only enable interpretation on the level of their parameters, not the distribution, which can be restrictive. In quantile regression (e.g., Koenker, 2005) no specific distribution family is assumed, but for each quantile of interest one model has to be estimated, which is potentially computationally demanding. Furthermore, the estimated quantiles may cross, which can be avoided, e.g., by monotonization or rearrangement (e.g., Chernozhukov et al., 2010). Conditional transformation models (CTMs, e.g., Hothorn et al., 2014) model a monotone transformation function, which transforms the conditional distribution function (cdf) of the response to an a priori specified reference distribution function, in terms of covariates. In distribution regression (e.g., Chernozhukov et al., 2013), the cdf is estimated pointwise, similarly as in quantile regression. It requires the choice of a link function between the conditional distribution and the parametric

covariate effects. Moreover, there are Bayesian (e.g., MacEachern, 1999), kernel estimation (e.g., Takeuchi et al., 2006), and machine learning approaches (e.g., Li et al., 2021) for modeling conditional response densities, suffering from two limitations: They work with a relatively large number of hyper-parameter(-distribution)s which influence the outcome; related to this is their lack of interpretability, in particular in terms of the covariate effects.
We aim to bridge this gap in the literature by directly modeling the response density on the one hand, while borrowing interpretable yet flexible additive models from functional data analysis on the other hand. To the best of the authors' knowledge, our approach is the first to cover continuous, discrete and mixed cases in a unified framework.
In the following, Section 2 summarizes the construction of Bayes Hilbert spaces. Section 3 introduces our density-on-scalar regression approach, where models are formulated in Bayes Hilbert spaces and estimated using a boosting algorithm. In the mixed case, we derive an orthogonal decomposition of the Bayes Hilbert space to facilitate (separate continuous/discrete) estimation. We develop an interpretation method of the estimated effects using odds-ratios. Section 4 involves a comprehensive application for the mixed case in analyzing the distribution of the woman's share in a couple's total labor income in Germany. Section 5 provides a small simulation study based on our application setting to validate our approach. We conclude with a discussion and an outlook in Section 6.

2 The Bayes Hilbert space

We briefly introduce Bayes spaces and summarize their basic vector space properties for a σ-finite reference measure as described in Boogaart et al. (2010). Refining these to Bayes Hilbert spaces (Boogaart et al., 2014), we have to restrict ourselves to finite reference measures. We provide proofs for all theorems in appendix A.1 since we take a slightly different point of view compared to Boogaart et al. (2010) and Boogaart et al. (2014).
Let $(\mathcal{T}, \mathcal{A})$ be a measurable space and μ a σ-finite measure on it, the so-called *reference measure*. Consider the set $\mathcal{M}(\mathcal{T}, \mathcal{A}, \mu)$, or short $\mathcal{M}(\mu)$, of σ-finite measures with the same null sets as μ. Such measures are mutually absolutely continuous to each other, i.e., by Radon-Nikodyms' theorem, the μ-density of ν or Radon-Nikodym derivative of ν with respect to μ, $f_\nu := \mathrm{d}\nu/\mathrm{d}\mu : \mathcal{T} \to \mathbb{R}$, exists for every $\nu \in \mathcal{M}(\mu)$. It is μ-almost everywhere (μ-a.e.) positive and unique. We write $f_\nu \cong \nu$ for a measure $\nu \in \mathcal{M}(\mu)$ and its corresponding μ-density f_ν. For measures $\nu_1, \nu_2 \in \mathcal{M}(\mu)$, let the equivalence relation $=_\mathcal{B}$ be given by $\nu_1 =_\mathcal{B} \nu_2$, iff there is a $c > 0$ such that $\nu_1(A) = c\,\nu_2(A)$ for every $A \in \mathcal{A}$, where $c\,(+\infty) = +\infty$. Respectively, we define $f_{\nu_1} =_\mathcal{B} f_{\nu_2}$, iff $f_{\nu_1} = c\,f_{\nu_2}$ for some $c > 0$. Here and in the following, pointwise identities have to be understood μ-a.e. Both definitions of $=_\mathcal{B}$ are compatible with the Radon-Nikodym identification $f_\nu \cong \nu$. The set of $(=_\mathcal{B})$-equivalence classes is called the *Bayes space (with reference measure μ)*, denoted by $\mathcal{B}(\mu) = \mathcal{B}(\mathcal{T}, \mathcal{A}, \mu)$. For equivalence classes containing finite measures, we choose the respective probability measure as representative in practice. Then, the corresponding μ-density is a probability density. However, mathematically it is

more convenient to use a non-normalized representative. For better readability, we omit the index $\mathcal{B}$ in $=_{\mathcal{B}}$ and the square brackets denoting equivalence classes in the following. For $f_{\nu_1} \cong \nu_1, f_{\nu_2} \cong \nu_2 \in \mathcal{B}(\mu)$, the addition or *perturbation* is given by the equivalent definitions

$$(\nu_1 \oplus \nu_2)(A) := \int_A \frac{\mathrm{d}\nu_1}{\mathrm{d}\mu} \frac{\mathrm{d}\nu_2}{\mathrm{d}\mu} \,\mathrm{d}\mu, \qquad\qquad f_{\nu_1} \oplus f_{\nu_2} := f_{\nu_1} f_{\nu_2}.$$

For $f_\nu \cong \nu \in \mathcal{B}(\mu)$ and $\alpha \in \mathbb{R}$, the scalar multiplication or *powering* is defined by

$$(\alpha \odot \nu)(A) := \int_A \left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right)^\alpha \mathrm{d}\mu, \qquad\qquad \alpha \odot f_\nu := (f_\nu)^\alpha.$$

Theorem 2.1 (Boogaart et al., 2010)**.** *The Bayes space $\mathcal{B}(\mu)$ with perturbation $\oplus$ and powering $\odot$ is a real vector space with additive neutral element $0 := \mu \cong 1$, additive inverse element $\ominus\nu := \int_A \mathrm{d}\mu/\mathrm{d}\nu \,\mathrm{d}\mu \cong 1/f_\nu$ for $\nu \in \mathcal{B}(\mu)$, and multiplicative neutral element $1 \in \mathbb{R}$.*

For subtraction, we write $\nu_1 \ominus \nu_2 := \nu_1 \oplus (\ominus\nu_2)$ and $f_{\nu_1} \ominus f_{\nu_2} := f_{\nu_1} \oplus (\ominus f_{\nu_2})$.
From now on, we restrict the reference measure μ to be finite, progressing to Bayes Hilbert spaces. This is similar to Boogaart et al. (2014) with some details different. In the style of the well-known L^p spaces, B^p spaces for $1 \leq p < \infty$ are defined as

$$B^p(\mu) = B^p(\mathcal{T}, \mathcal{A}, \mu) := \left\{ \nu \in \mathcal{B}(\mu) \;\middle|\; \int_{\mathcal{T}} \left|\log \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right|^p \mathrm{d}\mu < \infty \right\}.$$

We also say $f_\nu \in B^p(\mu)$ for $f_\nu \cong \nu \in B^p(\mu)$. This is equivalent to $\log f_\nu \in L^p(\mu)$, which gives us $B^q(\mu) \subset B^p(\mu)$ for $p, q \in \mathbb{R}$ with $1 \leq p < q$. Note that for every $p \in \mathbb{R}$ with $1 \leq p < \infty$, the space $B^p(\mu)$ is a vector subspace of $\mathcal{B}(\mu)$, see Boogaart et al. (2014). For $f_\nu \cong \nu \in B^p(\mu)$, the *centered log-ratio (clr) transformation of ν* is given by

$$\mathrm{clr}_{B^p(\mathcal{T},\mathcal{A},\mu)}[\nu] = \mathrm{clr}_{B^p(\mathcal{T},\mathcal{A},\mu)}[f_\nu] := \log f_\nu - \mathcal{S}_{B^p(\mathcal{T},\mathcal{A},\mu)}(f_\nu), \tag{2.1}$$

with $\mathcal{S}_{B^p(\mathcal{T},\mathcal{A},\mu)}(f_\nu) := 1/\mu(\mathcal{T}) \int_{\mathcal{T}} \log f_\nu \,\mathrm{d}\mu$ the mean logarithmic integral. We omit the indices $B^p(\mathcal{T}, \mathcal{A}, \mu)$ or shorten them to μ or $\mathcal{T}$, if the underlying space is clear from context.

Proposition 2.2 (For $p = 1$ shown in Boogaart et al., 2014)**.** *For $1 \leq p < \infty$, the clr transformation* $\mathrm{clr} : B^p(\mu) \to L^p_0(\mu) := \{\tilde{f} \in L^p(\mu) \mid \int_{\mathcal{T}} \tilde{f} \,\mathrm{d}\mu = 0\}$ *is an isomorphism with inverse transformation* $\mathrm{clr}^{-1}[\tilde{f}] = \exp \tilde{f}$.

Note that $L^p_0(\mu)$ is a closed subspace of $L^p(\mu)$. The space $B^2(\mu)$ is called the *Bayes Hilbert space (with reference measure μ)*. For $f_{\nu_1} \cong \nu_1, f_{\nu_2} \cong \nu_2 \in B^2(\mu)$, consider

$$\langle \nu_1, \nu_2 \rangle_{B^2(\mu)} := \langle f_{\nu_1}, f_{\nu_2} \rangle_{B^2(\mu)} := \int_{\mathcal{T}} \mathrm{clr}[f_{\nu_1}] \;\mathrm{clr}[f_{\nu_2}] \,\mathrm{d}\mu,$$

which is an inner product on $B^2(\mu)$, see Proposition A.1 in appendix A.1. It induces a norm on $B^2(\mu)$ by $\|\nu\|_{B^2(\mu)} := \|f_\nu\|_{B^2(\mu)} := \sqrt{\langle f_\nu, f_\nu \rangle_{B^2(\mu)}}$ for $f_\nu \cong \nu \in B^2(\mu)$. By definition, we have $\langle f_{\nu_1}, f_{\nu_2} \rangle_{B^2(\mu)} = \langle \mathrm{clr}[f_{\nu_1}], \mathrm{clr}[f_{\nu_2}] \rangle_{L^2(\mu)}$, which immediately implies that $\mathrm{clr} : B^2(\mu) \to L^2_0(\mu)$ is isometric. We now formulate the main statement of this section:

Theorem 2.3 (Boogaart et al., 2014)**.** *The Bayes Hilbert space $B^2(\mu)$ is a Hilbert space.*

Note that in Proposition A.2 in appendix A.1, we introduce a notion of canonical embedding, which enables us to identify the Bayes Hilbert space $B^2(\mathcal{T}_0, \mathcal{A} \cap \mathcal{T}_0, \mu)$ with a closed subspace of $B^2(\mathcal{T}, \mathcal{A}, \mu)$ for any $\mathcal{T}_0 \in \mathcal{A}$. Furthermore, we explicitly compute the orthogonal projection onto $B^2(\mathcal{T}_0, \mathcal{A} \cap \mathcal{T}_0, \mu)$. This construction is new to the best of the authors' knowledge. An important consequence of the properties of the orthogonal projection is that we may restrict linear problems (like regression models) onto subsets of $\mathcal{T}$ consistently with the geometry of the Bayes Hilbert spaces. For compositional data, the correspondence of subcompositions in $\mathcal{T}_0 \subset \mathcal{T}$ to subspaces of the Bayes Hilbert space is referred to as *subcompositional coherence* (Pawlowsky-Glahn et al., 2015).

3 Density-on-scalar regression

We consider regression models with a density as response and scalar covariates. More precisely, the response has to be an element of a Bayes Hilbert space $B^2(\mu) = B^2(\mathcal{T}, \mathcal{A}, \mu)$. This requires μ to be finite on $(\mathcal{T}, \mathcal{A})$, excluding, e.g., densities on the whole real line using the Lebesgue measure as reference or densities which are exactly zero in parts of $\mathcal{T}$. To consider $\mathcal{T} = \mathbb{R}$ with the Borel σ-algebra $\mathfrak{B}_{\mathbb{R}}$, a possible reference is the probability measure corresponding to the standard normal distribution (Boogaart et al., 2014). If a density is not directly observed but estimated from an observed sample, density values of zero can be avoided by choosing a density estimation method that yields a positive density. For discrete sets $\mathcal{T}$, one option is to replace observed density values of zero with small values (e.g., Pawlowsky-Glahn et al., 2015). The framework allows for a variety of different applications. Usually, we consider $\mathcal{T} \subset \mathbb{R}$ with three common cases: In the *continuous case*, we consider a nontrivial interval $\mathcal{T} = I$ with $\mathcal{A} = \mathfrak{B}$ the Borel σ-algebra restricted to I and $\mu = \lambda$ the Lebesgue measure. The *discrete case* refers to a discrete set $\mathcal{T} = \{t_1, \dots, t_D\}$ with $\mathcal{A} = \mathcal{P}(\mathcal{T})$ the power set of $\mathcal{T}$ and $\mu = \sum_{d=1}^{D} w_d \, \delta_{t_d}$ a weighted sum of Dirac measures, where $w_d > 0$. The *mixed case* is a mixture of both: As in the continuous case, we have $\mathcal{T} = I$ and $\mathcal{A} = \mathfrak{B}$, but some points $\mathcal{D} = \{t_1, \dots, t_D\} \subset I$ have positive probability mass. The corresponding reference measure is a mixture of weighted Dirac measures and the Lebesgue measure, i.e., $\mu = \sum_{d=1}^{D} w_d \, \delta_{t_d} + \lambda$. Note that the special case $\mathcal{D} = \emptyset$ yields the continuous case. Our application in Section 4 gives an example for the mixed case.

3.1 Regression model

Density-on-scalar regression is motivated by function-on-scalar regression. Both regression types are closely related (at least in the continuous case), as density-on-scalar models can be transformed to function-on-scalar models in $L_0^2(\mu)$ via the clr transformation. We formulate our model analogously to structured additive function-on-scalar regression models (Brockhaus et al., 2015), considering densities in a Bayes Hilbert space $B^2(\mu)$ instead of functions in $L^2(I, \mathfrak{B}, \lambda)$ and using the

corresponding operations. For data pairs $(y_i, \mathbf{x}_i) \in B^2(\mu) \times \mathbb{R}^K$, $K \in \mathbb{N}$, $i = 1, \ldots, N$, $N \in \mathbb{N}$, this yields the structured additive density-on-scalar regression model

$$y_i = h(\mathbf{x}_i) \oplus \varepsilon_i = \bigoplus_{j=1}^{J} h_j(\mathbf{x}_i) \oplus \varepsilon_i, \tag{3.1}$$

where $\varepsilon_i \in B^2(\mu)$ are functional error terms with $\mathbb{E}(\varepsilon_i) = 0 \in B^2(\mu)$ and $h_j(\mathbf{x}_i) \in B^2(\mu)$ are partial effects, $J \in \mathbb{N}$. The expectations of the $B^2(\mu)$-valued random elements ε_i are defined using the Bochner integral (e.g., Hsing and Eubank, 2015). Each partial effect $h_j(\mathbf{x}_i) \in B^2(\mu)$ in (3.1) models an effect of none, one or more covariates in $\mathbf{x}_i$.

Covariate(s)	Type of effect	$h_j(\mathbf{x})$
None	Intercept	β_0
One scalar covariate x	Linear effect	$x \odot \beta$
	Flexible effect	$g(x)$
Two scalar covariates x_1, x_2	Linear interaction	$x_1 \odot (x_2 \odot \beta)$
	Functional varying coefficient	$x_1 \odot g(x_2)$
	Flexible interaction	$g(x_1, x_2)$
Grouping variable k	Group-specific intercepts	β_k
Grouping variable k and scalar x	Group-specific linear effects	$x \odot \beta_k$
	Group-specific flexible effects	$g_k(x)$

Table 3.1: Partial effects for density-on-scalar regression.

Table 3.1 gives an overview of possible partial effects, inspired by Table 1 in Brockhaus et al. (2015). The upper part shows effects for up to two different scalar covariates. In the lower part, group-specific effects for categorical variables are presented. Interactions of the given effects are possible as well. Scalar covariates are denoted by x, densities in $B^2(\mu)$ by β and $g(\,)$. Note that constraints are necessary to obtain identifiable models. For a model with an intercept β_0, this is obtained by centering the partial effects:

$$\frac{1}{N} \odot \bigoplus_{i=1}^{N} h_j(\mathbf{x}_i) = 0. \tag{3.2}$$

More details about how to include this constraint in a functional linear array model for function-on-scalar regression can be found in appendix A of Brockhaus et al. (2015). A similar procedure can be used to obtain a centering of interaction effects around the main effects, see appendix A of Stöcker et al. (2021). Both approaches are based on Wood (2017, Section 1.8.1) and can be transferred straightforwardly to density-on-scalar regression.

3.2 Estimation by Gradient Boosting

To estimate the function $h(\mathbf{x}_i) \in B^2(\mu)$ in Equation (3.1), the sum of squared errors

$$\text{SSE}(h) := \sum_{i=1}^{N} \|\varepsilon_i\|^2_{B^2(\mu)} = \sum_{i=1}^{N} \|y_i \ominus h(\mathbf{x}_i)\|^2_{B^2(\mu)} = \sum_{i=1}^{N} \rho_{y_i}\left(h(\mathbf{x}_i)\right) \tag{3.3}$$

is minimized. Here, $\rho_{y_i} : B^2(\mu) \to \mathbb{R}$, $f_\nu \mapsto \|y_i \ominus f_\nu\|^2_{B^2(\mu)}$ is the quadratic loss functional. To simplify the minimization problem and to determine the type of an effect, compare Table 3.1, we consider a basis representation for each partial effect:

$$h_j(\mathbf{x}_i) = \left(\mathbf{b}_j(\mathbf{x}_i)^\top \circledcirc \mathbf{b}_Y^\top\right) \boldsymbol{\theta}_j = \bigoplus_{n=1}^{K_j} \bigoplus_{m=1}^{K_Y} b_{j,n}(\mathbf{x}_i) \odot b_{Y,m} \odot \theta_{j,n,m}, \tag{3.4}$$

where $\mathbf{b}_j = (b_{j,1}, \ldots, b_{j,K_j}) : \mathbb{R}^K \to \mathbb{R}^{K_j}$ is a vector of basis functions in direction of the covariates and $\mathbf{b}_Y = (b_{Y,1}, \ldots, b_{Y,K_Y}) \in B^2(\mu)^{K_Y}$ is a vector of basis functions over $\mathcal{T}$. With $\circledcirc$, we denote the Kronecker product of a real-valued with a $B^2(\mu)$-valued matrix. It is defined like the Kronecker product of two real-valued matrices, using $\odot$ instead of the usual multiplication. Similarly, matrix multiplication of a real-valued with a $B^2(\mu)$-valued matrix is defined by replacing sums with $\oplus$ and products with $\odot$ in the usual matrix multiplication. Our goal is to estimate the coefficient vector $\boldsymbol{\theta}_j = (\theta_{j,1,1}, \ldots, \theta_{j,K_j,K_Y}) \in \mathbb{R}^{K_j K_Y}$. To allow sufficient flexibility for h_j, the product $K_j K_Y$ can be chosen to be large. The necessary regularization can then be accomplished with a Ridge-type penalty term $\boldsymbol{\theta}_j^\top \mathbf{P}_{j,Y} \boldsymbol{\theta}_j$. For a basis representation as in equation (3.4), an anisotropic penalty matrix $\mathbf{P}_{j,Y} = \lambda_j(\mathbf{P}_j \otimes \mathbf{I}_{K_Y}) + \lambda_Y(\mathbf{I}_{K_j} \otimes \mathbf{P}_Y)$ can be used. Here, $\mathbf{P}_j \in \mathbb{R}^{K_j \times K_j}$ and $\mathbf{P}_Y \in \mathbb{R}^{K_Y \times K_Y}$ are suitable penalty matrices for $\mathbf{b}_j$ and $\mathbf{b}_Y$, respectively, and $\lambda_j, \lambda_Y \geq 0$ are smoothing parameters in the respective directions. Alternatively, a simplified isotropic penalty matrix $\mathbf{P}_{j,Y} = \lambda_j((\mathbf{P}_j \otimes \mathbf{I}_{K_Y}) + (\mathbf{I}_{K_j} \otimes \mathbf{P}_Y))$ with only one smoothing parameter is possible (Brockhaus et al., 2020). The basis representation framework might seem restrictive at first, but it indeed allows for very flexible modeling of the effects, as discussed below.

We fit model (3.1) using a component-wise gradient boosting algorithm, where the expected loss is minimized step-wise along the steepest gradient descent. It is an adaption of the algorithm presented in Brockhaus et al. (2015), which was modified from Hothorn et al. (2014). Advantages of this approach are that it can deal with a large number of covariates, it performs variable selection, and includes regularization. Bühlmann and Yu (2003) discuss theoretical properties of gradient boosting w.r.t. sum of squares errors, which is typically referred to as L_2-Boosting, for scalar responses. They show – simplifying to a single learner – that bias decays exponentially fast while estimator variance increases in exponentially small steps over the boosting iterations, which supports the general practice of stopping the algorithm early before it eventually reaches the standard (penalized) least squares estimate. Lutz and Bühlmann (2006) show consistency of component-wise L_2-Boosting for linear regression with both high-dimensional multivariate response and predictors. Similar to these predecessors, our L_2-Boosting algorithm for Bayes Hilbert spaces simplifies to repeated re-fitting of residuals – which, however, present densities in our case.

Algorithm: Bayes space L_2-Boosting for density-on-scalar models

1. Select vectors of basis functions $\mathbf{b}_Y, \mathbf{b}_j$, the starting coefficient vector $\boldsymbol{\theta}_j^{[0]} \in \mathbb{R}^{K_j K_Y}$, and penalty matrices $\mathbf{P}_{j,Y}, j = 1, \ldots, J$. Choose the step-length $\kappa \in (0, 1)$ and the stopping iteration m_{stop} and set the iteration number m to zero. We comment on a suitable selection of these quantities below.

2. Calculate the negative gradient of the empirical risk with respect to the Fréchet differential (see appendix A.2 for the proof of this equation)

$$U_i := \ominus \nabla \rho_{y_i}(f_\nu)\Big|_{f_\nu = \hat{h}^{[m]}(\mathbf{x}_i)} = 2 \odot \left(y_i \ominus \hat{h}^{[m]}(\mathbf{x}_i)\right), \tag{3.5}$$

   where $\hat{h}^{[m]}(\mathbf{x}_i) = \bigoplus_{j=1}^{J} \left(\mathbf{b}_j(\mathbf{x}_i)^\top \otimes \mathbf{b}_Y^\top\right) \boldsymbol{\theta}_j^{[m]}$. Fit the base-learners

$$\hat{\boldsymbol{\gamma}}_j = \underset{\boldsymbol{\gamma} \in \mathbb{R}^{K_j K_Y}}{\operatorname{argmin}} \sum_{i=1}^{N} \left\| U_i \ominus \left(\mathbf{b}_j(\mathbf{x}_i)^\top \otimes \mathbf{b}_Y^\top\right) \boldsymbol{\gamma} \right\|_{B^2(\mu)}^2 + \boldsymbol{\gamma}^\top \mathbf{P}_{jY} \boldsymbol{\gamma} \tag{3.6}$$

   for $j = 1, \ldots, J$ and select the best base-learner

$$j^* = \underset{j=1,\ldots,J}{\operatorname{argmin}} \sum_{i=1}^{N} \left\| U_i \ominus \left(\mathbf{b}_j(\mathbf{x}_i)^\top \otimes \mathbf{b}_Y^\top\right) \hat{\boldsymbol{\gamma}}_j \right\|_{B^2(\mu)}^2. \tag{3.7}$$

3. The coefficient vector corresponding to the best base-learner is updated, the others stay the same: $\boldsymbol{\theta}_{j^*}^{[m+1]} := \boldsymbol{\theta}_{j^*}^{[m]} + \kappa\, \hat{\boldsymbol{\gamma}}_{j^*}$, $\boldsymbol{\theta}_j^{[m+1]} := \boldsymbol{\theta}_j^{[m]}$ for $j \neq j^*$.

4. While $m < m_{\text{stop}}$, increase m by one and go back to step 2. Stop otherwise.

The resulting estimator of model (3.1) is $\hat{y}_i = \hat{\mathbb{E}}(y_i \mid \mathbf{X} = \mathbf{x}_i) = \bigoplus_{j=1}^{J} \hat{h}_j^{[m_{\text{stop}}]}(\mathbf{x}_i)$, with $\hat{h}_j^{[m_{\text{stop}}]}(\mathbf{x}_i) = (\mathbf{b}_j(\mathbf{x}_i)^\top \otimes \mathbf{b}_Y^\top)\boldsymbol{\theta}_j^{[m_{\text{stop}}]}$. In the following, we discuss the selection of parameters in step 1, see also Brockhaus et al. (2015) and Brockhaus et al. (2020). The choice of vectors of basis functions $\mathbf{b}_Y$ and $\mathbf{b}_j$ and their corresponding penalty matrices $\mathbf{P}_j$ and $\mathbf{P}_Y$ depends on the desired partial effect $h_j(\mathbf{x})$. Regarding the basis functions $\mathbf{b}_j$ in direction of the covariates, suitable selections for flexible effects are B-splines with a difference penalty. For a linear effect of one covariate, the vector of basis functions is chosen as $\mathbf{b}_j = (1, \text{id}) : \mathbb{R} \to \mathbb{R}^2,\ x \mapsto (1, x)$, resulting in the design matrix of a simple linear model. Here, a reasonable penalty matrix is $\mathbf{P}_j = \mathbf{I}_2$ corresponding to the Ridge penalty. A basis $\mathbf{b}_Y \in B^2(\mu)^{K_Y}$ can be obtained from a suitable basis $\bar{\mathbf{b}}_Y \in L^2(\mu)^{K_Y+1}$ as follows. Transforming $\bar{\mathbf{b}}_Y$ to $L_0^2(\mu)^{K_Y}$ yields a basis $\tilde{\mathbf{b}}_Y \in L_0^2(\mu)^{K_Y}$. The respective transformation matrix is constructed in appendix B. Applying the inverse clr transformation on each component of $\tilde{\mathbf{b}}_Y$ gives the desired basis $\mathbf{b}_Y$. For the continuous case, a reasonable choice for $\bar{\mathbf{b}}_Y \in L^2(\lambda)^{K_Y+1}$ is a B-spline basis with a difference penalty, allowing for flexible modeling of the response densities. For the discrete case, a suitable selection is $\bar{\mathbf{b}}_Y = (\mathbb{1}_{\{t_1\}}, \ldots, \mathbb{1}_{\{t_D\}}) \in L^2(\sum_{d=1}^{D} w_d\, \delta_{t_d})^D$, where $\mathbb{1}_A$ denotes the indicator function of $A \in \mathcal{A}$. Again, a difference penalty can be used to control the volatility of the estimates. The mixed case is not as straightforward. We show in Section 3.3 that it can be decomposed

into a continuous and a discrete component. Thus, it is not necessary to explicitly select basis functions $\mathbf{b}_Y \in B^2(\mu)^{K_Y}$ for the mixed case. However, they can be obtained by concatenating the basis functions of the continuous and the discrete components.
Selecting the smoothing parameters is also important for regularization. They are specified such that the degrees of freedom are equal for all base-learners, to ensure a fair base-learner selection in each iteration of the algorithm. Otherwise, selection of more flexible base-learners is more likely than that of less flexible ones, see Hofner et al. (2011). However, the effective degrees of freedom of an effect after m_{stop} iterations will in general differ from those preselected for the base learners in each single iteration. They are successively adapted to the data. The starting coefficient vectors $\boldsymbol{\theta}_j^{[0]}$ are usually all set to zero, enabling variable selection as an effect that is never selected stays at zero. Like in functional regression, a suitable offset can be used for the intercept to improve the convergence rate of the algorithm, e.g., the mean density of the responses in $B^2(\mu)$. Note that a scalar offset, which is another common choice in functional regression, equals zero in the Bayes Hilbert space and thus corresponds to no offset. The optimal number of boosting iterations m_{stop} can be found with cross-validation, sub-sampling or bootstrapping, with samples generated on the level of elements of $B^2(\mu)$. The early-stopping avoids overfitting. Finally, the value $\kappa = 0.1$ for the step-length is suitable in most applications for a quadratic loss function (Brockhaus et al., 2020). A smaller step-length usually requires a larger value for m_{stop}.
Note that the estimation problem can also be solved in $L_0^2(\mu)$ based on the clr transformed model, with the estimates in $B^2(\mu)$ obtained applying the inverse clr transformation, as proposed by Talská et al. (2018) for functional linear models on closed intervals. For our functional additive models, gradient boosting can be performed in $L_0^2(\mu)$ analogously to the algorithm described above. The results of both algorithms are equivalent via the clr transformation, which we show in appendix C. In the continuous case, this yields the functional boosting algorithm of Brockhaus et al. (2015) with the modification that the basis functions $\mathbf{b}_Y$ are constrained to be elements of $L_0^2(\lambda)$ instead of $L^2(\lambda)$.

3.3 Estimation in the mixed case

Recall the mixed case, i.e., $B^2(\mu) = B^2(I, \mathfrak{B}, \mu)$ with $\mu = \delta + \lambda$, where $\delta = \sum_{d=1}^{D} w_d \, \delta_{t_d}$ for $\{t_1, \ldots, t_D\} = \mathcal{D} \subset I$ and $w_d > 0$. Due to the mixed reference measure, the specification of suitable basis functions $\mathbf{b}_Y \in B^2(\mu)^{K_Y}$ is not straightforward. We simplify this by tracing the estimation problem back to two separate estimation problems – one continuous and one discrete. For the continuous one, consider the Bayes Hilbert space $B^2(\lambda) = B^2(\mathcal{C}, \mathfrak{B} \cap \mathcal{C}, \lambda)$, where $\mathcal{C} := I \setminus \mathcal{D} \in \mathfrak{B}$. Remarkably, its orthogonal complement in $B^2(\mu)$ is not the Bayes Hilbert space $B^2(\mathcal{D}, \mathfrak{B} \cap \mathcal{D}, \delta)$. Instead, an additional arbitrary discrete value $t_{D+1} \in \mathbb{R} \setminus \mathcal{D}$ is required, which can be considered the discrete equivalent of $\mathcal{C}$. Thus, an intuitive choice is $t_{D+1} \in \mathcal{C}$. Then, the orthogonal complement of $B^2(\lambda)$ in $B^2(\mu)$ is the Bayes Hilbert space $B^2(\delta^\bullet) = B^2\left(\mathcal{D}^\bullet, \mathcal{P}(\mathcal{D}^\bullet), \delta^\bullet\right)$, where $\mathcal{D}^\bullet := \mathcal{D} \cup \{t_{D+1}\}$ and $\delta^\bullet := \sum_{d=1}^{D+1} w_d \, \delta_{t_d}$ with $w_{D+1} := \lambda(I)$. The embeddings to consider $B^2(\lambda)$ and

$B^2(\delta^\bullet)$ as subspaces of $B^2(\mu)$ are $\iota_\mathrm{c} : B^2(\lambda) \hookrightarrow B^2(\mu)$ and $\iota_\mathrm{d} : B^2(\delta^\bullet) \hookrightarrow B^2(\mu)$, which are defined as $\iota_\mathrm{c}(f_c) = f_c$ and $\iota_\mathrm{d}(f_\mathrm{d}) = f_\mathrm{d}\,(t_{D+1})$ on $\mathcal{C}$, respectively, and $\iota_\mathrm{c}(f_c) = \exp \mathcal{S}_\lambda(f_\mathrm{c})$ and $\iota_\mathrm{d}(f_\mathrm{d}) = f_\mathrm{d}$ on $\mathcal{D}$. Here, $\mathcal{S}_\lambda(f_\mathrm{c})$ is the mean logarithmic integral as defined in (2.1). Note that $\exp \mathcal{S}_\lambda(f_\mathrm{c})$ corresponds to the geometric mean of f_c using the natural generalization of the usual definition of the geometric mean of a discrete set $\{g(s_1), \ldots, g(s_L)\}$, since $(\prod_{l=1}^L g(s_l))^{1/L} = \exp \mathcal{S}_{B^2(\mathcal{T}, \mathcal{P}(\mathcal{T}), \sum_{l=1}^L \delta_{s_l})}(g)$ for $\mathcal{T} = \{s_1, \ldots, s_L\}$. For $f \in B^2(\mu)$, the unique functions $f_\mathrm{c} \in B^2(\lambda), f_\mathrm{d} \in B^2(\delta^\bullet)$ such that $f = \iota_\mathrm{c}(f_\mathrm{c}) \oplus \iota_\mathrm{d}(f_\mathrm{d})$ are given by

$$f_\mathrm{c} : \mathcal{C} \to \mathbb{R}, \quad t \mapsto f(t), \qquad f_\mathrm{d} : \mathcal{D}^\bullet \to \mathbb{R}, \quad t \mapsto \begin{cases} 1, & t = t_{D+1} \\ \frac{f(t)}{\exp \mathcal{S}_\lambda(f)}, & t \in \mathcal{D}. \end{cases} \tag{3.8}$$

See Proposition A.3 in appendix A.2 for the proof that the orthogonal complement of $B^2(\lambda)$ in $B^2(\mu)$ is $B^2(\delta^\bullet)$, including (3.8). Then, we obtain $\|f\|^2_{B^2(\mu)} = \|f_\mathrm{c}\|^2_{B^2(\lambda)} + \|f_\mathrm{d}\|^2_{B^2(\delta^\bullet)}$ implying that minimizing the sum of squared errors (3.3) is equivalent to minimizing its discrete and continuous components separately and then combining the solutions $\hat{f}_\mathrm{c}$ and $\hat{f}_\mathrm{d}$ in the overall solution $\hat{f} = \iota_\mathrm{c}(\hat{f}_\mathrm{c}) \oplus \iota_\mathrm{d}(\hat{f}_\mathrm{d})$.
Equivalently, we can decompose the Hilbert space $L^2_0(I, \mathfrak{B}, \mu)$ such that embeddings and clr transformations commute. See Proposition A.4 in appendix A.2 for details and proof.

3.4 Interpretation

The interpretation of the estimated effects $\hat{h}_j := \hat{h}_j^{[m_\mathrm{stop}]}(\mathbf{x}_i) \in B^2(\mu),\ j = 1, \ldots, J$, has to respect the special structure of Bayes Hilbert spaces. In particular, it should be independent of the selected representative of an equivalence class in $B^2(\mu)$. Naturally, interpretation in a Bayes Hilbert space is relative. Accordingly, the shape of clr transformed effects can be interpreted using differences, resulting in an interpretation analogous to the well-known odds ratios. For two effects $\hat{h}_j$ and $\hat{h}_k$ for $j \neq k \in \{1, \ldots, J\}$ and $s, t \in \mathcal{T}$, we have

$$\exp\left(\mathrm{clr}[\hat{h}_j](t) - \mathrm{clr}[\hat{h}_j](s) - \left(\mathrm{clr}[\hat{h}_k](t) - \mathrm{clr}[\hat{h}_k](s)\right)\right) = \frac{\hat{h}_j(t) \,/\, \hat{h}_j(s)}{\hat{h}_k(t) \,/\, \hat{h}_k(s)}. \tag{3.9}$$

The compound fraction on the right is called *odds ratio of $\hat{h}_j$ and $\hat{h}_k$ for t compared to s*, its numerator is called *odds of $\hat{h}_j$ for t compared to s*. Thus, the log odds ratio corresponds to the difference of the differences of the clr transformed effects evaluated at t and s. In a reference coding setting, this reduces to a simple difference as the clr transformed effect for the reference category is 0. Considering additional effects $\hat{h}_\mathcal{J} = \bigoplus_{l \in \mathcal{J}} \hat{h}_l$ with $\mathcal{J} \subset \{1, \ldots, J\} \setminus \{j, k\}$, the odds ratio of $\hat{h}_\mathcal{J} \oplus \hat{h}_j$ and $\hat{h}_\mathcal{J} \oplus \hat{h}_k$ for t compared to s is equal to the odds ratio of $\hat{h}_j$ and $\hat{h}_k$ for t compared to s, enabling a ceteris paribus interpretation.
The odds ratio (3.9) is a ratio of density values, which depending on the case (discrete, continuous, mixed) is identical to or approximates a usual ratio of probabilities: Let $\mathbb{P}_j \cong \hat{h}_j$ and $\mathbb{P}_k \cong \hat{h}_k$ be the corresponding probability measures in $B^2(\mu)$ of the estimated effects. In the discrete case, i.e., $\mathcal{T} = \{t_1, \ldots, t_D\}$ and

$\mu = \sum_{d=1}^{D} w_d \, \delta_{t_d}$, we have $\mathbb{P}_j(\{t_d\})/\mu(\{t_d\}) = [w_d \, \hat{h}_j(t_d)]/w_d = \hat{h}_j(t_d)$ for every $t_d \in \mathcal{T}$. Then, the odds ratio of $\hat{h}_j$ and $\hat{h}_k$ for $t_{d_1} \in \mathcal{T}$ compared to $t_{d_2} \in \mathcal{T}$ equals $[\mathbb{P}_j(\{t_{d_1}\})/\mathbb{P}_j(\{t_{d_2}\})] \,/\, [\mathbb{P}_k(\{t_{d_1}\})/\mathbb{P}_k(\{t_{d_2}\})]$, i.e., the odds ratio of $\mathbb{P}_j$ and $\mathbb{P}_k$ for $\{t_{d_1}\}$ compared to $\{t_{d_2}\}$. In the continuous and mixed cases, i.e., $\mathcal{T} = I \subset \mathbb{R}$ and $\mu = \sum_{d=1}^{D} w_d \, \delta_{t_d} + \lambda$ for $\mathcal{D} = \{t_1, \ldots, t_D\} \subset I$ (continuous case: $\mathcal{D} = \emptyset$), the relation holds approximately: For $s, t \in I$, let $A_n, B_n \subseteq I$ be two nested sequences of intervals centered at s and t for all $n \in \mathbb{N}$, whose intersection is $\{s\}$ and $\{t\}$, respectively. Then,

$$\frac{\hat{h}_j(t)}{\hat{h}_j(s)} = \lim_{n\to\infty} \frac{\mathbb{P}_j(B_n) \,/\, \mu(B_n)}{\mathbb{P}_j(A_n) \,/\, \mu(A_n)} \quad \text{and thus} \quad \frac{\hat{h}_j(t) \,/\, \hat{h}_j(s)}{\hat{h}_k(t) \,/\, \hat{h}_k(s)} = \lim_{n\to\infty} \frac{\mathbb{P}_j(B_n) \,/\, \mathbb{P}_j(A_n)}{\mathbb{P}_k(B_n) \,/\, \mathbb{P}_k(A_n)}, \tag{3.10}$$

i.e., the odds ratio of density values approximates the odds ratio of probabilities for small neighborhoods of s and t. We prove (3.10) in appendix A.2, where we also show that if there exist $I_t, I_s \subset I$ with $\hat{h}_j(t)/\hat{h}_j(s) < \hat{h}_k(t)/\hat{h}_k(s)$ for all $t \in I_t, s \in I_s$, then, $\mathbb{P}_j(I_t)/\mathbb{P}_j(I_s) < \mathbb{P}_k(I_t)/\mathbb{P}_k(I_s)$. Further ideas of interpreting effects are developed in appendix D.

4 Application

With our modeling approach, we analyze the distribution of the female share in a couple's total labor income in Germany. Note that for simplicity we use the terms East/West Germany also after reunification. Although we refer to Bertrand et al. (2015), we do not focus on the question of whether there is actually a decline in density at 0.5.

4.1 Background and hypotheses

There is a larger share fraction in Germany below 0.5 (as in Bertrand et al., 2015) reflecting the gender pay gap, but there is no consensus in the literature regarding a discontinuous drop at 0.5 (Sprengholz et al., 2020; Kuehnle et al., 2021). The employment and earnings of female partners show a strong childhood penalty (Kleven et al., 2019; Fitzenberger et al., 2013). The social norm in West Germany used to be that mothers should stay at home with their children. Institutionalized child care was scarce and there are strong financial incentives for part-time work for the second earner. Together, this results in part-time employment increasing strongly for women after having their first child. Thus, we expect that the income share of the woman is lower in the presence of children reflecting a childhood penalty.
Due to changing social norms, the female employment increases strongly over time. However, occupational segregation by gender is persistent (Cortes and Pan, 2018) with men being more likely to work in better paying occupations. Still, occupations with a higher share of women seem to benefit from technological change (Black and Spitz-Oener, 2010). Thus, the income share of female partners without children is predicted to grow over time.

Ex ante reasoning suggests an ambiguous effect on the childhood penalty. On the one hand, the incentives for part-time work especially for female partners with young children may prevent an increase in the income share. Thus, the childhood penalty in the income share may even grow over time. On the other hand, growing female employment may actually increase the female income share, especially among female partners with older children.
Turning to the comparison between East and West Germany, the literature emphasizes that social norms are likely to differ between the two parts of the country (Beblo and Görges, 2018). Before reunification, it was basically mandatory for women to work in East Germany and comprehensive institutionalized child care was available. This suggests that the female income share in East Germany is higher than in West Germany.
After reunification, social norms have been converging between the East and the West. In East Germany, female employment may have fallen more strongly than for males due to the strong economic transformation and the lower mobility of female partners after job loss. Part-time employment is likely to become more prevalent in East Germany, and over time mothers more often drop out of the labor force. While we expect the childhood penalty to be lower in East Germany than in West Germany, it is ex ante ambiguous whether the East-West gap in the childhood penalty decreases over time, a question of interest.

4.2 Data and descriptive evidence on response densities

Our data set derived from the German Socio-Economic Panel (see appendix E for details) contains 154,924 observations of couples of opposite sex living together in a household, where at least one partner reports positive labor income. We include cohabitating couples in addition to married ones as there is a strong tax incentive to get married in case of unequal incomes, leading to a bias. The women's *share* in the couple's total gross labor income together with the household's sample *weight* yields the response densities. Four variables serve as covariates. First, the binary covariate *West_East* specifies whether the couple lives in *West* Germany or in *East* Germany (including Berlin). A second finer disaggregation distinguishes six *regions* (two in *East* and four in *West* Germany, see appendix E.1). The third covariate *c_age* is a categorical variable for the age range (in years) of the couple's youngest child living in the household: *0-6*, *7-18*, and *other* (i.e., couples without minor children). Finally, *year* ranges from 1984 (*West* Germany)/1991 (*East* Germany) to 2016.
A response density $f_{region,\ c\_age,\ year} : [0,1] \to \mathbb{R}^+, s \mapsto f_{region,\ c\_age,\ year}(s)$ is estimated for each combination of covariate values (note that *region* determines *West_East*), with s denoting the woman's income share. In total, this yields 552 response densities. Often, we just write f and omit the indices. Before elaborating on the estimation, we determine a suitable underlying Bayes Hilbert space $B^2(\mu) = B^2(\mathcal{T}, \mathcal{A}, \mu)$. Since s denotes a share, we consider $\mathcal{T} = [0,1]$ with $\mathcal{A} = \mathfrak{B}$. The Lebesgue measure is no appropriate reference, as the boundary values 0 and 1 correspond to single-earner households and thus have positive probability mass (see appendix E.2 for exemplary barplots). A suitable reference measure respecting this structure is $\mu := \delta_0 + \lambda + \delta_1$, i.e., the mixed case with $D = 2,\ t_1 = 0,\ t_2 = 1$, and $w_1 = 1 = w_2$,

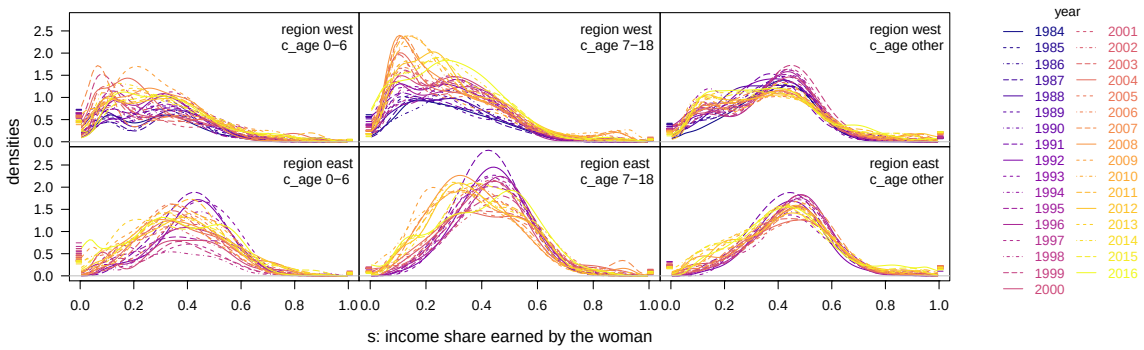

Figure 4.1: Response densities for *regions west* and *east* [rows] for all three values of *c_age* [columns].

see Section 3. The values $f(0)$ and $f(1)$ are the (weighted) relative frequencies for shares of 0 and 1, denoted by p_0 and p_1, respectively. To estimate f on $(0, 1)$, we compute continuous densities based on dual-earner households, and multiply them by $p_{(0,1)} = 1 - p_0 - p_1$. For this purpose, weighted kernel density estimation with beta-kernels (Chen, 1999) is used to preserve the support $(0, 1)$, see appendix E.3 for details.

The response densities are very similar in the different *regions* within *West* and *East* Germany, respectively. Thus, we restrict visualization in Figure 4.1 to the exemplary *regions west* (North Rhine-Westphalia) for *West* Germany and *east* (Saxony-Anhalt, Thuringia, Saxony) for *East* Germany. See Figure E.7 in appendix E.4 for the corresponding figure for all six *regions*, with additional illustration of the respective relative frequencies p_0, $p_{(0,1)}$, p_1 over time. Figure 4.1 depicts the response densities for all *years* by the *c_age* groups, for the *regions west* and *east*, with a color gradient and different line types distinguishing the *year*. The density values $f(0)$ and $f(1)$ are represented as dashes, shifted slightly outwards for better visibility. Consider the continuous parts ($s \in (0, 1)$): In *west* (first row), the densities differ between couples with (*0-6* and *7-18*) and without minor children (*other*), with the latter lying more to the right reflecting lower female shares in the presence of children. In *east*, the shapes are more egalitarian and vary much less with the age of the youngest child. In all cases, the fraction of couples with a share less than 0.5 exceeds the fraction with a share larger than 0.5. Over time, the probability mass for a small share increases and the share of non-working women declines, reflecting the increase in female part-time employment. These findings show the importance of considering the mixed densities. The shares of dual-earner households and non-working women evolve in opposite direction over time, while the share of single-earner women remains small.

4.3 Model specification

Based on the empirical response densities $f_{region,\ c\_age,\ year}$, we estimate the model

$$\begin{aligned} f_{region,\ c\_age,\ year} = \beta_0 &\oplus \beta_{West\_East} \oplus \beta_{region} \oplus \beta_{c\_age} \oplus \beta_{c\_age,\ West\_East} \\ &\oplus g(year) \oplus g_{West\_East}(year) \oplus g_{c\_age}(year) \\ &\oplus g_{c\_age,\ West\_East}(year) \oplus \varepsilon_{region,\ c\_age,\ year}. \end{aligned} \tag{4.1}$$

All summands are densities of the share $s \in [0, 1]$ and elements of the Bayes Hilbert space $B^2(\mu)$. The model is reference coded with reference categories $West\_East = West$, $c\_age = other$, and $year = 1991$. The corresponding effect for the reference is given by the intercept β_0. The effect for the six regions β_{region} is centered around the respective β_{West_East}. The smooth year effect $g(year)$ describes the deviation for each *year* from the reference 1991 (for *West* Germany and *c_age other*). Finally, several interaction terms are included with a group-specific intercept $\beta_{c\_age,\ West\_East}$ as well as group-specific flexible terms $g_{West\_East}(year)$, $g_{c\_age}(year)$, and $g_{c\_age,\ West\_East}(year)$. They are constrained to be orthogonal to the respective main effects using a similar constraint as (3.2) to ensure identifiability. Due to reference coding, all partial effects for the reference categories are zero.
As described in Section 3.3, we decompose the Bayes Hilbert space $B^2(\mu)$ into two orthogonal subspaces $B^2(\lambda) = B^2((0,1), \mathfrak{B} \cap (0,1), \lambda)$ and $B^2(\delta^\bullet) = B^2(\mathcal{D}^\bullet, \mathcal{P}(\mathcal{D}^\bullet), \delta^\bullet)$, where $\mathcal{D}^\bullet = \{t_1, t_2, t_3\}$ and $\delta^\bullet = \sum_{d=1}^{3} \delta_{t_d}$. We choose $t_3 = 1/2$ to represent the continuous component in between the boundary values $t_1 = 0$ and $t_2 = 1$. For every f we generate the unique functions $f_\mathrm{c} \in B^2(\lambda)$ and $f_\mathrm{d} \in B^2(\delta^\bullet)$ as in (3.8). As proposed in Section 3.2, we choose transformed cubic B-splines as basis functions $\mathbf{b}_Y$ for the continuous component and a transformed basis of indicator functions for the discrete component. The remaining specification is identical in both models. We use an anisotropic penalty without penalizing in direction of the share, i.e., $\lambda_Y = 0$, to ensure the necessary flexibility towards the boundaries. For the flexible nonlinear effects, the selected basis functions are cubic B-splines with penalization of second order differences. We set the degrees of freedom to 2 for all effects but β_0 and β_{West_East}, as these only allow for a maximum value of 1. Regarding base-learner selection, β_{West_East} thus is at a slight disadvantage compared to other main effects. However, in a sensitivity check imposing equal degrees of freedom for all base-learners by adjusting λ_Y to 1 for all effects, we do not observe large deviations in the selection frequencies while the fit to the data is better with unequal degrees of freedom, see appendix E.4. Note that the intercept as well as the interaction effects are separated from the main effects due to the orthogonalizing constraints, ensuring a fair selection for the remaining base-learners. The starting coefficients are set to zero in every component and we set the step-length κ to 0.1. We obtain a stopping iteration value of 262 for the continuous model and 731 for the discrete model based on 25 bootstrap samples, respectively.

4.4 Regression Results

All effects in our regression model (4.1) are selected by the algorithm (see appendix E.5). The predictions in Figure E.8 in appendix E.4 mostly show a good fit.

In the following, we discuss the key findings.

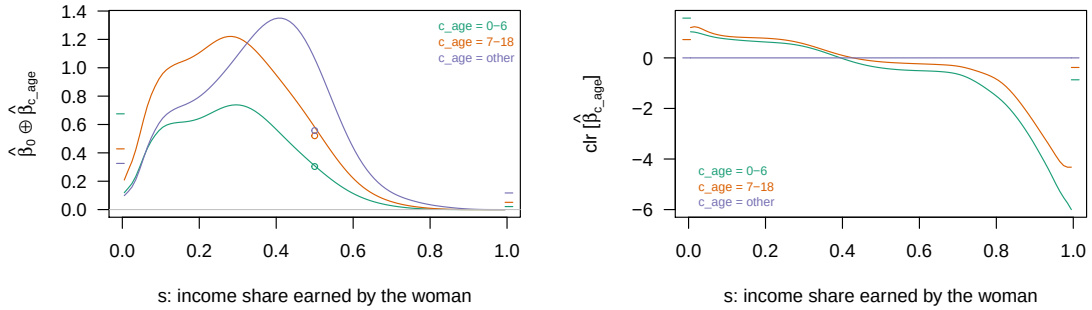

Figure 4.2: Expected densities for couples living in *West* Germany in 1991 for all three values of *c_age* [left] and clr transformed estimated effects of *c_age* [right].

The left part of Figure 4.2 shows the perturbation of the intercept by the *c_age* effect, i.e., the expected densities for couples without minor children (*c_age other*), for couples with children aged *0-6*, and for couples with children aged *7-18* living in *West* Germany in 1991. The circles at 0.5 represent the expected relative frequency of dual-earner households. Our main finding is that the expected density on $(0, 1)$ for *c_age other* is unimodal with a maximum above 0.4, while the densities for *c_age 0-6* and *7-18* are bimodal with both maxima to the left of 0.4. The latter show a similar shape, but are scaled differently. The relative frequencies of dual-earner households (circles at 0.5) and the two types of single-earner households (dashes at 0, 1) are similar for couples with children aged *7-18* years and couples without minor children, respectively. In contrast, the relative frequency of non-working women is much higher and the relative frequency of dual-earner households is much lower for couples with children aged *0-6*. The right part of the figure shows the clr transformed effect for interpretation via (log) odds ratios, see Section 3.4. As *c_age=other* is the reference category, we have $\text{clr}[\hat{\beta}_{other}] = 0$. The clr transformed effects of *c_age 0-6* and *7-18* again show similar shapes on $(0, 1)$, but shifted vertically. As the log odds ratio of $\hat{\beta}_k$ and $\hat{\beta}_{other}$ for s compared to t corresponds to vertical differences within $\text{clr}[\hat{\beta}_k]$, $k \in \{$*0-6*, *7-18*$\}$, the log odds ratio of $\hat{\beta}_{0\text{-}6}$ and $\hat{\beta}_{other}$ is similar to the one of $\hat{\beta}_{7\text{-}18}$ and $\hat{\beta}_{other}$. This implies they have similar impact on the shape of a density. Both log odds ratios are always negative for $s < t \in (0, 1)$, i.e., the odds for a larger versus a smaller income share are always smaller for couples with minor children than for couples without minor children, reflecting the strong childhood penalty in *West* Germany in 1991. See Appendix E.5 for quantitative examples of concrete odds ratios.
Figure 4.3 shows the expected densities for four selected *years*, separately for couples with and without minor children (see Figure E.16 in appendix E.5 for all *years*). For *other*, the frequency of non-working women ($s = 0$) falls continuously over time and the density becomes more dispersed with a lower maximum around 0.4 in 2016 than in 1993 and 2003 (however, it was even lower in 1984). In fact, by 2016 the expected density tends to have a second maximum further left and a heavier tail on the right, most likely due to the strong growth of part-time employment even among women without minor children. Furthermore, the frequency of single-earner

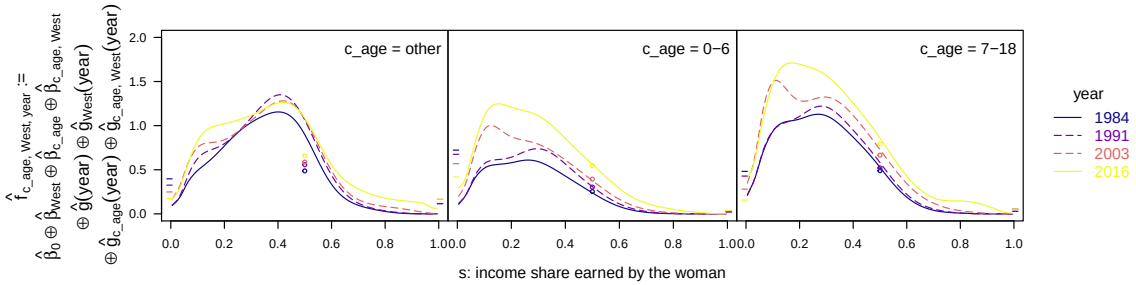

Figure 4.3: Expected densities in the *years* 1984, 1991, 2003, and 2016 for *West* Germany for all three values of *c_age*: *other* [left], *0-6* [middle], *7-18* [right].

women ($s = 1$) increases to a level similar to the frequency of non-working women. For *0-6* and *7-18*, we also observe a fall in the frequency of non-working women and a stronger concentration around the larger mode until 1991. However, up to 2016 the distributions show more probability mass for small shares, reflecting the even larger growth of part-time employment among women with minor children.

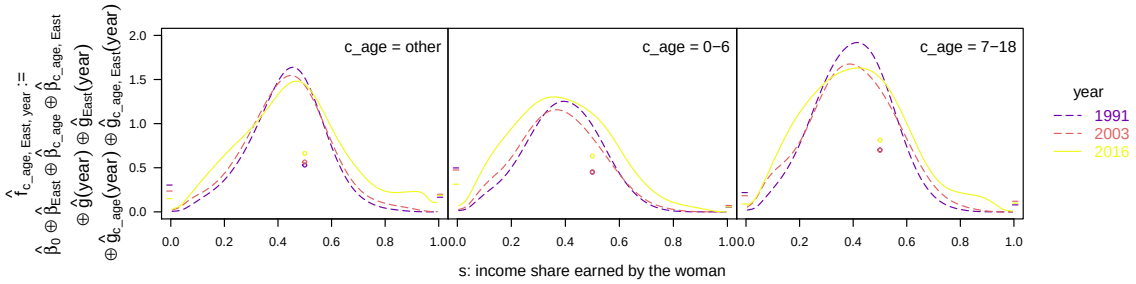

Figure 4.4: Expected densities in the *years* 1991, 2003, and 2016 for *East* Germany for all three values of *c_age*: *other* [left], *0-6* [middle], *7-18* [right].

Figure 4.4 shows the expected densities in *East* Germany for selected *years* (see Figure E.16 in appendix E.5 for all *years*). In all three cases, the share distribution has a unique mode at or above 0.4. The distribution becomes more dispersed over time, with more probability mass moving to the left and a growing right tail. The frequency of non-working women is falling over time. While showing a similar trend as in *West* Germany, there remain persistent differences. In *East* Germany, the frequency of non-working women for couples with minor children remains much lower and the shape of the distribution shows no trend towards a second maximum at a low share. Hence, there remains a considerable West-East gap in the childhood penalty.
To address this issue explicitly, the West-East gap in the childhood penalty is calculated by the difference-in-differences (DiD) effect for $year \in \{1991, 2016\}$ and $c\_age \in \{0\text{-}6, 7\text{-}18\}$:

$$DiD_{c\_age,\ year} = (\hat{f}_{c\_age,\ West,\ year} \ominus \hat{f}_{other,\ West,\ year}) \ominus (\hat{f}_{c\_age,\ East,\ year} \ominus \hat{f}_{other,\ East,\ year}).$$

Figure 4.5 shows the log odds

$$LO_{c\_age,\ year}(t, s) := \log\left([DiD_{c\_age,\ year}](t)/[DiD_{c\_age,\ year}](s)\right)$$

of $DiD_{c\_age,\ year}$ for t compared to s for pairs $(t, s) \in [0, 1]^2$, see Section 3.4, as heat maps. We omit the index *c_age, year* in the following. The inner quadrant shows the respective heat map for $t, s \in (0, 1)$. The log odds involving the two mass points 0 and 1 are given by the band around the inner quadrants. The top-left corner concerns the log odds for $t = 0$ (non-working woman) compared to $s = 1$ (single-earner woman). The inner bands around the inner quadrant correspond to the log odds between a mass point $0, 1$ and a share in $(0, 1)$. The outer bands show the constant log odds between one of the mass points and the event dual-earner household $(0 < s, t < 1)$. A positive (negative) value implies that the log odds for shares t versus s are higher (lower) in the *West* than in the *East*. Thus, $LO(t, s) > 0$ for $t < s$ implies that the child penalty (lower share t is more likely relative to s in the presence of children) is more pronounced (stronger) in the *West* than in the *East*. For 1991, the vertical band for $t = 0$ to the left of the heatmap is quite red $(LO(0, s) > 0)$, implying that it is much more likely that women in the *West* compared to the *East* stop working in the presence of a child, relative to all other shares. This holds both for *c_age 0-6* (top panel) and *c_age 7-18* (bottom panel). However, the entire heatmap shows positive (negative) values above (below) the 45-degree-line implying that the shift to lower shares compared to higher shares in the presence of children is stronger in the *West* than in the *East*, with the West-East gap in the child penalty being even larger for *c_age 7-18*.

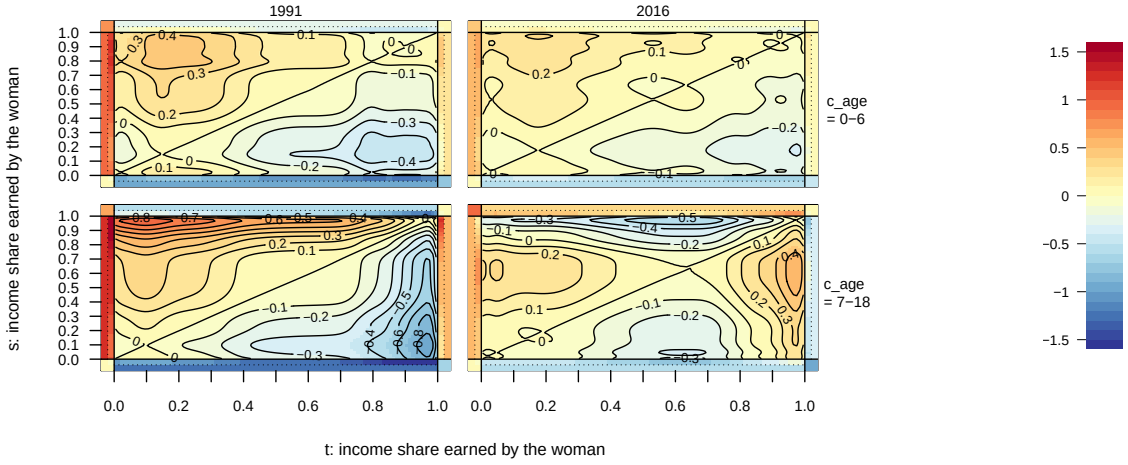

Figure 4.5: Log Odds $LO_{c\_age,\ year}(t, s)$ of the West-East gap in the childhood penalty (DiD effects) for *c_age 0-6* [top] and *7-18* [bottom] for the *years* 1991 [left] and 2016 [right].

The comparison between the two years is informative about the change in the West-East gap in the childhood penalty over time. In 2016, the childhood penalty remains larger in the *West* compared to the *East* over almost the entire share distribution – only for *c_age 7-18* is there a reversal for very large shares compared to medium share levels. However, since the absolute log odds have become much smaller, especially for non-working women, the West-East gap in the childhood penalty has decreased considerably over time.
Summarizing our main findings, the frequency of non-working women and women with a lower income share is higher in *West* Germany than in *East* Germany and

these differences are larger for couples with children. Over time, the share of non-working women decreased. Among dual-earner households the dispersion of the share distribution increased over time with both a growing lower and higher tail. Despite persistent East-West differences in the share distributions and the child penalty until the end of the observation period, the West-East gap in the childhood penalty fell considerably over time.

5 Simulation study

The gradient boosting approach has already been tested extensively in several simulation studies for scalar and functional data (e.g., Brockhaus et al. (2015) and references therein). For completeness and to validate our modified approach for density-on-scalar models, we present a small simulation study for this case. It is based on the results of our analysis in Section 4. The predictions obtained there serve as true mean response densities for the simulation and are denoted by $F_i \in B^2(\mu),\ i = 1, \dots, 552$, where each i corresponds to one combination of values for the covariates *region*, *c_age*, and *year* and $B^2(\mu)$ is the Bayes Hilbert space from Section 4. To simulate data, we perform a functional principal component (PC) analysis (e.g. Ramsay and Silverman, 2005) on the clr transformed functional residuals $\text{clr}[\hat{\varepsilon}_i] = \text{clr}[f_i \ominus F_i] = \text{clr}[f_i] - \text{clr}[F_i]$, with $f_i \in B^2(\mu)$ the response densities from the application. Let ψ_m denote the PC functions corresponding to the descending ordered eigenvalues ξ_m and let ρ_{im} denote the PC scores for $i = 1, \dots, 552$ and $m \in \mathbb{N}$. Then, the truncated Karhunen-Loève expansion for $M \in \mathbb{N}$ yields an approximation of the functional residuals: $\text{clr}[\hat{\varepsilon}_i] \approx \sum_{m=1}^{M} \rho_{im}\psi_m$. The PC scores can be viewed as realizations of uncorrelated random variables ρ_m with zero-mean and covariance $\text{Cov}(\rho_m, \rho_n) = \xi_m \delta_{mn}$, where δ_{mn} denotes the Kronecker delta and $m, n = 1, \dots, M$. We simulate residuals $\tilde{\varepsilon}_i$ by drawing uncorrelated random $\tilde{\rho}_{im}$ from mean zero normal distributions with variance ξ_m and applying the inverse clr transformation to the truncated Karhunen-Loève expansion, $\tilde{\varepsilon}_i = \text{clr}^{-1}[\sum_{m=1}^{M} \tilde{\rho}_{im}\psi_m] = \bigoplus_{m=1}^{M} \tilde{\rho}_{im} \odot \text{clr}^{-1}[\psi_m]$. Adding these to the mean response densities yields the simulated data: $\tilde{f}_i = F_i \oplus \tilde{\varepsilon}_i,\ i = 1, \dots, 552$. Using these as observed response densities, we then estimate model (4.1) and denote the resulting predictions with $\hat{f}_i \in B^2(\mu),\ i = 1, \dots, 552$. We replicate this approach 200 times with $M = 102$, which is the maximal possible value due to the number of available grid points per density.

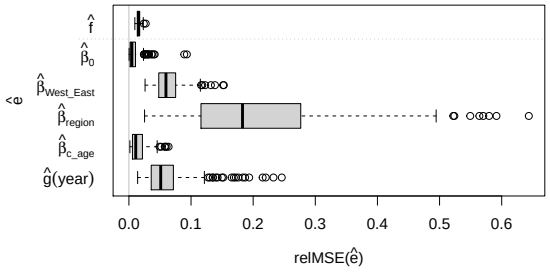

Figure 5.1: RelMSE for predictions [top] and main effects [bottom].

To evaluate the goodness of the estimation results, we use the relative mean squared error (relMSE; defined in appendix F.1) motivated by Brockhaus et al. (2015), standardizing the mean squared error with respect to the global variability of the true density. Figure 5.1 shows the boxplots of the relMSEs (200 each) of the predictions and the main effects. All effects are illustrated in appendix F.2. The distribution of relMSE($\hat{f}$) over the 200 simulation runs shows good estimation quality, with a median of 1.55%. Regarding the main effects, the relMSEs are the smallest for $\hat{\beta}_0$ and $\hat{\beta}_{c\_age}$ with medians of 0.48% and 1.1%, respectively. For $\hat{\beta}_{West\_East}$ and $\hat{g}(year)$, the values tend to be slightly larger (medians: 5.96% and 5.12%) while they are clearly larger for $\hat{\beta}_{region}$ (median: 18.28%). However, the larger relative values, especially for $\hat{\beta}_{region}$, arise from the variability of the true effects being small, not from the mean squared errors being large. This is also the case for the interaction effects, see appendix F.2. Regarding model selection, the main effects are all selected in each simulation run, while the smaller interaction effects are not, see appendix F.3 for details. Overall, the estimates capture the true means F_i and all effects that are pronounced very well. Small effects in the model are estimated well in absolute, but badly in relative terms.

6 Conclusion

We presented a flexible framework for density-on-scalar regression models by formulating them in a Bayes Hilbert space $B^2(\mu)$, which respects the nature of probability densities and allows for a unified treatment of arbitrary finite measure spaces. This covers in particular the common discrete, continuous, and mixed cases. To estimate the covariate effects in $B^2(\mu)$, we introduced a gradient boosting algorithm. We used our approach to analyze the distribution of the woman's share in a couple's total labor income, an example of the challenging mixed case, for which we developed a decomposition into a continuous and a discrete estimation problem. We observe strong differences between West and East Germany and between couples with and without children. Among dual-earner households the dispersion of the share distribution increased over time. Despite persistent East-West differences in the share distributions and the child penalty until the end of the observation period, the West-East gap in the childhood penalty fell considerably over time. Finally, we performed a small simulation study justifying our approach in a setting motivated by our application.

Density regression has particular advantages in terms of interpretation compared to approaches considering equivalent functions like quantile functions (e.g., Yang et al., 2018; Koenker, 2005) or distribution functions (CTMs, e.g., Hothorn et al., 2014; distribution regression, e.g., Chernozhukov et al., 2013), as shifts in probability masses or bimodality are easily visible in densities. Odds-ratio-type interpretations of effect functions further add to the interpretability of our model. A crucial part in our approach is played by the clr transformation, which simplifies among other things estimation, as gradient boosting can be performed equivalently on the clr transformed densities in $L_0^2(\mu)$. This allows taking advantage of and extending existing implementations for function-on-scalar regression like the `R` add-on package `FDboost` (Brockhaus and Rügamer, 2018), see the github repository *FDboost* for

our enhanced version of the package and in particular our vignette "density-on-scalar_birth". The idea to transform the densities to (a subspace of) the well-known L^2 space with its metric is also used by other approaches. Besides the clr transformation, the square root velocity transformation (Srivastava et al., 2007) as well as the log hazard and log quantile density transformations (e.g., Han et al., 2020) are popular choices. The approach of Petersen and Müller (2019) does not use a transformation, but also computes the applied Wasserstein metric via the L^2 metric. What is special about the clr transformation based Bayes Hilbert space approach, is the embedding of the untransformed densities in a Hilbert space structure. It is the extension of the well-established Aitchison geometry (Aitchison, 1986), which provides a reasonable framework for compositional data – the discrete equivalent of densities – fulfilling appealing properties like subcompositional coherence. The clr transformation helps to conveniently interpret covariate effects via ratios of density values (odds-ratios), which approximate or are equal to ratios of probabilities in three common cases (discrete, continuous, mixed). Modeling those three cases in a unified framework is a novelty to the best of the authors' knowledge, and a contribution of our approach to the literature on density regression.

Due to the gradient boosting algorithm used for estimation, our method includes variable selection and regularization, while it can deal with numerous covariates. However, like all gradient boosting approaches, it is limited by not naturally yielding inference – unlike some existing approaches (e.g., Petersen and Müller, 2019). This might be developed using a bootstrap-based approach or selective inference (Rügamer and Greven, 2020) in the future. Alternatively, other estimation methods allowing for formal inference could be derived.

The (current) definition of Bayes Hilbert spaces, which only allows finite reference measures, does not cover the interesting case of the measurable space $(\mathbb{R}, \mathfrak{B}_\mathbb{R})$ with Lebesgue measure λ. Though $(\mathbb{R}, \mathfrak{B}_\mathbb{R})$ can still be considered using, e.g., the probability measure corresponding to the standard normal distribution (Boogaart et al., 2014) as reference, it would be desirable to extend Bayes Hilbert spaces to σ-finite reference measures, allowing for $B^2(\mathbb{R}, \mathfrak{B}_\mathbb{R}, \lambda)$. Moreover, Bayes Hilbert spaces include only (μ-a.e.) positive densities. While in the continuous case, values of zero can in many cases be avoided using a suitable density estimation method, they are often replaced with small values in the discrete case (see Pawlowsky-Glahn et al., 2015). In contrast, the square root velocity transformation (Srivastava et al., 2007) allows density values of zero and may be an alternative in such cases, at the price of loosing the Hilbert space structure for the untransformed densities.

Finally, while in practice densities are sometimes directly reported, one often does not observe the response densities directly, but has to first estimate them from individual data to enable the use of density-on-scalar regression. This causes two problems. First, when treating estimated densities as observed, like also in other approaches such as Petersen and Müller (2019) and Han et al. (2020), estimation uncertainty is not accounted for in the analysis. Second, the number of individual observations for each covariate value combination which is available for density estimation can limit the number of covariates that can be included in the model. In the future, we thus aim to extend our approach to also model conditional densities for individual observations, still allowing flexibility in the covariate effects, but with-

out restrictive assumptions such as a particular distribution family as in GAMLSS (Rigby and Stasinopoulos, 2005).

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London, UK, UK: Chapman & Hall, Ltd.

Beblo, M. and Görges, L. (2018). On the nature of nurture. The malleability of gender differences in work preferences. *Journal of Economic Behavior & Organization* **151**, 19–41.

Bertrand, M., Kamenica, E., and Pan, J. (2015). Gender Identity and Relative Income within Households. *The Quarterly Journal of Economics* **130**, 571–614.

Black, S. E. and Spitz-Oener, A. (2010). Explaining women's success: technological change and the skill content of women's work. *The Review of Economics and Statistics* **92**, 187–194.

Boogaart, K. G. van den, Egozcue, J. J., and Pawlowsky-Glahn, V. (2010). Bayes linear spaces. *SORT: statistics and operations research transactions* **34**, 201–222.

– (2014). Bayes Hilbert Spaces. *Australian & New Zealand Journal of Statistics* **56**, 171–194.

Boogaart, K. G. van den, Tolosana-Delgado, R., and Templ, M. (2015). Regression with compositional response having unobserved components or below detection limit values. *Statistical Modelling* **15**, 191–213.

Brockhaus, S. and Rügamer, D. (2018). *FDboost: Boosting Functional Regression Models*. R package version 0.3-2.

Brockhaus, S., Rügamer, D., and Greven, S. (2020). Boosting Functional Regression Models with FDboost. *Journal of Statistical Software* **94**, 1–50.

Brockhaus, S., Scheipl, F., Hothorn, T., and Greven, S. (2015). The functional linear array model. *Statistical Modelling* **15**, 279–300.

Bühlmann, P. and Yu, B. (2003). Boosting with the L2 loss: regression and classification. *Journal of the American Statistical Association* **98**, 324–339.

Chen, S. X. (1999). Beta kernel estimators for density functions. *Computational Statistics & Data Analysis* **31**, 131–145.

Chernozhukov, V., Fernández-Val, I., and Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica* **78**, 1093–1125.

Chernozhukov, V., Fernández-Val, I., and Melly, B. (2013). Inference on counterfactual distributions. *Econometrica* **81**, 2205–2268.

Cortes, P. and Pan, J. (2018). Occupation and gender. *The Oxford handbook of women and the economy*, 425–452.

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.

Fitzenberger, B., Sommerfeld, K., and Steffes, S. (2013). Causal effects on employment after first birth—A dynamic treatment approach. *Labour Economics* **25**, 49–62.

Han, K., Müller, H.-G., and Park, B. U. (2020). Additive functional regression for densities as responses. *Journal of the American Statistical Association* **115**, 997–1010.

Happ, C., Scheipl, F., Gabriel, A., and Greven, S. (2019). A general framework for multivariate functional principal component analysis of amplitude and phase variation. *Stat* **8**, e220.

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.

Hothorn, T., Kneib, T., and Bühlmann, P. (2014). Conditional transformation models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 3–27.

Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley Series in Probability and Statistics. Chichester: John Wiley & Sons, Ltd.

Kleven, H., Landais, C., Posch, J., Steinhauer, A., and Zweimüller, J. (2019). Child penalties across countries: Evidence and explanations. *AEA Papers and Proceedings*. Vol. 109, 122–26.

Koenker, R. (2005). *Quantile Regression*. Econometric Society Monographs. Cambridge University Press.

Kuehnle, D., Oberfichtner, M., and Ostermann, K. (2021). Revisiting Gender Identity and Relative Income within Households: A Cautionary Tale on the Potential Pitfalls of Density Estimators. *Journal of Applied Econometrics*, to appear.

Li, R., Reich, B. J., and Bondell, H. D. (2021). Deep distribution regression. *Computational Statistics & Data Analysis* **159**, 107203.

Lutz, R. W. and Bühlmann, P. (2006). Boosting for high-multivariate responses in high-dimensional linear regression. *Statistica Sinica*, 471–494.

MacEachern, S. N. (1999). Dependent nonparametric processes. *ASA proceedings of the section on Bayesian statistical science*. Vol. 1. Alexandria, Virginia. Virginia: American Statistical Association; 1999, 50–55.

Ollivier, Y., Pajot, H., and Villani, C. (2014). *Optimal Transport: Theory and Applications*. Vol. 413. Cambridge University Press.

Park, J. Y. and Qian, J. (2012). Functional regression of continuous state distributions. *Journal of Econometrics* **167**, 397–412.

Pawlowsky-Glahn, V., Egozcue, J. J., and Tolosana-Delgado, R. (2015). *Modeling and analysis of compositional data*. John Wiley & Sons.

Petersen, A. and Müller, H.-G. (2019). Fréchet regression for random objects with Euclidean predictors. *The Annals of Statistics* **47**, 691–719.

Ramsay, J. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer-Verlag New York.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **54**, 507–554.

Rügamer, D. and Greven, S. (2020). Inference for L2-Boosting. *Statistics and Computing* **30**, 279–289.

Sprengholz, M., Wieber, A., and Holst, E. (2020). Gender identity and wives' labor market outcomes in West and East Germany between 1983 and 2016. *Socio-Economic Review*, to appear.

Srivastava, A., Jermyn, I., and Joshi, S. (2007). Riemannian Analysis of Probability Density Functions with Applications in Vision. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8.

Stöcker, A., Brockhaus, S., Schaffer, S., Bronk, B. von, Opitz, M., and Greven, S. (2021). Boosting Functional Response Models for Location, Scale and Shape with an Application to Bacterial Competition. *Statistical Modelling*, to appear.

Takeuchi, I., Le, Q. V., Sears, T. D., and Smola, A. J. (2006). Nonparametric quantile estimation. *Journal of machine learning research* **7**, 1231–1264.

Talská, R., Menafoglio, A., Machalová, J., Hron, K., and Fišerová, E. (2018). Compositional regression with functional response. *Computational Statistics & Data Analysis* **123**, 66–85.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*. 2nd ed. Boca Raton: Chapman and Hall/CRC.

Yang, H., Baladandayuthapani, V., Rao, A. U. K., and Morris, J. S. (2018). Regression Analyses of Distributions using Quantile Functional Regression. *arXiv preprint arXiv:1810.03496*.

Part II.

Modeling Multidimensional Curves and their Shape

4. Multivariate Functional Additive Mixed Models

In this chapter, we propose a multivariate functional mixed model for irregularly/sparsely sampled functional data and for nested/crossed experimental designs. This is required in two data problems we consider: phonetic data of acoustic measurements and EPG measurements during speech production and human motion trajectories in billiards. The multidimensional covariance surface is hereby estimated via symmetric covariance smoothing and allows for interpretation of the correlation within and across curves via multivariate functional principal component analysis, in addition to interpretation of covariate effects.

Contributing article:
Volkmann, A., Stöcker, A., Scheipl, F., and Greven, S. (2021). Multivariate Functional Additive Mixed Models. *Statistical Modelling*. DOI: 10.1177/1471082X211056158

Declaration on personal contributions:
This project arose from a master's thesis topic that the author of this thesis proposed, prototyped and supervised together with Fabian Scheipl and Sonja Greven. When Alexander Volkmann continued and substantially extended his master's thesis project during his own doctoral studies, the author of this thesis contributed the data collaboration for the novel billard's application (including substantial pre-processing of the data, later also together with Alexander Volkmann) and was passively/supportingly involved throughout the process taking a consulting role.

Statistical Modelling xxxx; **xx(x)**: 1–24

Multivariate functional additive mixed models

Alexander Volkmann[1], Almond Stöcker[1], Fabian Scheipl[2] and Sonja Greven[1]
[1]Chair of Statistics, School of Business and Economics, Humboldt-Universität zu Berlin, Germany
[2]Department of Statistics, Ludwig-Maximilians-Universität München, Germany

Abstract: Multivariate functional data can be intrinsically multivariate like movement trajectories in 2D or complementary such as precipitation, temperature and wind speeds over time at a given weather station. We propose a multivariate functional additive mixed model (multiFAMM) and show its application to both data situations using examples from sports science (movement trajectories of snooker players) and phonetic science (acoustic signals and articulation of consonants). The approach includes linear and nonlinear covariate effects and models the dependency structure between the dimensions of the responses using multivariate functional principal component analysis. Multivariate functional random intercepts capture both the auto-correlation within a given function and cross-correlations between the multivariate functional dimensions. They also allow us to model between-function correlations as induced by, for example, repeated measurements or crossed study designs. Modelling the dependency structure between the dimensions can generate additional insight into the properties of the multivariate functional process, improves the estimation of random effects, and yields corrected confidence bands for covariate effects. Extensive simulation studies indicate that a multivariate modelling approach is more parsimonious than fitting independent univariate models to the data while maintaining or improving model fit.

Key words: functional additive mixed model; multivariate functional principal components; multivariate functional data; snooker trajectories; speech production

Received March 2021; **revised** July 2021; **accepted** October 2021

1 Introduction

With the technological advances seen in recent years, functional datasets are increasingly multivariate. They can be multivariate with respect to the domain of a function, its codomain, or both. Here, we focus on multivariate functions with a one-dimensional domain $\boldsymbol{f} = (f^{(1)}, ..., f^{(D)})$: $\mathcal{I} \subset \mathbb{R} \to \mathbb{R}^D$ with square-integrable components $f^{(d)} \in L^2(\mathcal{I})$, $d = 1, ..., D$. For this type of data, we can distinguish two subclasses: One has interpretable separate dimensions and can be seen as several complementary modes of a common phenomenon ('multimodal' data, cf. Uludağ and Roebroeck, 2014) as in the analysis of acoustic signals and articulation

Address for correspondence: Alexander Volkmann, Humboldt-Universität zu Berlin, School of Business and Economics, Chair of Statistics, Unter den Linden 6, Berlin 10099, Germany.
E-mail: alexander.volkmann@hu-berlin.de

 10.1177/1471082X211056158

processes in speech production in one of our data examples. The codomain then simply is the Cartesian product $\mathcal{S} = \mathcal{S}^{(1)} \times ... \times \mathcal{S}^{(D)}$ of interpretable univariate codomains $\mathcal{S}^{(d)} \subset \mathbb{R}$. The other subclass is more 'intrinsically' multivariate insofar as univariate analyses would not yield meaningful results. Consider for example two-dimensional movement trajectories as in one of our motivating applications, where the function measures Cartesian coordinates over time: for fixed trajectories, rotation or translation of the essentially arbitrary coordinate system would change the results of univariate analyses. For intrinsically multivariate functional data a multivariate approach is the natural and preferred mode of analysis, yielding interpretable results on the observation level. Even for multimodal functional data, a joint analysis may generate additional insight by incorporating the covariance structure between the dimensions. This motivates the development of statistical methods for multivariate functional data. We here propose multivariate functional additive mixed models to model potentially sparsely observed functions with flexible covariate effects and crossed or nested study designs.

Multivariate functional data have been the interest in different statistical fields such as clustering (Jacques and Preda, 2014; Park and Ahn, 2017), functional principal component analysis (FPCAs) (Chiou et al., 2014; Happ and Greven, 2018; Backenroth et al., 2018; Li et al., 2020), and registration (Carroll et al., 2021; Steyer et al., 2021). There is also ample literature on multivariate functional data regression such as graphical models (Zhu et al., 2016), reduced rank regression (Liu et al., 2020), or varying coefficient models (Zhu et al., 2012; Li et al., 2017). Yet, so far, there are only few approaches that can handle multilevel regression when the functional response is multivariate. In particular, Goldsmith and Kitago (2016) propose a hierarchical Bayesian multivariate functional regression model that can include subject level and residual random effect functions to account for correlation between and within functions. They work with bivariate functional data observed on a regular and dense grid and assume *a priori* independence between the different dimensions of the subject-specific random effects. Thus, they model the correlation between the dimensions only in the residual function. As our approach explicitly models the dependencies between dimensions for multiple functional random effects and also handles data observed on sparse and irregular grids on more than two dimensions, the model proposed by Goldsmith and Kitago (2016) can be seen as a special case of our more general model class.

Alternatively, Zhu et al. (2017) use a two-stage transformation with basis functions for the multivariate functional mixed model. This allows the estimation of scalar regression models for the resulting basis coefficients that are argued to be approximately independent. The proposed model is part of the so-called functional mixed model (FMM) framework (Morris, 2017). While FMMs use basis transformations of functional responses (observed on equal grids) at the start of the analysis, we propose a multivariate model in the functional additive mixed model (FAMM) framework, which uses basis representations of all (effect) functions in the model (Scheipl et al., 2015). The differences between these two functional regression frameworks have been extensively discussed before (Greven and Scheipl, 2017; Morris, 2017).

The main advantages of our multivariate regression model, also compared to Goldsmith and Kitago (2016) and Zhu et al. (2017), are that it is readily available for sparse and irregular functional data and that it allows to include multiple nested or crossed random processes, both of which are required in our data examples. Another important contribution is that our approach directly models the multivariate covariance structure of all random effects included in the model using multivariate functional principal components (FPCs) and thus implicitly models the covariances between the dimensions. This makes the model representation more parsimonious, avoids assumptions difficult to verify, and allows further interpretation of the random effect processes, such as their relative importance and their dominating modes. As part of the FAMM framework, our model provides a vast toolkit of modelling options for covariate and random effects, of estimation and inference (Wood, 2017). The proposed multivariate functional additive mixed model (multiFAMM) extends the FAMM framework combining ideas from multilevel modelling (Cederbaum et al., 2016) and multivariate functional data (Happ and Greven, 2018) to account for sparse and irregular functional data and different study designs.

We illustrate the multiFAMM on two motivating examples. The first (intrinsically multivariate) data stem from a study on the effect of a training programme for snooker players with a nested study design (shots within sessions within players) (Enghofer, 2014). The movement trajectories of a player's elbow, hand, and shoulder during a snooker shot are recorded on camera, yielding six-dimensional multivariate functional data (see Figure 1). In the second data example, we analyse multimodal data from a speech production study with a crossed study design (speakers crossed with words) (Pouplier and Hoole, 2016) on so-called 'assimilation' of consonants. The two measured modes (acoustic and articulatory, see Figure 3) are expected to be closely related but joint analyses have not yet incorporated the functional nature of the data.

These two examples motivate the development of a regression model for sparse and irregularly sampled multivariate functional data that can incorporate crossed or nested functional random effects as required by the study design in addition to flexible covariate effects. The proposed approach is implemented in `R` (R Core Team, 2020) in package `multifamm` (Volkmann, 2021). The article is structured as follows: Section 2 specifies the multiFAMM and Section 3 its estimation process. Section 4 presents the application of the multiFAMM to the data examples and Section 5 shows the estimation performance of our proposed approach in simulations. Section 6 closes with a discussion and outlook.

2 Multivariate functional additive mixed model

2.1 General model

Let $\boldsymbol{y}_i^*(t) = (y_i^{*(1)}(t), ..., y_i^{*(D)}(t))^\top$ be the multivariate functional response of unit $i = 1, ..., N$ over $t \in \mathcal{I}$, consisting of dimensions $d = 1, ..., D$. Without loss of generality, we assume a common one-dimensional interval domain $\mathcal{I} = [0, 1]$ for all

dimensions, and square-integrable $y_i^{*(d)} \in L^2(\mathcal{I})$. Define $L_D^2(\mathcal{I}) := L^2(\mathcal{I}) \times ... \times L^2(\mathcal{I})$ so that $\boldsymbol{y}_i^* \in L_D^2(\mathcal{I})$. The underlying smooth function $\boldsymbol{y}_i^*$, however, is only evaluated at (potentially sparse or dimension specific) points $\boldsymbol{y}_{it}^* = (y_{it}^{*(1)}, ..., y_{it}^{*(D)})^\top$ and the evaluation is subject to white noise, that is, $\boldsymbol{y}_{it} = \boldsymbol{y}_{it}^* + \boldsymbol{\epsilon}_{it}$. The residual term $\boldsymbol{\epsilon}_{it}$ reflects additional uncorrelated white noise measurement error, following a D-dimensional multivariate normal distribution $\mathcal{N}_D$ with zero-mean and diagonal covariance matrix $\tilde{\boldsymbol{\Sigma}} = \text{diag}(\sigma_1^2, ..., \sigma_D^2)$ with dimension-specific variances σ_d^2. We construct a multivariate functional mixed model as

$$\begin{aligned} \boldsymbol{y}_{it} &= \boldsymbol{y}_i^*(t) + \boldsymbol{\epsilon}_{it} = \boldsymbol{\mu}(\boldsymbol{x}_i, t) + \boldsymbol{U}(t)\boldsymbol{z}_i + \boldsymbol{\epsilon}_{it} \\ &= \boldsymbol{\mu}(\boldsymbol{x}_i, t) + \sum_{j=1}^{q} \boldsymbol{U}_j(t)\boldsymbol{z}_{ij} + \boldsymbol{E}_i(t) + \boldsymbol{\epsilon}_{it}, \quad t \in \mathcal{I}, \end{aligned} \tag{2.1}$$

where

$$\begin{aligned} \boldsymbol{U}_j(t) &= (\boldsymbol{U}_{j1}(t), ..., \boldsymbol{U}_{jV_{U_j}}(t)); j = 1, ..., q, \\ \boldsymbol{U}_{jv}(t) &\overset{\text{ind.c.}}{\sim} MGP\left(\boldsymbol{0}, K_{U_j}\right); v = 1, ..., V_{U_j}; \forall j = 1, ..., q, \\ \boldsymbol{E}_i(t) &\overset{\text{ind.c.}}{\sim} MGP\left(\boldsymbol{0}, K_E\right); i = 1, ..., N, \text{ and} \\ \boldsymbol{\epsilon}_{it} &\overset{\text{i.i.d.}}{\sim} \mathcal{N}_D\left(\boldsymbol{0}, \tilde{\boldsymbol{\Sigma}} = \text{diag}(\sigma_1^2, ..., \sigma_D^2)\right); i = 1, ..., N; \quad t \in \mathcal{I}. \end{aligned}$$

We assume an additive predictor $\boldsymbol{\mu}(\boldsymbol{x}_i, \cdot) = \sum_{l=1}^{p} \boldsymbol{f}_l(\boldsymbol{x}_i, \cdot)$ of fixed effects, which consists of partial predictors $\boldsymbol{f}_l(\boldsymbol{x}_i, \cdot) = (f_l^{(1)}(\boldsymbol{x}_i, \cdot), ..., f_l^{(D)}(\boldsymbol{x}_i, \cdot))^\top \in L_D^2(\mathcal{I}),\ l = 1, ..., p$, that are multivariate functions depending on a subset of the vector of scalar covariates $\boldsymbol{x}_i$. This allows to include linear or smooth covariate effects as well as interaction effects between multiple covariates as in the univariate FAMM (Scheipl et al., 2015). Partial predictors may also depend on dimension-specific subsets of covariates.

For random effects $\boldsymbol{U}$, we focus on model scenarios with q independent multivariate functional random intercepts for crossed and/or nested designs. For group level $v = 1, \ldots, V_{U_j}$ within grouping layer $j = 1, \ldots, q$, these take the value $\boldsymbol{U}_{jv} \in L_D^2(\mathcal{I})$. For each layer, the $\boldsymbol{U}_{j1}, ..., \boldsymbol{U}_{jV_{U_j}}$ present independent copies of a multivariate smooth zero-mean Gaussian random process. Analogously to scalar linear mixed models, the $\boldsymbol{U}_{jv}$ model correlations between different response functions $\boldsymbol{y}_i^*$ within the same group as well as variation across groups. By arranging them in a $(D \times V_{U_j})$ matrix $\boldsymbol{U}_j(t)$ per t, the jth random intercept can be expressed in the common mixed model notation in (2.1) using appropriate group indicators $\boldsymbol{z}_{ij} = (z_{ij1}, \ldots, z_{ijV_{U_j}})^\top$ for the respective design.

Although technically a curve-specific functional random intercept, we distinguish the smooth residuals $\boldsymbol{E}_i \in L_D^2(\mathcal{I})$ in the notation, as they model correlation within

rather than between response functions. We write $\boldsymbol{E}_\nu \in L^2_D(\mathcal{I}), \nu = 1, ..., V_E$ with $V_E = N$. The $\boldsymbol{E}_i$ capture smooth deviations from the group-specific mean $\boldsymbol{\mu}(\boldsymbol{x}_i, \cdot) + \sum_{j=1}^{q} \boldsymbol{U}_j(\cdot)\boldsymbol{z}_{ij}$.

For a more compact representation, we can arrange all $\boldsymbol{U}_j(t)$ and $\boldsymbol{E}_i(t)$ together in a $(D \times (\sum_{j=1}^{q} V_{U_j} + N))$ matrix $\boldsymbol{U}(t)$ per t, and the group indicators for all layers in a corresponding vector $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}^\top, \ldots, \boldsymbol{z}_{iq}^\top, \boldsymbol{e}_i^\top)^\top$ with $\boldsymbol{e}_i$ the i-th unit vector. The resulting model term $\boldsymbol{U}(t)\boldsymbol{z}_i$ then comprises all smooth random functions, accounting for all correlation between/within response functions $\boldsymbol{y}_i^*$ given the covariates $\boldsymbol{x}_i$ as required by the respective experimental design.

$\boldsymbol{E}_i$ and $\boldsymbol{U}_{j\nu}$ are independent copies (ind. c.) of random processes having multivariate $D \times D$ covariance kernels K_E, K_{U_j}, with univariate covariance surfaces $K_E^{(d,e)}(t, t') = \text{Cov}\left[E_i^{(d)}(t), E_i^{(e)}(t')\right]$ and $K_{U_j}^{(d,e)}(t, t') = \text{Cov}\left[U_{j\nu}^{(d)}(t), U_{j\nu}^{(e)}(t')\right]$ reflecting the covariance between the process dimensions d and e at t and t'. We call these auto-covariances for $d = e$ and cross-covariance otherwise. The multivariate Gaussian processes are uniquely defined by their multivariate mean function, here the null function $\mathbf{0}$, and the multivariate covariance kernels K_g and we write $MGP\left(\mathbf{0}, K_g\right), g \in \{U_1, \ldots, U_q, E\}$. Note that vectorizing the matrix $\boldsymbol{U}(t)$ allows to formulate the joint distribution assumption $\text{vec}(\boldsymbol{U}(t)) \sim MGP\left(\mathbf{0}, K_U\right)$ with $K_U(t, t')$ having a block-diagonal structure repeating each $K_{U_j}(t, t')$ for V_{U_j} times and $K_E(t, t')$ for N times.

We assume that the different sources of variation $\boldsymbol{U}_j(t), j = 1, ..., q, \boldsymbol{E}_i(t)$, and $\boldsymbol{\epsilon}_{it}$ are mutually uncorrelated random processes to assure model identification. Assuming smoothness of the covariance kernel K_E further guarantees that the residual process $\boldsymbol{E}_i(t)$ can be separated from the white noise $\boldsymbol{\epsilon}_{it}$, removing the error variance from the diagonal of the smooth covariance kernel (e.g., Yao et al., 2005).

2.2 FPC representation of the random effects

Model (2.1) specifies a univariate functional linear mixed model (FLMM) as given in Cederbaum et al. (2016) for each dimension d. The main difference lies in the multivariate random processes that introduce dependencies between the dimensions. In order to avoid restrictive assumptions about the structure of these multivariate covariance operators, which would typically be very difficult to elicit *a priori* or verify *ex post*, we estimate them directly from the data. The main difficulty then becomes computationally efficient estimation, which is already costly in the univariate case. Especially for higher dimensional multivariate functional data, accounting for the cross-covariances can become a complex task, which we tackle with multivariate functional principal component analysis (MFPCA).

Given the covariance operators (see Section 3), we represent the multivariate random effects in Model (2.1) using truncated multivariate Karhunen-Loève (KL)

expansions

$$\begin{aligned} \boldsymbol{U}_{jv}(t) &\approx \sum_{m=1}^{M_{U_j}} \rho_{U_j vm} \boldsymbol{\psi}_{U_j m}(t), \; j = 1, ..., q; \; v = 1, ..., V_{U_j}, \\ \boldsymbol{E}_v(t) &\approx \sum_{m=1}^{M_E} \rho_{Evm} \boldsymbol{\psi}_{Em}(t), \; v = 1, ..., N, \end{aligned} \tag{2.2}$$

where the orthonormal multivariate eigenfunctions $\boldsymbol{\psi}_{gm} = (\psi_{gm}^{(1)}, ..., \psi_{gm}^{(D)})^\top \in L_D^2(\mathcal{I})$, $m = 1, ..., M_g$, $g \in \{U_1, ..., U_q, E\}$ of the corresponding covariance operators with truncation order M_g are used as basis functions and the random scores $\rho_{gvm} \sim N(0, \nu_{gm})$ are independent and identically distributed (i.i.d.) with ordered eigenvalues ν_{gm} of the corresponding covariance operator. Note that the assumption of Gaussianity for the random processes can be relaxed. For non-Gaussian random processes, the KL expansion still gives uncorrelated (but non-normal) scores and estimation based on a penalized least squares (PLS) criterion (see Section 3.2) remains reasonable.

Using KL expansions gives a parsimonious representation of the multivariate random processes that is an optimal approximation with respect to the integrated squared error (cf. Ramsay and Silverman, 2005), as well as interpretable basis functions capturing the most prominent modes of variation of the respective process. The distinct feature of this approach is that the multivariate FPCs directly account for the dependency structure of each random process across the dimensions. If, by contrast, for example, splines were used in the basis representation of the random effects, it would be necessary to explicitly model the cross-covariances of each random process in the model (cf. Li et al., 2020). Multivariate eigenfunctions, however, are designed to incorporate the dependency structure between dimensions and allow the assumption of independent (univariate) basis coefficients ρ_{gvm} via the KL theorem (see, e.g., Happ and Greven, 2018). This leads to a parsimonious multivariate basis for the random effects, where a typically small vector of scalar scores ρ_{gvm} common to all dimensions represents nearly the entire information about these D-dimensional processes.

3 Estimation

We use a two-step approach to estimate the multiFAMM and the respective multivariate covariance operators. In a first step (Section 3.1), the D-dimensional eigenfunctions $\boldsymbol{\psi}_{gm}(t)$ with their corresponding eigenvalues ν_{gm} are estimated from their univariate counterparts following Cederbaum et al. (2018) and Happ and Greven (2018). These estimates are then plugged into (2.2) and we represent the multiFAMM as part of the general FAMM framework (Section 3.2) by suitable

re-arrangement. We can view the estimated $\boldsymbol{\psi}_{gm}(t)$ simply as an empirically derived basis that parsimoniously represents the patterns in the observed data. While their estimation adds uncertainty, we are not interested in inferential statements for the variance modes and our simulations (see Section 5) suggest that the estimated eigenfunctions are reasonable approximations that work well as a basis.

3.1 Step 1: Estimation of the eigenfunction basis

3.1.1 Step 1 (i): Univariate mean estimation

In a first step, we obtain preliminary estimates of the dimension-specific means $\mu^{(d)}(\boldsymbol{x}_i, t) = \sum_{l=1}^{p} f_l^{(d)}(\boldsymbol{x}_{il}, t)$ using univariate FAMMs. We model

$$y_{it}^{(d)} = \mu^{(d)}(\boldsymbol{x}_i, t) + \epsilon_{it}^{(d)};\ d = 1, \ldots, D \tag{3.1}$$

independently for all d with i.i.d. Gaussian random variables $\epsilon_{it}^{(d)}$. The estimation of $\mu^{(d)}(\boldsymbol{x}_i, t)$ proceeds analogously to the estimation of the multiFAMM described in Section 3.2. It is based on the evaluation points of the $y_i^{*(d)}(t)$, whose locations on the interval $\mathcal{I}$ can vary across dimensions. Model (3.1) thus accommodates sparse and irregular multivariate functional data and implies a working independence assumption across scalar observations within and across functions.

3.1.2 Step 1 (ii): Univariate covariance estimation

This preliminary mean function is used to centre the data $\tilde{y}_{it}^{(d)} = y_{it}^{(d)} - \hat{\mu}^{(d)}(\boldsymbol{x}_i, t)$ in order to obtain noisy evaluations of the detrended functions $\tilde{y}_i^{*(d)}(t) = y_i^{*(d)}(t) - \mu^{(d)}(\boldsymbol{x}_i, t)$ for covariance estimation. Cederbaum et al. (2016) already find that for this purpose, the working independence assumption within functions across evaluation points in (3.1) gives reasonable results. The expectation of the crossproducts of the centred functions then coincides with the auto-covariance, that is, $\mathbb{E}\left(\tilde{y}_{it}^{(d)}\tilde{y}_{i't'}^{(d)}\right) \approx \text{Cov}\left[y_{it}^{(d)}, y_{i't'}^{(d)}\right]$. For the independent random components specified in Model (2.1), this overall covariance decomposes additively into contributions from each random process as

$$\mathbb{E}\left(\tilde{y}_{it}^{(d)}\tilde{y}_{i't'}^{(d)}\right) \approx \sum_{j=1}^{q} K_{U_j}^{(d,d)}(t, t')\delta_{v_j v_j'} + \left(K_E^{(d,d)}(t, t') + \sigma_d^2\delta_{tt'}\right)\delta_{ii'}, \tag{3.2}$$

using indicators $\delta_{xx'}$ that equal one for $x = x'$ and zero otherwise. The indicator $\delta_{v_j v_j'}$ thus identifies if the curves in the crossproduct belong to the same group v_j of the jth layer. Using t, t', and the indicators $\delta_{v_j v_j'}$, $\delta_{tt'}$, $\delta_{ii'}$ as covariates and the crossproducts of the centred data as responses, we can estimate the auto-covariances $K_{U_1}^{(d,d)}, \ldots, K_{U_q}^{(d,d)}$,

and $K_E^{(d,d)}$ of the random processes using symmetric additive covariance smoothing (Cederbaum et al., 2018). This extends the univariate approach proposed by Cederbaum et al. (2016). In particular, we also allow a nested random effects structure as required for the snooker training application in Section 4.1 by specifying the indicator of the nested effect as the product of subject-and-session indicators. Note that estimating (3.2) also yields estimates of the dimension-specific error variances σ_d^2 as a byproduct.

3.1.3 Step 1 (iii): Univariate eigenfunction estimation

Based on the covariance kernel estimates, we apply separate univariate FPCAs for each random process by conducting an eigendecomposition of the respective linear integral operator. Practically, each estimated process- and dimension-specific auto-covariance is re-evaluated on a dense grid so that a univariate functional principal component analysis (FPCA) can be conducted. Alternatively, Reiss and Xu (2020) provide an explicit spline representation of the estimated eigenfunctions. Eigenfunctions with non-positive eigenvalues are removed to ensure positive definiteness, and further regularization by truncation based on the proportion of variance explained is possible (see, e.g., Di et al., 2009; Peng and Paul, 2009; Cederbaum et al., 2016). However, we suggest to keep all univariate FPCs with positive eigenvalues for the computation of the MFPCA in order to preserve all important modes of variation and cross-correlation in the data.

3.1.4 Step 1 (iv): Multivariate eigenfunction estimation

The estimated univariate eigenfunctions and scores are then used to conduct an MFPCA for each of the g multivariate random processes separately. The MFPCA exploits correlations between univariate FPC scores across dimensions to reduce the number of basis functions needed to sufficiently represent the random processes. We base the MFPCA on the following definition of a (weighted) scalar product

$$\langle\langle \boldsymbol{f}, \boldsymbol{g} \rangle\rangle := \sum_{d=1}^{D} w_d \int_{\mathcal{I}} f^{(d)}(t) g^{(d)}(t) dt, \quad \boldsymbol{f}, \boldsymbol{g} \in L_D^2(\mathcal{I}), \tag{3.3}$$

for the response space with positive weights w_d, $d = 1, ..., D$ and the induced norm denoted by $|||\cdot|||$. The corresponding covariance operators $\Gamma_g : L_D^2(\mathcal{I}) \to L_D^2(\mathcal{I})$ of the multivariate random processes $\boldsymbol{U}_{jv}$ and $\boldsymbol{E}_v$ are then given by $(\Gamma_g \boldsymbol{f})(t) = \langle\langle \boldsymbol{f}, K_g(t, \cdot)\rangle\rangle$, $g \in \{U_1, ..., U_q, E\}$. The standard choice of weights in our applications is $w_1 = ... = w_D = 1$ (unweighted scalar product) but other choices are possible. Consider for example a scenario where dimensions are observed with different amounts of measurement error. If variation in dimensions with a large proportion of measurement error is to be downweighted, we propose to use $w_d = \frac{1}{\hat{\sigma}_d^2}$ with the dimension-specific measurement error variance estimates $\hat{\sigma}_d^2$ obtained from (3.2).

Happ and Greven (2018) show that estimates of the multivariate eigenvalues ν_{gm} of Γ_g can be obtained from an eigenanalysis of a covariance matrix of the univariate random scores. The corresponding multivariate eigenfunctions $\boldsymbol{\psi}_{gm}$ can be obtained as linear combinations of the univariate eigenfunctions with the weights given by the resulting eigenvectors. The estimates $\hat{\boldsymbol{\psi}}_{gm}$ are then substituted for the basis functions of the truncated multivariate KL expansions of the random effects $\boldsymbol{U}_{jv}$ and $\boldsymbol{E}_v$ in (2.2). Note that for each random process g, the maximum number of FPCs is given by the total number of univariate eigenfunctions included in the estimation process of the MFPCA of g. To achieve further regularization and analogously to Cederbaum et al. (2016), we propose to choose truncation orders M_g for each KL expansion of the multivariate random processes using a prespecified proportion of explained variation.

3.1.5 Step 1 (v): Multivariate truncation order

We offer two different approaches for the choice of truncation orders M_g based on different variance decompositions (derivation in Supplementary Material A):

$$\mathbb{E}\left(|||\boldsymbol{y}_i - \boldsymbol{\mu}(\boldsymbol{x}_i)|||^2\right) = \sum_{d=1}^{D} w_d \int_{\mathcal{I}} \mathrm{Var}\big(y_i^{(d)}(t)\big)dt = \sum_g \sum_{m=1}^{\infty} \nu_{gm} + \sum_{d=1}^{D} w_d \sigma_d^2 |\mathcal{I}|, \tag{3.4}$$

$$\text{and} \quad \int_{\mathcal{I}} \mathrm{Var}\big(y_i^{(d)}(t)\big)dt = \sum_g \sum_{m=1}^{\infty} \nu_{gm} ||\psi_{gm}^{(d)}||^2 + \sigma_d^2 |\mathcal{I}| \tag{3.5}$$

with $|\mathcal{I}|$ the length of the interval $\mathcal{I}$ (here equal to one) and $||\cdot||$ the L^2 norm. Multivariate variance decomposition (3.4) uses the (weighted) sum of total variation in the data across dimensions. We select the FPCs with highest associated eigenvalues ν_{gm} over all random processes g until their sum reaches a prespecified proportion (e.g., 0.95) of the total variation, thus approximating the infinite sums in (3.4) with M_g summands. For the approach based on the univariate variance (3.5), we require M_g to be the smallest truncation order for which at least a prespecified proportion of variance is explained on every dimension d. This second choice of M_g might be preferable in situations where the variation is considerably different (in amount or structure) across dimensions, whereas the first approach gives a more parsimonious representation of the random effects. Note that both approaches can lead to a simplification of the multiFAMM if $M_g = 0$ is chosen for some g. The simulation results of Section 5 suggest that increasing the number of FPCs improves model accuracy which is why sensitivity analyses with regard to the truncation order are recommended.

3.2 Step 2: Estimation of the multiFAMM

In the following, we discuss estimating the multiFAMM given the estimated multivariate FPCs. We base the proposed model on the general FAMM framework of Scheipl et al. (2015), which models functional responses using basis representations. To make the extension of the FAMM framework to multivariate functional data more apparent, the multivariate response vectors and the respective model matrices are stacked over dimensions, so that every block has the structure of a univariate FAMM over all observations i. This gives concatenated basis functions with discontinuities between the dimensions. The fixed effects are modelled analogously to the univariate case by interacting all covariate effects with a dimension indicator. The random effects are based on the parsimonious, concatenated multivariate FPC basis.

3.2.1 Matrix representation

For notational simplicity we assume that the functions are evaluated on a fixed grid of time points $\boldsymbol{t} = (\boldsymbol{t}^{(1)\top}, ..., \boldsymbol{t}^{(D)\top})^\top$ with $\boldsymbol{t}^{(d)\top} = (\boldsymbol{t}_1^{(d)}, ..., \boldsymbol{t}_N^{(d)\top})$ and identical $\boldsymbol{t}_i^{(d)} \equiv (t_1, ..., t_T)^\top$ over all N individuals and D dimensions. However, our framework allows for sparse functional data using different grids per dimension and per observed function as in the two applications (Section 4). Correspondingly, $\boldsymbol{y} = (\boldsymbol{y}^{(1)\top}, ..., \boldsymbol{y}^{(D)\top})^\top$ is the DNT-vector of stacked evaluation points with $\boldsymbol{y}^{(d)} = (\boldsymbol{y}_1^{(d)\top}, ..., \boldsymbol{y}_N^{(d)\top})^\top$ and $\boldsymbol{y}_i^{(d)} = (y_{i1}^{(d)}, ..., y_{iT}^{(d)})^\top$. Model (2.1) on this grid can be written as

$$\boldsymbol{y} = \boldsymbol{\Phi\theta} + \boldsymbol{\Psi\rho} + \boldsymbol{\epsilon} \tag{3.6}$$

with $\boldsymbol{\Phi}, \boldsymbol{\Psi}$ the model matrices for the fixed and random effects, respectively, $\boldsymbol{\theta}, \boldsymbol{\rho}$ the vectors of coefficients and random effect scores to be estimated, and $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}^{(1)\top}, ..., \boldsymbol{\epsilon}^{(D)\top})^\top$, $\boldsymbol{\epsilon}^{(d)} = (\epsilon_{11}^{(d)}, ..., \epsilon_{1T}^{(d)}, ..., \epsilon_{NT}^{(d)})^\top$ the vector of residuals. We have $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, ..., \sigma_D^2) \otimes \boldsymbol{I}_{NT}$, the Kronecker product denoted by $\otimes$, and the $(NT \times NT)$ identity matrix $\boldsymbol{I}_{NT}$.

We estimate $\boldsymbol{\theta}$ and $\boldsymbol{\rho}$ by minimizing the PLS criterion

$$(\boldsymbol{y} - \boldsymbol{\Phi\theta} - \boldsymbol{\Psi\rho})\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\Phi\theta} - \boldsymbol{\Psi\rho})^\top + \sum_{l=1}^{p} \boldsymbol{\theta}_l^\top \boldsymbol{P}_l(\boldsymbol{\lambda}_{xl}, \boldsymbol{\lambda}_{tl})\boldsymbol{\theta}_l + \sum_g \lambda_g \boldsymbol{\rho}_g^\top \boldsymbol{P}_g \boldsymbol{\rho}_g \tag{3.7}$$

using appropriate penalty matrices $\boldsymbol{P}_l(\boldsymbol{\lambda}_{xl}, \boldsymbol{\lambda}_{tl})$ and $\boldsymbol{P}_g$ for the fixed effects and random effects, respectively, and smoothing parameters $\boldsymbol{\lambda}_{xl} = \left(\lambda_{xl}^{(1)}, ..., \lambda_{xl}^{(D)}\right)$, $\boldsymbol{\lambda}_{tl} = \left(\lambda_{tl}^{(1)}, ..., \lambda_{tl}^{(D)}\right)$, and λ_g. The model and penalty matrices as well as the parameter vectors of (3.6) and (3.7) are discussed in detail below.

3.2.2 Modelling of fixed effects

The block-diagonal matrix $\boldsymbol{\Phi} = \text{diag}\left(\boldsymbol{\Phi}^{(1)}, ..., \boldsymbol{\Phi}^{(D)}\right)$ models the fixed effects separately on each dimension as in a FAMM (Scheipl et al., 2015). The $(DNT \times b)$ matrix $\boldsymbol{\Phi}$ consists of the design matrices $\boldsymbol{\Phi}^{(d)} = (\boldsymbol{\Phi}_1^{(d)} \,|\, ... \,|\, \boldsymbol{\Phi}_p^{(d)})$ that are constructed for the partial predictors $f_l^{(d)}(\boldsymbol{x}, \boldsymbol{t}^{(d)})$, $l = 1, ..., p$, which correspond to the NT-vectors of evaluations of the effect functions $f_l^{(d)}$. The vectors of scalar covariates $\boldsymbol{x}_i$ are repeated T times to form the matrix of covariate information $\boldsymbol{x} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_1, ..., \boldsymbol{x}_N)^\top$. We use the basis representations

$$f_l^{(d)}(\boldsymbol{x}, \boldsymbol{t}^{(d)}) \approx \boldsymbol{\Phi}_l^{(d)} \boldsymbol{\theta}_l^{(d)} = (\boldsymbol{\Phi}_{xl}^{(d)} \odot \boldsymbol{\Phi}_{tl}^{(d)}) \boldsymbol{\theta}_l^{(d)},$$

where $\boldsymbol{A} \odot \boldsymbol{B}$ denotes the row tensor product $(\boldsymbol{A} \otimes \mathbf{1}_b^\top) \cdot (\mathbf{1}_a^\top \otimes \boldsymbol{B})$ of the $(h \times a)$ matrix $\boldsymbol{A}$ and the $(h \times b)$ matrix $\boldsymbol{B}$ with element-wise multiplication $\cdot$ and $\mathbf{1}_c$ the c-vector of ones. This modelling approach combines the $(NT \times b_{xl}^{(d)})$ basis matrix $\boldsymbol{\Phi}_{xl}^{(d)}$ with the $(NT \times b_{tl}^{(d)})$ basis matrix $\boldsymbol{\Phi}_{tl}^{(d)}$. These matrices contain the evaluations of suitable marginal bases in $\boldsymbol{x}$ and $\boldsymbol{t}^{(d)}$, respectively. For a linear effect, for example, the basis matrix $\boldsymbol{\Phi}_{xl}^{(d)}$ is specified as the familiar linear model design matrix $\boldsymbol{x}$ for the linear effect $f_l^{(d)}(\boldsymbol{x}, \boldsymbol{t}^{(d)}) = \boldsymbol{x}\boldsymbol{\beta}_l^{(d)}(\boldsymbol{t}^{(d)})$ with coefficient function $\boldsymbol{\beta}_l^{(d)}(\boldsymbol{t}^{(d)})$. For a nonlinear effect $f_l^{(d)}(\boldsymbol{x}, \boldsymbol{t}^{(d)}) = g_l^{(d)}(\boldsymbol{x}, \boldsymbol{t}^{(d)})$, the basis matrix $\boldsymbol{\Phi}_{xl}^{(d)}$ contains an (e.g., B-spline) basis representation analogously to a scalar additive model. For the functional intercept, $\boldsymbol{\Phi}_{xl}^{(d)}$ is a vector of ones, and we generally use a spline basis for $\boldsymbol{\Phi}_{tl}^{(d)}$. For a complete list of possible effect specifications with examples, we refer to Scheipl et al. (2015). The tensor product basis is weighted by the $b_{xl}^{(d)} b_{tl}^{(d)}$ unknown basis coefficients in $\boldsymbol{\theta}_l^{(d)}$. Stacking the vectors $\boldsymbol{\theta}_l^{(d)}$ gives $\boldsymbol{\theta}^{(d)} = (\boldsymbol{\theta}_1^{(d)\top}, ..., \boldsymbol{\theta}_p^{(d)\top})^\top$ and finally the b-vector $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)\top}, ..., \boldsymbol{\theta}^{(D)\top})^\top$ with $b = \sum_d \sum_l b_{xl}^{(d)} b_{tl}^{(d)}$.

Choosing the number of basis functions is a well known challenge in the estimation of nonlinear or functional effects. We introduce regularization by a corresponding quadratic penalty term in (3.7). Let $\boldsymbol{\theta}_l$ contain the coefficients corresponding to the partial predictor l and order it by dimensions. The penalty $\boldsymbol{P}_l(\boldsymbol{\lambda}_{xl}, \boldsymbol{\lambda}_{tl})$ is then constructed from the penalty on the marginal basis for the covariate effect, $\boldsymbol{P}_{xl}^{(d)}$, and the penalty on the marginal basis over the functional index, $\boldsymbol{P}_{tl}^{(d)}$. Specifically, $\boldsymbol{P}_l(\boldsymbol{\lambda}_{xl}, \boldsymbol{\lambda}_{tl})$ is a block-diagonal matrix with blocks for each d corresponding to the Kronecker sums of the marginal penalty matrices $\lambda_{xl}^{(d)} \boldsymbol{P}_{xl}^{(d)} \otimes \boldsymbol{I}_{b_{tl}^{(d)}} + \lambda_{tl}^{(d)} \boldsymbol{I}_{b_{xl}^{(d)}} \otimes \boldsymbol{P}_{tl}^{(d)}$ (Wood, 2017). A standard choice for these marginal penalty matrices given a B-splines basis representation are second or third order difference penalties, thus approximately penalizing squared second or third derivatives of the respective functions (Eilers and

Marx, 1996). For unpenalized effects such as a linear effect of a scalar covariate, the corresponding $\boldsymbol{P}_{xl}^{(d)}$ is simply a matrix of zeroes.

3.2.3 Modelling of random effects

We represent the DNT-vectors $\boldsymbol{U}_j(\boldsymbol{t}) = (\boldsymbol{U}_j(\boldsymbol{t}^{(1)})^\top, ..., \boldsymbol{U}_j(\boldsymbol{t}^{(D)})^\top)^\top$, $\boldsymbol{E}(\boldsymbol{t}) = (\boldsymbol{E}(\boldsymbol{t}^{(1)})^\top, ..., \boldsymbol{E}(\boldsymbol{t}^{(D)})^\top)^\top$ with $\boldsymbol{U}_j(\boldsymbol{t}^{(d)})$, $\boldsymbol{E}(\boldsymbol{t}^{(d)})$ containing the evaluations of the univariate random effects for the corresponding groups and time points using the basis approximations

$$\boldsymbol{U}_j(\boldsymbol{t}) \approx \boldsymbol{\Psi}_{U_j}\boldsymbol{\rho}_{U_j} = (\boldsymbol{\delta}_{U_j} \odot \tilde{\boldsymbol{\Psi}}_{U_j})\boldsymbol{\rho}_{U_j}, \quad \boldsymbol{E}(\boldsymbol{t}) \approx \boldsymbol{\Psi}_E\boldsymbol{\rho}_E = (\boldsymbol{\delta}_E \odot \tilde{\boldsymbol{\Psi}}_E)\boldsymbol{\rho}_E.$$

The vth column in the $(DNT \times V_g)$, $g \in \{U_1, ..., U_q, E\}$ indicator matrix $\boldsymbol{\delta}_g$ indicates whether a given row is from the vth group of the corresponding grouping layer. Thus, the rows of the indicator matrix $\boldsymbol{\delta}_g$ contain repetitions of the group indicators $\boldsymbol{z}_{ij}^\top$ and $\boldsymbol{e}_i^\top$ in model (2.1). For the smooth residual, $\boldsymbol{\delta}_E$ simplifies to $\boldsymbol{1}_D \otimes (\boldsymbol{I}_N \otimes \boldsymbol{1}_T)$. The $(DNT \times M_g)$ matrix $\tilde{\boldsymbol{\Psi}}_g = (\tilde{\boldsymbol{\Psi}}_g^{(1)\top}|...|\tilde{\boldsymbol{\Psi}}_g^{(D)\top})^\top$ comprises the evaluations of the M_g multivariate eigenfunctions $\psi_{gm}^{(d)}(t)$ on dimensions $d = 1, ..., D$ for the NT time points contained in the $(NT \times M_g)$ matrix $\tilde{\boldsymbol{\Psi}}_g^{(d)}$. The M_gV_g vector $\boldsymbol{\rho}_g = (\boldsymbol{\rho}_{g1}^\top, ..., \boldsymbol{\rho}_{gV_g}^\top)^\top$ with $\boldsymbol{\rho}_{gv} = (\rho_{gv1}, ..., \rho_{gvM_g})^\top$ stacks all the unknown random scores for the functional random effect g. The $(DNT \times \sum_g M_gV_g)$ model matrix $\boldsymbol{\Psi} = (\boldsymbol{\Psi}_{U_1}|...|\boldsymbol{\Psi}_{U_q}|\boldsymbol{\Psi}_E)$ then combines all random effect design matrices. Stacking the vectors of random scores in a $\sum_g M_gV_g$ vector $\boldsymbol{\rho} = (\boldsymbol{\rho}_{U_1}^\top, ..., \boldsymbol{\rho}_{U_q}^\top, \boldsymbol{\rho}_E^\top)^\top$ lets us represent all functional random intercepts in the model via $\boldsymbol{\Psi}\boldsymbol{\rho}$.

For a given functional random effect, the penalty takes the form $\boldsymbol{\rho}_g^\top\boldsymbol{P}_g\boldsymbol{\rho}_g = \boldsymbol{\rho}_g^\top(\boldsymbol{I}_{V_g} \otimes \tilde{\boldsymbol{P}}_g)\boldsymbol{\rho}_g$, where $\boldsymbol{I}_{V_g}$ corresponds to the assumed independence between the V_g different groups. The diagonal matrix $\tilde{\boldsymbol{P}}_g = \text{diag}(\nu_{g1}, ..., \nu_{gM_g})^{-1}$ contains the (estimated) eigenvalues ν_{gm} of the associated multivariate FPCs. This quadratic penalty is mathematically equivalent to a normal distribution assumption on the scores $\boldsymbol{\rho}_{gv}$ with mean zero and covariance matrix $\tilde{\boldsymbol{P}}_g^{-1}$, as implied by the KL theorem for Gaussian random processes. Note that the smoothing parameter λ_g allows for additional scaling of the covariance of the corresponding random process.

3.2.4 Estimation

We estimate the unknown smoothing parameters in $\boldsymbol{\lambda}_{xl}$, $\boldsymbol{\lambda}_{tl}$, and λ_g using fast restricted maximum likelihood (REML)-estimation (Wood, 2017). The standard identifiability constraints of FAMMs are used (Scheipl et al., 2015). In particular, in addition to the constraints for the fixed effects, the multivariate random intercepts are subject to a sum-to-zero constraint over all evaluation points as given by, for example, Goldsmith et al. (2016).

We propose a weighted regression approach to handle the heteroscedasticity assumption contained in $\mathbf{\Sigma}$. We weigh each observation proportionally to the inverse of the estimated univariate measurement error variances $\hat{\sigma}_d^2$ from the estimation of the univariate covariances (3.2). Alternatively, updated measurement error variances can be obtained from fitting separate univariate FAMMs on the dimensions using the univariate components of the multivariate FPCs basis. In practice, we found that the less computationally intensive former option gives reasonable results.

As our proposed model is part of the FAMM framework, inference for the multiFAMM is readily available based on inference for scalar additive mixed models (Wood, 2017). Note, however, that all inferential statements do not incorporate uncertainty due to the estimated multivariate eigenfunction bases, nor in the chosen smoothing parameters. The estimation process readily provides, amongst other things, standard errors for the construction of point-wise univariate confidence bands (CBs).

3.3 Implementation

We provide an implementation of the estimation of the proposed multiFAMM in the `multifamm` `R`-package (Volkmann, 2021). It is possible to include up to two functional random intercepts in $\boldsymbol{U}(t)$, which can have a nested or crossed structure, in addition to the curve-specific random intercept $\boldsymbol{E}_i(t)$. While including, for example, functional covariates is conceptually straightforward (see Scheipl et al., 2015), our implementation is restricted to scalar covariates and interactions thereof. We provide different alternatives for specifying the multivariate scalar product, the multivariate cut-off criterion, and the covariance matrix of the white noise error term. Note that the estimated univariate error variances have been proposed as weights for two separate and independent modelling decisions: as weights in the scalar product of the MFPCA and as regression weights under heteroscedasticity across dimensions.

4 Applications

We illustrate the proposed multiFAMM for two different data applications corresponding to intrinsically multivariate and multimodal fuctional data. The presentation focuses on the first application with a detailed description of the multimodal data application in Supplementary Material C. We provide the data and the code to produce all analyses in the Supplementary Material (`http://www.statmod.org/smij/archive.html`).

4.1 Snooker training data

4.1.1 Data set and preprocessing

In a study by Enghofer (2014), 25 recreational snooker players split into two groups, one of which had instructions to follow a self-administered training schedule over

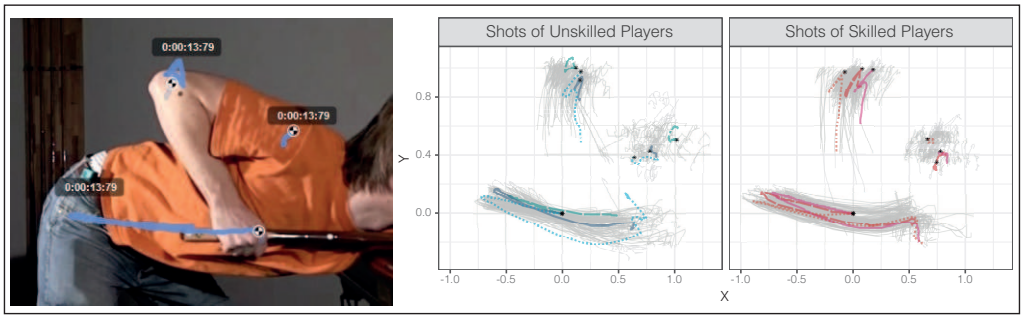

Figure 1 Screenshot of software for tracking (lines) the points of interest (circles) (left), two-dimensional trajectories of the snooker training data set (grey curves, right). For both groups of skilled and unskilled participants, three randomly selected observations are highlighted and every line type corresponds to one multivariate observation, that is, one observation consists of three trajectories: elbow (top), shoulder (right) and hand (bottom). The start of the exemplary trajectories are marked with a black asterisk with the hand trajectory centred at the origin

the next six weeks consisting of exercises aimed at improving snooker specific muscular coordination. The second was a control group. Before and after the training period, both groups were recorded on high-speed digital camera under similar conditions to investigate the effects of the training on their snooker shot of maximal force. In each of the two recording sessions, six successful shots per participant were videotaped. The recordings were then used to manually locate points of interest (a participant's shoulder, elbow, and hand) and track them on a two-dimensional grid over the course of the video. This yields a six-dimensional functional observation per snooker shot $\boldsymbol{y}^* = (y^{*(\text{elbow.x})}, \ldots, y^{*(\text{shoulder.y})}) : \mathcal{I} = [0, 1] \rightarrow \mathbb{R}^6$, that is, a two-dimensional movement trajectory for each point of interest (see Figure 1).

In their starting position (hand centred at the origin), the snooker players are positioned centrally in front of the snooker table aiming at the cue ball. From their starting position, the players draw back the cue, then accelerate it forwards and hit the cue ball shortly after their hands enter the positive range of the horizontal x-axis. After the impulse onto the cue ball, the hand movement continues until it is stopped at a player's chest. Enghofer (2014) identify two underlying techniques that a player can apply: dynamic and fixed elbow. With a dynamic elbow, the cue can be moved in an almost straight line (piston stroke) whereas additionally fixing the elbow results in a pendular motion (pendulum stroke). In both cases, the shoulder serves as a fixed point and should be positioned close to the snooker table.

We adjust the data for differences in body height and relative speed (Steyer et al., 2021) and apply a coarsening method to reduce the number of redundant data points, thereby lowering computational demands of the analysis. Supplementary Material B provides a detailed description of the data preprocessing. As some recordings and evaluations of bivariate trajectories are missing, the final dataset contains 295 functional observations with a total of 56,910 evaluation points. These multivariate

functional data are irregular and sparse, with a median of 30 evaluation points per functional observation (minimum 8, maximum 80) for each of the six dimensions.

4.1.2 Model specification

We estimate the following model

$$\boldsymbol{y}_{\ddot{\imath}jht} = \boldsymbol{\mu}(\boldsymbol{x}_{\ddot{\imath}j}, t) + \boldsymbol{B}_i(t) + \boldsymbol{C}_{\ddot{\imath}j}(t) + \boldsymbol{E}_{\ddot{\imath}jh}(t) + \boldsymbol{\epsilon}_{\ddot{\imath}jht}, \tag{4.1}$$

with $i = 1, ..., 25$ the index for the snooker player, $j = 1, 2$ the index for the session, $h = 1, ..., H_{\ddot{\imath}j}$ the index for the typically six snooker shot repetitions in a session, and $t \in [0, 1]$ relative time. Correspondingly, $\boldsymbol{B}_i(t)$ is a subject-specific random intercept, $\boldsymbol{C}_{\ddot{\imath}j}(t)$ is a nested subject-and-session-specific random intercept, and $\boldsymbol{E}_{\ddot{\imath}jh}(t)$ is the shot-specific random intercept (smooth residual). The nested random effect $\boldsymbol{C}_{\ddot{\imath}j}(t)$ is supposed to capture the variation within players between sessions (e.g., differences due to players having a good or bad day). Different positioning of participants with respect to the recording equipment or the snooker table as well as shot to shot variation are captured by the smooth residual $\boldsymbol{E}_{\ddot{\imath}jh}(t)$. The white noise measurement error $\boldsymbol{\epsilon}_{\ddot{\imath}jht}$ is assumed to follow a zero-mean multivariate normal distribution with covariance matrix $\sigma^2\boldsymbol{I}_6$, as all six dimensions are measured with the same set-up. The additive predictor is defined as

$$\begin{aligned}\boldsymbol{\mu}(\boldsymbol{x}_{\ddot{\imath}j}, t) = \boldsymbol{f}_0(t) &+ \texttt{skill}_i \cdot \boldsymbol{f}_1(t) + \texttt{group}_i \cdot \boldsymbol{f}_2(t) + \texttt{session}_j \cdot \boldsymbol{f}_3(t) \\ &+ \texttt{group}_i \cdot \texttt{session}_j \cdot \boldsymbol{f}_4(t).\end{aligned}$$

The dummy covariates $\texttt{skill}_i$ and $\texttt{group}_i$ indicate whether player i is an advanced snooker player and belongs to the treatment group (i.e., receives the training programme), respectively. Note that the snooker players self-select into training and control group to improve compliance with the training programme, which is why we include a group effect in the model. The dummy covariate $\texttt{session}_j$ indicates whether the shot j is recorded after the training period. The effect function $\boldsymbol{f}_4(t)$ can thus be interpreted as the treatment effect of the training programme.

Cubic P-splines with first-order difference penalty, penalizing deviations from constant functions over time, with 8 basis functions are used for all effect functions in the preliminary mean estimation as well as in the final multiFAMM. For the estimation of the auto-covariances of the random processes, we use cubic P-splines with first-order difference penalty on five marginal basis functions. We use an unweighted scalar product (3.3) for the MFPCA to give equal weight to all spatial dimensions, as we can assume that the measurement error mechanism is similar across dimensions. Additionally, we find that hand, elbow, and shoulder contribute roughly the same amount of variation to the data, cf. Table 1 in Supplementary Material B.3, where we also discuss potential weighting schemes for the MFPCA. The multivariate truncation order is chosen such that 95% of the (unweighted) sum of variation (3.4) is explained.

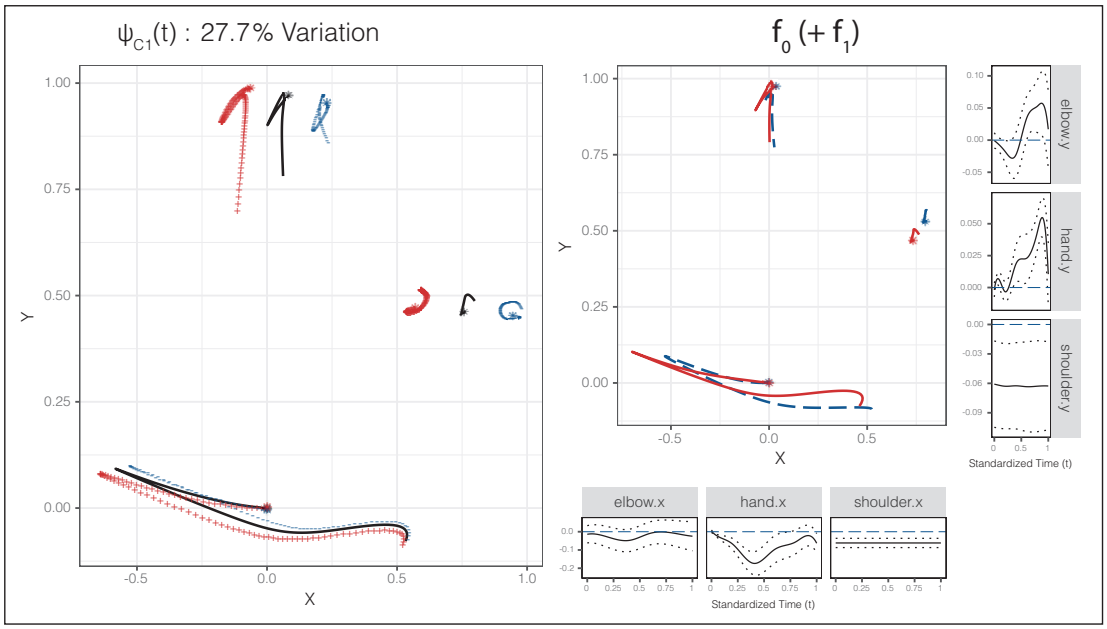

Figure 2 Dominant mode (ψ_{C1}) of the subject-and-session-specific random effect, explaining 27.7% of total variation and shown as mean trajectory (black solid) plus (+) or minus (−) $2\sqrt{\nu_{C1}}$ times the first FPC (left). An asterisk marks the start of a trajectory. Estimated covariate effect functions for skill (right). The central plot shows the effect of the coefficient function (solid) on the two-dimensional trajectories for the reference group (dashed). The marginal plots show the estimated univariate effect functions (solid) with pointwise 95% CBs (dotted) and the baseline (dashed)

4.1.3 Results

The MFPCA gives sets of five (for $\boldsymbol{C}$ and $\boldsymbol{E}$) and six (for $\boldsymbol{B}$) multivariate FPCs that explain 95% of the total variation. The estimated eigenvalues allow to quantify their relative importance. Approximately 41% of the total variation (conditional on covariates) can be attributed to the nested subject-and-session-specific random intercept $\boldsymbol{C}_{ij}(t)$, 33% to the subject-specific random intercept $\boldsymbol{B}_i(t)$, 14% to the shot-specific $\boldsymbol{E}_{ijh}(t)$, and 7% to white noise. This suggests that day to day variation within a snooker player is larger than the variation between snooker players. Note that these proportions are based on estimation step 1 (see Section 3.1).

The left plot of Figure 2 displays the first FPC for $\boldsymbol{C}$, which explains about 28% of total variation. A suitable multiple of the FPCs is added (+) to and subtracted (−) from the overall mean function (black solid line, all covariate values set to 0.5). We find that the dominant mode of the random subject-and-session-specific effect influences the relative positioning of a player's elbow, shoulder, and hand, thus suggesting a strong dependence between the dimensions. Enghofer (2014) argue from a theoretical viewpoint that the ideal starting position should place elbow and hand in a line perpendicular to the plane of the snooker table (corresponding to the x-axis). The most prominent mode of variation captures deviations from this ideal starting position found in the overall mean. The next most important

FPC ψ_{B1} of the subject-specific random effect, which explains about 15% of total variation, represents a subject's tendency towards the piston or pendulum stroke (see Supplementary Material Figure 4). This additional insight into the underlying structure of the variance components might be helpful for, for example, developing personalized training programmes.

The central plot on the right of Figure 2 compares the estimated mean movement trajectory for advanced snooker players (solid line) to that in the reference group (dashed). It suggests that more experienced players tend towards the dynamic elbow technique, generating a hand trajectory resembling a straight line (piston stroke). Uncertainties in the trajectory could be represented by pointwise ellipses, but inference is more straightforward to obtain from the univariate effect functions. The marginal plots display the estimated univariate effects with pointwise 95% confidence intervals. Even though we find only little statistical evidence for increased movement of the elbow (horizontal-left and vertical-top marginal panels), the hand and shoulder movements (horizontal centre and right, vertical centre and bottom) strongly suggest that the skill level indeed influences the mean movement trajectory of a snooker player. Further results indicate that the mean hand trajectories might slightly differ between treatment and control group at baseline as well as between sessions ($f_2(t)$ and $f_3(t)$, see Supplementary Material Figure 8). The estimated treatment effect $f_4(t)$ (Supplementary Material Figure 7), however, suggests that the training programme did not change the participants' mean movement trajectories substantially. Supplementary Material B.3 contains a detailed discussion of all model terms as well as some model diagnostics and sensitivity analyses.

4.2 Consonant assimilation data

4.2.1 Data set and model specification

Pouplier and Hoole (2016) study the assimilation of the German /s/ and /sh/ sounds such as the final consonant sounds in 'Kürbis' (English example: 'haggis') and 'Gemisch' (English example: 'dish'), respectively. The research question is how these sounds assimilate in fluent speech when combined across words such as in 'Kürbis-Schale' or 'Gemisch-Salbe', denoted as /s#sh/ and /sh#s/ with # the word boundary. The 9 native German speakers in the study repeated a set of 16 selected word combinations five times. Two different types of functional data, that is, acoustic (ACO) and electropalatographic (EPG) data, were recorded for each repetition to capture the acoustic (produced sound) and articulatory (tongue movements) aspects of assimilation over (relative) time t within the consonant combination.

Each functional index varies roughly between +1 and −1 and measures how similar the articulatory or acoustic pattern is to its reference patterns for the first (+1) and second (−1) consonant at every observed time point (Cederbaum et al., 2016). Without assimilation, the data are thus expected to shift from positive to negative values in a sinus-like form (see Figure 3). The dataset contains 707 bivariate functional observations with differently spaced grids of evaluation points per curve and dimension, with the number of evaluation points ranging from 22 to 59 with a median of 35. Note that the consonant assimilation data are unaligned as registration

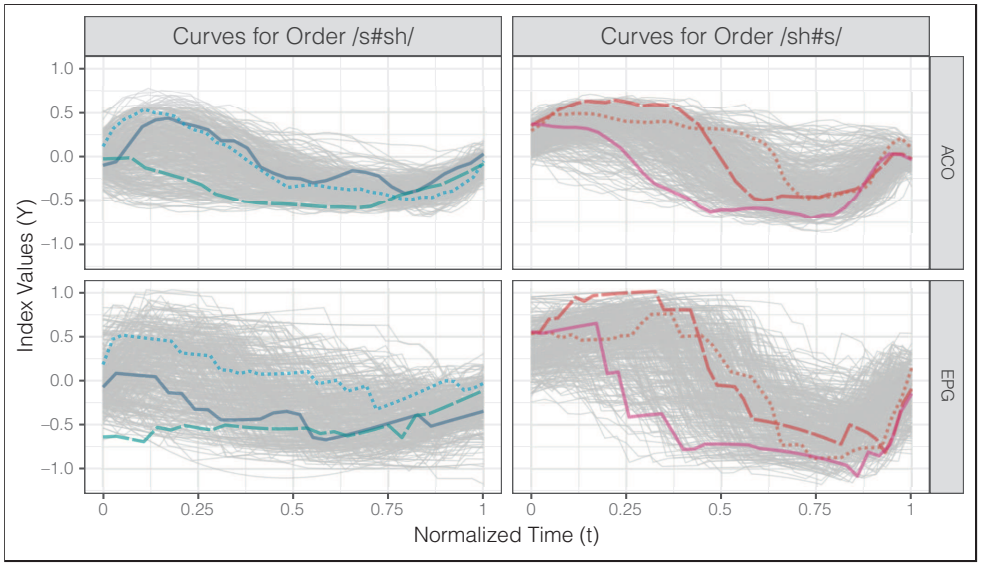

Figure 3 Index curves of the consonant assimilation dataset for both ACO and EPG data as a function of standardized time t (grey curves). For every consonant order, three randomly selected observations have been highlighted and every line type corresponds to one multivariate observation, that is, one observation consists of two index curves

of the time domain would mask transition speeds between the consonants, which are an interesting part of assimilation.

For comparability, we follow the model specification of Cederbaum et al. (2016), who analyse only the ACO dimension and ignore the second mode EPG. Our specified multivariate model is similar to (4.1) with $i = 1, ..., 9$ the speaker index, $j = 1, ..., 16$ the word combination index, $h = 1, ..., H_{ij}$ the repetition index and $t \in [0, 1]$ relative time. Note that the nested effect $\mathbf{C}_{ij}(t)$ is replaced by the crossed random effect $\mathbf{C}_j(t)$ specific to the word combinations. The additive predictor $\boldsymbol{\mu}(\boldsymbol{x}_j, t)$ now contains eight partial effects: the functional intercept plus main and interaction effects of scalar covariates describing characteristics of the word combination such as the order of the consonants /s/ and /sh/. The white noise measurement error $\boldsymbol{\epsilon}_{ijht}$ is assumed to follow a zero-mean bivariate normal distribution with diagonal covariance matrix $\text{diag}(\sigma^2_{\text{ACO}}, \sigma^2_{\text{EPG}})$. The basis and penalty specifications follow the univariate analysis in Cederbaum et al. (2016). Given different sampling mechanisms, we also compare the multiFAMM based on weighted and unweighted scalar products for the MFPCA.

4.2.2 Results

The multivariate analysis supports the findings of Cederbaum et al. (2016) that assimilation is asymmetric (different mean patterns for /s#sh/ and /sh#s/). Overall, the estimated fixed effects are similar across dimensions as well as comparable to the univariate analysis. Hence, the multivariate analysis indicates that previous results for the acoustics are consistently found also for the articulation. Compared to univariate

analyses, our approach reduces the number of FPC basis functions and thus the number of parameters in the analysis. The multiFAMM can improve the model fit and can provide smaller CBs for the ACO dimension compared to the univariate model in Cederbaum et al. (2016) due to the strong cross-correlation between the dimensions. We find similar modes of variation for the multivariate and the univariate analysis as well as across dimensions. In particular, the word combination-specific random effect $\boldsymbol{C}_j(t)$ is dropped from the model as much of the between-word variation is already explained by the included fixed effects. The definition of the scalar product has little effect on the estimated fixed effects but changes the interpretation of the FPCs. Supplementary Material C contains a more in-depth description of this application.

5 Simulations

5.1 Simulation set-up

We conduct an extensive simulation study to investigate the performance of the multiFAMM depending on different model specifications and data settings (over 20 scenarios total), and to compare it to univariate regression models as proposed by Cederbaum et al. (2016), estimated on each dimension independently. Given the broad scope of analysed model scenarios, we refer the interested reader to Supplementary Material D for a detailed report and restrict the presentation here to the main results.

We mimic our two presented data examples (Section 4) and simulate new data based on the respective multiFAMM-fit. Each scenario consists of model fits to 500 generated datasets, where we randomly draw the number and location of the evaluation points, the random scores, and the measurement errors according to different data settings. The accuracy of the estimated model components is measured by the root relative mean squared error (rrMSE) based on the unweighted multivariate norm but otherwise as defined by Cederbaum et al. (2016), see Supplementary Material D.1. The rrMSE takes on (unbounded) positive values with smaller values indicating a better fit.

5.2 Simulation results

Figure 4 compares the rrMSE values over selected modelling scenarios based on the consonant assimilation data. We generate a benchmark scenario (far left boxplots), which imitates the original data without misspecification of any model component. In particular, the number of FPCs is fixed to avoid truncation effects. Comparing this scenario to the two scenarios left and centre illustrates the importance of the number of FPCs in the accuracy of the estimation. Choosing the truncation order via the proportion of univariate variance explained (Cut-Off Uni) as in (3.5) gives models with roughly the same number of FPCs (mean $\boldsymbol{B}: 2.8, \boldsymbol{E}: 5$) as is used for the data generation ($\boldsymbol{B}: 3, \boldsymbol{E}: 5$). The cut-off criterion based on the multivariate variance (Cut-Off Mul) given by (3.4) results in more parsimonious models (mean

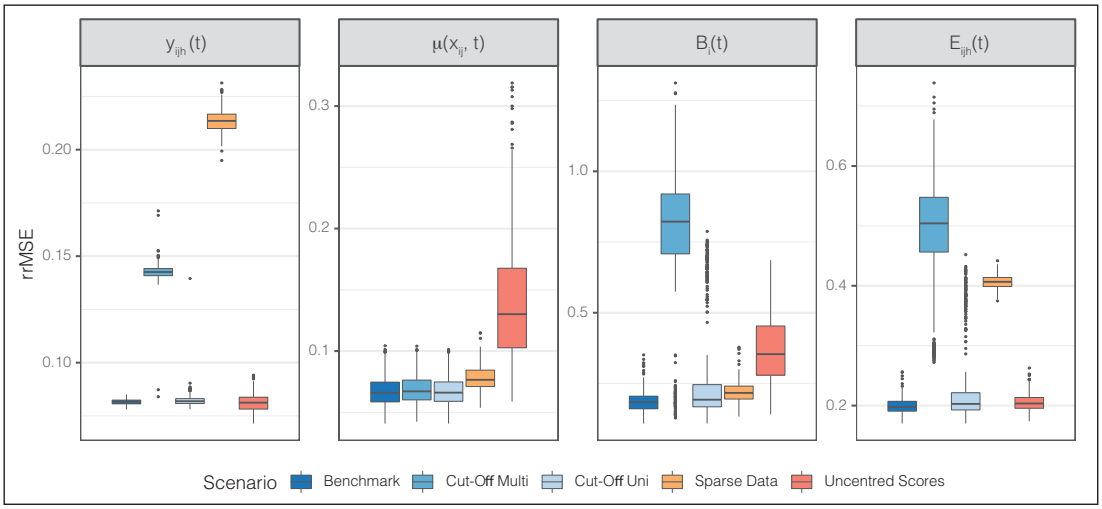

Figure 4 rrMSE values of the fitted curves $y_{ijh}(t)$, the mean $\boldsymbol{\mu}(\boldsymbol{x}_{ij}, t)$, and the random effects $B_i(t)$ and $E_{ijh}(t)$ for different modelling scenarios. The three leftmost scenarios correspond to different model specifications in the same data setting

$\boldsymbol{B}$: 2.15, $\boldsymbol{E}$: 4) and thus considerably higher rrMSE values. The increased variation in the rrMSE values can also be attributed to variability in the truncation orders (cf. Supplementary Material Figure 19), leading to a mixture distribution. Comparing the benchmark scenario to more sparsely observed functional data (ceteris paribus) suggests a lower estimation accuracy for the Sparse Data scenario (right), especially for the curve-specific random effect $\boldsymbol{E}_{ijh}(t)$ and resultingly the fitted curves $\boldsymbol{y}_{ijh}(t)$, but pooling the information across functions helps the estimation of $\boldsymbol{\mu}(\boldsymbol{x}_{ij}, t)$ and $\boldsymbol{B}_i(t)$. In particular, the estimation of the mean $\boldsymbol{\mu}(\boldsymbol{x}_{ij}, t)$ is quite robust against the increased uncertainty of these three scenarios. Only when the random scores are not centred and decorrelated as in the benchmark scenario do we find an increase in rrMSE values for the mean (Uncentred Scores, far right). This corresponds to a departure from the modelling assumptions likely to occur in practice when only few levels of a random effect are available (here for the subject-specific $\boldsymbol{B}_i(t)$). The model then has difficulties to correctly separate the intercept in $\boldsymbol{\mu}(\boldsymbol{x}_{ij}, t)$ and the random effects $\boldsymbol{B}_i(t)$. The empirical (non-zero) mean of the $\boldsymbol{B}_i(t)$ is then absorbed by the intercept in $\boldsymbol{\mu}(\boldsymbol{x}_{ij}, t)$, resulting in higher rrMSE values for both of these model terms. However, this shift does not affect the overall fit to the data $\boldsymbol{y}_{ijh}(t)$ nor the estimation of the other fixed effects (cf. Supplementary Material Figure 27). Note that the rrMSE values of the Sparse Data and Uncentred Scores scenarios are based on slightly different normalizing constants (i.e., different true data) and cannot be directly compared except for the mean.

Our simulation study thus suggests that basing the truncation orders on the proportion of explained variation on each dimension (3.5) gives parsimonious and well-fitting models. If interest lies mainly in the estimation of fixed effects, the alternative cut-off criterion based on the total variation in the data (3.4)

allows even more parsimonious models at the cost of a less accurate estimation of the random effects and overall model fit. Furthermore, the results presented in Supplementary Material D show that the mean estimation is relatively stable over different model scenarios including misspecification of the measurement error variance structure or of the multivariate scalar product, as well as in scenarios with strong heteroscedasticity across dimensions. In our benchmark scenario, the CBs cover the true effect 89 – 94% of the time but coverage can further decrease with additional uncertainty, for example, about the number of FPCs. Overall, the covariance structure such as the leading FPCs can be recovered well, also for a nested random effect such as in the snooker training application. The comparison to the univariate modelling approach suggests that the multiFAMM can improve the mean estimation but is especially beneficial for the prediction of the random effects while reducing the number of parameters to estimate. In some cases like strong heteroscedasticity, including weights in the multivariate scalar product might further improve the modelling.

6 Discussion

The proposed multivariate functional regression model is an additive mixed model, which allows to model flexible covariate effects for sparse or irregular multivariate functional data. It uses FPC based functional random effects to model complex correlations within and between functions and dimensions. An important contribution of our approach is estimating the parsimonious multivariate FPC basis from the data. This allows us to account not only for auto-covariances, but also for non-trivial cross-covariances over dimensions, which are difficult to adequately model using alternative approaches such as parametric covariance functions like the Matèrn family or penalized splines, which imply a parsimonious covariance only within but not necessarily between functions. As a FAMM-type regression model, a wide range of covariate effect types is available, also providing pointwise CBs. Our applications show that the multiFAMMs can give valuable insight into the multivariate correlation structure of the functions in addition to the mean structure.

An apparent benefit of multivariate modelling is that it allows to answer research questions simultaneously relating to different dimensions. In addition, using multivariate FPCs reduces the number of parameters compared to fitting comparable univariate models while improving the random effects estimation by incorporating the cross-covariance in the multivariate analysis. The added computational costs are small: For our multimodal application, the multivariate approach prolongs the computation time by only 5% (104 vs. 109 minutes on a 64-bit Linux platform).

We find that the average point-wise coverage of the point-wise CBs can in some cases lie considerably below the nominal value. There are two main reasons for this: One, the CBs presented here do not incorporate the uncertainty of the eigenfunction estimation nor of the smoothing parameter selection. Two, coverage issues can arise in (scalar) mixed models, if effect functions are estimated as constant when in truth they are not (e.g., Wood, 2017; Greven and Scheipl, 2016). To resolve these issues,

further research on the level of scalar mixed models might be needed. A large body of research covering CB estimation for functional data (e.g., Goldsmith et al., 2013; Choi and Reimherr, 2018; Liebl and Reimherr, 2019) suggests that the construction of CBs is an interesting and complex problem, also outside of the FAMM framework.

It would be interesting to extend the multiFAMM to more general scenarios of multivariate functional data such as observations consisting of functions with different dimensional domains, for example, functions over time and images as in Happ and Greven (2018). This would require adapting the estimation of the univariate auto-covariances for spatial arguments t, t'. Exploiting properties of dense functional data, such as the block structure of design matrices for functions observed on a grid, could help to reduce computational cost in this case. Future research could further generalize the covariance structure of the multiFAMM by allowing for additional covariate effects. In our snooker training application, for example, a treatment effect of the snooker training might show itself in the form of reduced intra-player variance (cf. Backenroth et al., 2018). Ideas from distributional regression could be incorporated to jointly model the mean trajectories and covariance structure conditional on covariates.

Acknowledgements

We thank Timon Enghofer, Phil Hoole, and Marianne Pouplier for providing access to their data and for fruitful discussions. We also thank Lisa Steyer for contributing the data registration of the snooker training data and the reviewers and editors for their helpful suggestions.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship and/or publication of this article: Sonja Greven, Almond Stöcker, and Alexander Volkmann were funded by grant GR 3793/3-1 from the German research foundation (DFG). Fabian Scheipl was funded by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A.

Supplemental material

Supplementary materials for this article are available from
http://www.statmod.org/smij/archive.html

References

Backenroth D, Goldsmith J, Harran MD, Cortes JC, Krakauer JW and Kitago T (2018) Modelling motor learning using heteroscedastic functional principal components analysis. *Journal of the American Statistical Association*, **113**, 1003–1015.

Carroll C, Müller H-G and Kneip A (2021) Cross-component registration for multivariate functional data, with application to growth curves. *Biometrics*, **77**, 839–51.

Cederbaum J, Pouplier M, Hoole P and Greven S (2016) Functional linear mixed models for irregularly or sparsely sampled data. *Statistical Modelling*, **16**, 67–88.

Cederbaum J, Scheipl F and Greven S (2018) Fast symmetric additive covariance smoothing. *Computational Statistics & Data Analysis*, **120**, 25–41.

Chiou J-M, Chen Y-T and Yang Y-F (2014) Multivariate functional principal component analysis: A normalization approach. *Statistica Sinica*, **24**, 1571–96.

Choi H and Reimherr M (2018) A geometric approach to confidence regions and bands for functional parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **80**, 239–60.

Di C-Z, Crainiceanu CM, Caffo BS and Punjabi NM (2009) Multilevel functional principal component analysis. *The Annals of Applied Statistics*, **3**, 458.

Eilers PH and Marx BD (1996) Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**, 89–102.

Enghofer T (2014) *Überblick über die Sportart snooker, Entwicklung eines Muskeltraining und Untersuchung dessen Einflusses auf die Stoßtechnik* [Overview of snooker as a sport, development of a muscular training programme, and analysis of this training programme's influence on the Snooker shot]. Unpublished thesis. Technische Universität München.

Goldsmith J and Kitago T (2016) Assessing systematic effects of stroke on motor control by using hierarchical function-on-scalar regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **65**, 215–36.

Goldsmith J, Greven S and Crainiceanu C (2013) Corrected confidence bands for functional data using principal components. *Biometrics*, **69**, 41–51.

Goldsmith J, Scheipl F, Huang L, Wrobel J, Gellar J, Harezlak J, McLean MW, Swihart B, Xiao L, Crainiceanu C and Reiss, PT (2016) *refund: Regression with Functional Data*. URL https://cran.r-project.org/web/packages/refund/refund.pdf (last accessed 25 October 2021).

Greven S and Scheipl F (2016) Comment. *Journal of the American Statistical Association*, **111**, 1568–1573.

Greven S and Scheipl F (2017) A general framework for functional regression modelling. *Statistical Modelling*, **17**, 1–35, 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**, 649–59.

Jacques J and Preda C (2014) Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis*, **71**, 92–106.

Li C, Xiao L and Luo S (2020) Fast covariance estimation for multivariate sparse functional data. *Stat*, **9**, e245.

Li J, Huang C, Hongtu Z and Alzheimer's Disease Neuroimaging Initiative (2017) A functional varying-coefficient single-index model for functional response data. *Journal of the American Statistical Association*, **112**, 1169–81.

Liebl D and Reimherr M (2019) Fast and fair simultaneous confidence bands for functional parameters. *arXiv preprint arXiv:1910.00131*.

Liu Y, Yan B, Merikangas K and Shou H (2020) Graph-fused multivariate regression via total variation regularization. *arXiv preprint arXiv:2001.04968*.

Morris JS (2017) Comparison and contrast of two general functional regression modelling

frameworks. *Statistical Modelling*, **17**, 59–85.

Park J and Ahn J (2017) Clustering multivariate functional data with phase variation. *Biometrics*, **73**, 324–33.

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**, 995–1015.

Pouplier M and Hoole P (2016) Articulatory and acoustic characteristics of German fricative clusters. *Phonetica*, **73**, 52–78.

R Core Team (2020) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org (last accessed 25 October 2021).

Ramsay JO and Silverman BW (2005) *Functional data analysis, 2nd edition*. Springer Science & Business Media.

Reiss PT and Xu M (2020) Tensor product splines and functional principal components. *Journal of Statistical Planning and Inference*, **208**, 1–12.

Scheipl F, Staicu A-M and Greven S (2015) Functional additive mixed models. *Journal of Computational and Graphical Statistics*, **24**, 477–501.

Steyer L, Stöcker A and Greven S (2021) Elastic analysis of irregularly or sparsely sampled curves. *arXiv preprint arXiv:2104.11039*.

Uludağ K and Roebroeck A (2014) General overview on the merits of multimodal neuroimaging data fusion. *Neuroimage*, **102**, 3–10.

Volkmann A (2021) *multifamm: Multivariate Functional Additive Mixed Models*. URL https://cran.r-project.org/web/packages/multifamm (last accessed 25 October 2021).

Wood SN (2017) *Generalized additive models: An introduction with R, 2nd edition*. Chapman and Hall/CRC Press.

Yao F, Müller H-G and Wang J-L (2005) Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, **100**, 577–90.

Zhu H, Li R and Kong L (2012) Multivariate varying coefficient model for functional responses. *Annals of Statistics*, **40**, 2634.

Zhu H, Morris JS, Wei F and Cox DD (2017) Multivariate functional response regression, with application to fluorescence spectroscopy in a cervical pre- cancer study. *Computational Statistics & Data Analysis*, **111**, 88–101.

Zhu H, Strawn N and Dunson DB (2016) Bayesian graphical models for multivariate functional data. *The Journal of Machine Learning Research*, **17**, 7157–83.

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

While Chapters 3 and 4 consider flat geometries of functional data in Hilbert spaces, we generalize FAM to model shapes of planar curves as elements of a Riemannian manifold (Kendall's shape space in infinite dimensions) in this contribution. We model the mean shape via a geodesic response function inspired by geodesic regression and fit the model with respect to the squared Riemannian distance, for which we propose a novel Riemannian L^2-Boosting algorithm. Besides considering the shape of curves modulo translation, rotation and scale, we also model the form of curves modulo translation and rotation but preserving the size to offer more flexibility in adapting to the data problem at hand. While known from literature in the shape case, we derive the parallel transport in the form space which is required for model formulation and fitting. Moreover, we establish a novel model effect visualization based on a suitable tensor-product factorization, which allows for systematic graphical interpretation also in the multidimensional functional case. The proposed methods are illustrated in a morphological study on sheep bone shapes and for inferring cell forms generated in biophysical simulations.

Contributing article:
Stöcker, A., Steyer, L., and Greven, S. (2022). Functional additive models on manifolds of planar shapes and forms. *arXiv pre-print*. Licensed under CC BY 4.0. Copyright © 2022 The Authors. DOI: 10.48550/ARXIV.2109.02624. *Tentatively accepted for publication in the Journal of Computational and Graphical Statistics.*
Supplementary material provided in Appendix B.

Declaration on personal contributions:
Major parts of this project were conducted by the author of this thesis with important and detailed advise and discussion from Sonja Greven and Lisa Steyer.

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

July 8, 2022

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 non-linear) covariate effects in the shape/form space via suitable tensor-product factorization. The usefulness of the proposed framework is illustrated in an analysis of 1) astragalus shapes of wild and domesticated sheep and 2) cell forms generated in a biophysical model, as well as 3) in a realistic simulation study with response shapes and forms motivated from a dataset on bottle outlines.

Keywords: functional regression, boosting, 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 re-scaling, or as its *form* (or *size-and-shape*) invariant under translation and rotation. This paper 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 et al., 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 et al., 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 then 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 et al., 2010), polynomial regression (Hinkle et al., 2014), smoothing splines (Kume et al., 2007), regression along geodesic paths with non-constant speed (Hong et al., 2014), or kernel regression (Davis et al., 2010) and Kriging (Pigoli et al., 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); 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. Recently, Lin et al. (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 different strategies existing to model a) response functions and b) smooth covariate effects. For a), basis expansions in spline (Brockhaus et al., 2015), functional principal component (FPC) bases (Morris and Carroll, 2006) or both (Scheipl et al., 2015) are employed 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. Other approaches effectively evaluate curves on grids or apply pre-smoothing techniques instead (e.g., Jeon and Park, 2020). For b), again penalized spline basis approaches are employed (Scheipl et al., 2015; Brockhaus et al., 2015), or local linear/polynomial (Müller and Yao, 2008; Jeon et al., 2022) or other kernel-based approaches (Jeon and Park, 2020; Jeon et al., 2021). The different approaches come with different theoretical and practical advantages, but similiarities such as regarding asymptotic behavior are also known from scalar nonparametric regression (Li and Ruppert,

2008). Advantages of the fully basis expansion based approach summarized in Greven and Scheipl (2017) include its appropriateness for sparse irregular functional data and its modular extensibility to functional mixed models (Scheipl et al., 2015; Meyer et al., 2015) and non-standard response distributions (Brockhaus et al., 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 et al. (2012); Olsen et al. (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. Volkmann 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 paper 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 et al., 2015) to Riemannian manifolds for model estimation. The three major contributions of our regression framework are: 1. We introduce additive regression with shapes/forms of planar curves and/or landmarks as response, extending FAMs to non-linear response spaces or, vice versa, extending GLM-type regression on manifolds for landmark shapes both to functional shape manifolds and to include (non-linear) additive model effects. 2. We propose a novel Riemannian L_2-Boosting algorithm for estimating regression models for this type of manifold response, and 3. 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, i.a., in control engineering (Baranyi et al., 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 ver-

satile 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 re-scaling, 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 forms and shapes

Riemannian manifolds of planar shapes (and forms) are discussed in various textbooks at different levels of generality, in finite (Dryden and Mardia, 2016; Kendall et al., 1999) 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 re-scaling 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 $y : \mathbb{R} \supset \mathcal{T} \to \mathbb{C}$, 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 $\text{Trl} = \{y \overset{\text{Trl}_\gamma}{\longmapsto} y + \gamma \mathbb{1} : \gamma \in \mathbb{C}\}$ with $\mathbb{1} \in \mathcal{Y}$ canonically given by $\mathbb{1} : t \mapsto \frac{1}{\|t \mapsto 1\|}$ the real constant function of unit norm; re-scaling $\text{Scl} = \{y \overset{\text{Scl}_\lambda}{\longmapsto} \lambda \cdot (y - \mathcal{O}_y) + \mathcal{O}_y : \lambda \in \mathbb{R}^+\}$ around

the centroid $0_y = \langle 1, y\rangle 1$ (which we consider more natural than using 0, the zero element of $\mathcal{Y}$, mostly chosen in the literature); and rotation $\mathrm{Rot} = \{y \overset{\mathrm{Rot}_u}{\longmapsto} u \cdot (y - 0_y) + 0_y : u \in \mathbb{S}^1\}$ around 0_y with $\mathbb{S}^1 = \{u \in \mathbb{C} : |u| = 1\} = \{\exp(\omega\sqrt{\text{-}1}) : \omega \in \mathbb{R}\}$ reflecting counterclockwise rotations by ω radian measure. Concatenation yields combined group actions G as direct products, such as the rigid motions $G = \mathrm{Trl} \times \mathrm{Rot} = \{\mathrm{Trl}_\gamma \circ \mathrm{Rot}_u : \gamma \in \mathbb{C}, u \in \mathbb{S}^1\} \cong \mathbb{C} \times \mathbb{S}^1$ (see Supplement S.1.1 for more details). The two real-valued component functions of y are identified with the real part $\mathrm{Re}(y) : \mathcal{T} \to \mathbb{R}$ and imaginary part $\mathrm{Im}(y) : \mathcal{T} \to \mathbb{R}$ of $y = \mathrm{Re}(y) + \mathrm{Im}(y)\sqrt{\text{-}1}$. While the complex setup is used for convenience, the real part of $\langle \cdot, \cdot \rangle$ constitutes an inner product $\mathrm{Re}(\langle y_1, y_2\rangle) = \langle \mathrm{Re}(y_1), \mathrm{Re}(y_2)\rangle + \langle \mathrm{Im}(y_1), \mathrm{Im}(y_2)\rangle$ for $y_1, y_2 \in \mathcal{Y}$ on the underlying real vector space of planar curves. Typically $\mathrm{Re}(y)$, $\mathrm{Im}(y)$ are assumed square-intregrable 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, i.e. $y^\dagger(t) = \mathrm{Re}(y)(t) - \mathrm{Im}(y)(t)\sqrt{\text{-}1}$ 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 $\mathcal{T} = \{1, \ldots, k\}$.

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 $G = \mathrm{Trl} \times \mathrm{Rot} \times \mathrm{Scl}$, $[y] = [y]_{\mathrm{Trl} \times \mathrm{Rot} \times \mathrm{Scl}} = \{\lambda u\, y + \gamma 1 : \lambda \in \mathbb{R}^+, u \in \mathbb{S}^1, \gamma \in \mathbb{C}\}$ is referred to as the *shape* of y and, for $G = \mathrm{Trl} \times \mathrm{Rot}$, $[y] = [y]_{\mathrm{Trl} \times \mathrm{Rot}} = \{uy + \gamma 1 : u \in \mathbb{S}^1, \gamma \in \mathbb{C}\}$ 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.a., be thought of as gradients $\dot{c}(0)$ of paths $c : \mathbb{R} \supset (-\delta, \delta) \to \mathcal{M}$ 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 $\widetilde{y}^G \in [y]_G \subset \mathcal{Y}$ (or short $\widetilde{y}$) 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 $[y]_{\mathrm{Trl}}$ can be one-to-one identified with its centered representative $\widetilde{y}^{\mathrm{Trl}} = y - \langle y, \mathbb{1}\rangle \mathbb{1}$ yielding an identification $\mathcal{Y}_{/\,\mathrm{Trl}} \cong \{y \in \mathcal{Y} : \langle y, \mathbb{1}\rangle = 0\}$ with a linear subspace of $\mathcal{Y}$. Hence, also $T_{[y]}\mathcal{Y}_{/\,\mathrm{Trl}} = \{y \in \mathcal{Y} : \langle y, \mathbb{1}\rangle = 0\}$. For rotation, by contrast, we can only find local identifications with Hilbert subspaces (i.e. charts) around reference points $[p]_{\mathrm{Trl}\times\mathrm{Rot}}$ we refer to as "poles". Moreover, we restrict to $y, p \in \mathcal{Y}^* = \mathcal{Y} \setminus [\mathcal{O}]_{\mathrm{Trl}}$ eliminating constant functions as degenerate special cases in the translation orbit of zero. For each $[y]_{\mathrm{Trl}\times\mathrm{Rot}}$ in an open neighborhood around $[p]_{\mathrm{Trl}\times\mathrm{Rot}}$ which can be chosen with $\langle \widetilde{y}^{\mathrm{Trl}}, \widetilde{p}^{\mathrm{Trl}}\rangle \neq 0$, y can be uniquely rotation aligned to p, yielding a one-to-one identification of the form $[y]_{\mathrm{Trl}\times\mathrm{Rot}}$ with the aligned representative given by $\widetilde{y}^{\mathrm{Trl}\times\mathrm{Rot}} = \frac{\langle \widetilde{y}^{\mathrm{Trl}}, \widetilde{p}^{\mathrm{Trl}}\rangle}{|\langle \widetilde{y}^{\mathrm{Trl}}, \widetilde{p}^{\mathrm{Trl}}\rangle|}\widetilde{y}^{\mathrm{Trl}} = \underset{y' \in [y]_{\mathrm{Trl}\times\mathrm{Rot}}}{\operatorname{argmin}} \|y' - p\|$ (compare Fig. 1). While $\widetilde{y}^{\mathrm{Trl}\times\mathrm{Rot}}$ depends on p, we omit this in the notation for simplicity. All $\widetilde{y}^{\mathrm{Trl}}$ rotation aligned to $\widetilde{p}^{\mathrm{Trl}}$ lie on the hyper-plane determined by $\mathrm{Im}\big(\langle \widetilde{y}^{\mathrm{Trl}}, \widetilde{p}^{\mathrm{Trl}}\rangle\big) = 0$ (Figure 1), which yields $T_{[p]}\mathcal{Y}^*_{/\,\mathrm{Trl}+\mathrm{Rot}} = \{y \in \mathcal{Y} : \langle y, \mathbb{1}\rangle = 0,\ \mathrm{Im}(\langle y, p\rangle) = 0\}$ with normal vectors $\zeta^{(1)} = \mathbb{1}, \zeta^{(2)} = \sqrt{\text{-}1}\,\mathbb{1}, \zeta^{(3)} = \sqrt{\text{-}1}\,p$. Note that, despite the use of complex arithmetic, $T_{[p]}\mathcal{Y}^*_{/\,\mathrm{Trl}\times\mathrm{Rot}}$ is a real vector space not closed under complex scalar multiplication. The geodesic distance of $[y]_{\mathrm{Trl}\times\mathrm{Rot}}$ to the pole $[p]_{\mathrm{Trl}\times\mathrm{Rot}}$ is given by $d([y]_{\mathrm{Trl}\times\mathrm{Rot}}, [p]_{\mathrm{Trl}\times\mathrm{Rot}}) = \|\widetilde{y}^{\mathrm{Trl}\times\mathrm{Rot}} - \widetilde{p}^{\mathrm{Trl}}\| = \underset{y' \in [y]_{\mathrm{Trl}\times\mathrm{Rot}}, p' \in [p]_{\mathrm{Trl}\times\mathrm{Rot}}}{\operatorname{argmin}} \|y' - p'\|$. 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 $[y]_{\mathrm{Trl}\times\mathrm{Rot}\times\mathrm{Scl}}$, they are identified with normalized representatives $\widetilde{y}^{\mathrm{Trl}\times\mathrm{Rot}\times\mathrm{Scl}} = \frac{\widetilde{y}^{\mathrm{Trl}\times\mathrm{Rot}}}{\|\widetilde{y}^{\mathrm{Trl}\times\mathrm{Rot}}\|}$. Motivated by the normalization, we borrow the well-known geometry of the sphere $\mathbb{S} = \{y \in \mathcal{Y} : \|y\| = 1\}$, where $T_p\mathbb{S} = \{y \in \mathcal{Y} : \mathrm{Re}(\langle y, p\rangle) = 0\}$ 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 $T_{[p]}\mathcal{Y}^*_{/\,\mathrm{Trl}\times\mathrm{Rot}\times\mathrm{Scl}} = T_{[p]}\mathcal{Y}^*_{/\,\mathrm{Trl}\times\mathrm{Rot}} \cap T_p\mathbb{S} = \{y \in \mathcal{Y} : \langle y, \mathbb{1}\rangle = 0,\ \langle y, p\rangle = 0\}$ with normal vector $\zeta^{(4)} = p$ in addition to $\zeta^{(1)}, \zeta^{(2)}, \zeta^{(3)}$ above. The geodesic distance $d([p]_{\mathrm{Trl}\times\mathrm{Rot}\times\mathrm{Scl}}, [y]_{\mathrm{Trl}\times\mathrm{Rot}\times\mathrm{Scl}}) = \arccos|\langle \widetilde{y}^{\mathrm{Trl}\times\mathrm{Rot}\times\mathrm{Scl}}, \widetilde{p}^{\mathrm{Trl}\times\mathrm{Rot}\times\mathrm{Scl}}\rangle|$ corresponds to the arc-length between the representatives. This distance is often referred to as *Procrustres distance* in statistical shape analysis.

We may now define the maps needed for the regression model formulation. Let $\widetilde{y}$ and $\widetilde{p}$ 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 $c : (-\delta, \delta) \to \mathcal{M}$ 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* $\mathrm{Exp}_q : T_q\mathcal{M} \to \mathcal{M}$ 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 $\mathrm{Exp}_q(\beta) \in \mathcal{M}$ located $d(q, \mathrm{Exp}_q(\beta)) = \|\beta\|$ apart of the pole q in the direction of β. On the form space $\mathcal{Y}_{/\,\mathrm{Trl}\times\mathrm{Rot}}$, the exponential map is simply given by $\mathrm{Exp}_{[p]_{\mathrm{Trl}\times\mathrm{Rot}}}(\beta) = \left[\widetilde{p}^{\,\mathrm{Trl}\times\mathrm{Rot}} + \beta\right]_{\mathrm{Trl}\times\mathrm{Rot}}$. On the shape space $\mathcal{Y}_{/\,\mathrm{Trl}\times\mathrm{Rot}\times\mathrm{Scl}}$, identification with exponential maps on the sphere yields $\mathrm{Exp}_{[p]_G}(\beta) = \left[\cos(\|\beta\|)\widetilde{p}^{\,G} + \sin(\|\beta\|)\frac{\beta}{\|\beta\|}\right]_G$ with $G = \mathrm{Trl}\times\mathrm{Rot}\times\mathrm{Scl}$. In an open neighborhood $\mathcal{U}$, $q \in \mathcal{U} \subset \mathcal{M}$, Exp_q is invertible yielding the $\mathrm{Log}_q : \mathcal{U} \to T_q\mathcal{M}$ map from the manifold to the tangent space at q. For forms, it is given by $\mathrm{Log}_{[p]_{\mathrm{Trl}\times\mathrm{Rot}}}([y]_{\mathrm{Trl}\times\mathrm{Rot}}) = \widetilde{y}^{\,\mathrm{Trl}\times\mathrm{Rot}} - \widetilde{p}^{\,\mathrm{Trl}\times\mathrm{Rot}}$ and, for shapes, by $\mathrm{Log}_{[p]_G}([y]_G) = d([p]_G, [y]_G)\frac{\widetilde{y}^G - \langle\widetilde{p}^G, \widetilde{y}^G\rangle\widetilde{p}^G}{\|\widetilde{y}^G - \langle\widetilde{p}^G, \widetilde{y}^G\rangle\widetilde{p}^G\|}$ with $G = \mathrm{Trl}\times\mathrm{Rot}\times\mathrm{Scl}$. Finally, $\mathrm{Transp}_{q,q'} : T_q\mathcal{M} \to T_{q'}\mathcal{M}$ parallel transports tangent vectors $\varepsilon \mapsto \varepsilon'$ isometrically along a geodesic $c(\tau)$ connecting q and $q' \in \mathcal{M}$ such that the slopes $\mathrm{Transp}_{q,q'}(\dot{c}(q))=\dot{c}(q')$ are identified and all angles are preserved. For shapes, $\mathrm{Transp}_{[y]_G,[p]_G}(\varepsilon) = \varepsilon - \langle\varepsilon, \widetilde{p}^G\rangle\frac{\widetilde{y}^G + \widetilde{p}^G}{1+\langle\widetilde{y}^G, \widetilde{p}^G\rangle}$, with $G = \mathrm{Trl}\times\mathrm{Rot}\times\mathrm{Scl}$, 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 $\mathrm{Im}(\langle\varepsilon, \widetilde{p}\rangle)$ coordinate orthogonal to the real $\widetilde{y}$-$\widetilde{p}$-plane as in the shape case, while the remainder of ε is left unchanged as in a linear space. This yields $\mathrm{Transp}_{[y]_G,[p]_G}(\varepsilon) = \varepsilon - \mathrm{Im}\big(\langle\widetilde{p}^G/\|\widetilde{p}^G\|, \varepsilon\rangle\big)\frac{\widetilde{y}^G/\|\widetilde{y}^G\| + \widetilde{p}^G/\|\widetilde{p}^G\|}{1+\langle\widetilde{y}^G/\|\widetilde{y}^G\|, \widetilde{p}^G/\|\widetilde{p}^G\|\rangle}\sqrt{\text{-1}}$, with $G = \mathrm{Trl}\times\mathrm{Rot}$, for form tangent vectors. While equivalent expressions for the parallel transport in the shape case can be found, e.g., in Dryden and Mardia (2016); Huckemann et al. (2010), a corresponding derivation for the form case is given in Supplement S.1.2 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 $y_1, \ldots, y_n \in \mathcal{Y}$ representing orbits $[y_1], \ldots, [y_n]$ with respect to group actions G. In the functional case, with the domain $\mathcal{T} = [0, 1]$, these curves are usually observed as evaluations $\mathbf{y}_i = (y_i(t_{i1}), \ldots, y_i(t_{ik_i}))^\top$ on a finite grid $t_{i1} < \cdots < t_{ik_i} \in \mathcal{T}$

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 $\langle y_i, y_i' \rangle_i = \mathbf{y}_i^\dagger \mathbf{W}_i \mathbf{y}_i'$ providing inner products on the k_i-dimensional space $\mathcal{Y}_i = \mathbb{C}^{k_i}$ of evaluations $\mathbf{y}_i, \mathbf{y}_i'$ 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) $y_1, \ldots, y_n$ along the lines of Yao et al. (2005); Stöcker et al. (2022). While this is beyond the scope of this paper, potential procedures are sketched in Supplement S.7. With the inner products given for $i = 1, \ldots, n$, the sample space naturally arises as the Riemannian product $\mathcal{Y}_{1/G}^* \times \cdots \times \mathcal{Y}_{n/G}^*$ 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 $y_i : \mathcal{T} \to \mathbb{C}$, $i = 1, \ldots, n$, belonging to a Hilbert space $\mathcal{Y}$ defined as above and potentially irregularly measured on individual grids $t_{i1} < \cdots < t_{ik_i} \in \mathcal{T}$. The response object $[Y]$ is the equivalence class of Y with respect to translation, rotation and possibly scale and the sample $[y_1], \ldots, [y_n]$ is equipped with the respective Riemannian manifold geometry introduced in the previous section. For $i = 1, \ldots, n$, 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

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

with an additive predictor $h : \mathcal{X} \to T_{[p]}\mathcal{Y}_{/G}^*$ acting in the tangent space at an "intercept" $[p] \in \mathcal{Y}_{/G}^*$. Generalizing an additive model "$Y = \mu + \epsilon = p + h(\mathbf{x}) + \epsilon$" in a linear

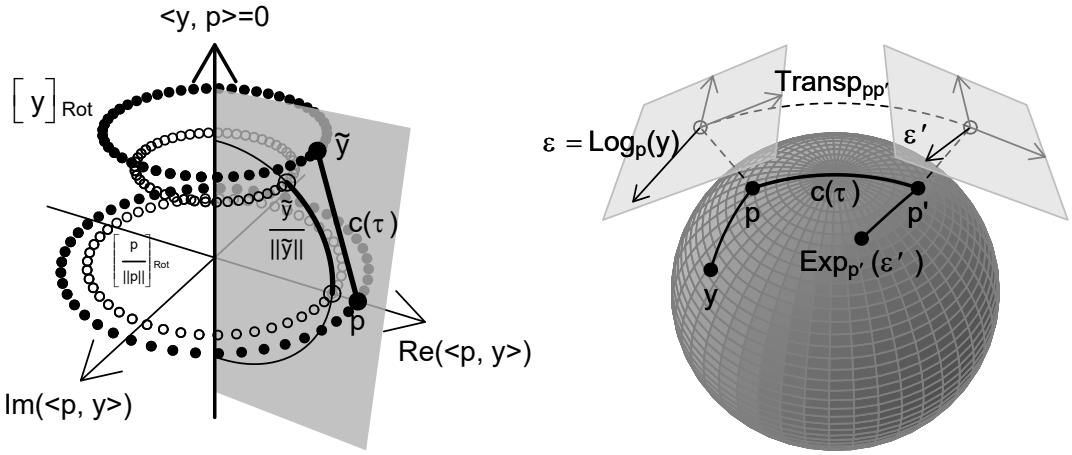

Figure 1: *Left:* Quotient space geometry: assuming p and y centered, translation invariance is not further considered in the plot; given pole representative p, we express $y = \frac{\mathrm{Re}\langle p,y\rangle}{\|p\|^2}p + \frac{\mathrm{Im}\langle p,y\rangle}{\|p\|^2}ip + (y - \frac{\langle p,y\rangle}{\|p\|^2}p) \in \mathcal{Y}$ in its coordinates in p and ip direction, subsuming all orthogonal directions in the third dimension. In this coordinate system, the rotation orbit $[y]_{\mathrm{Rot}}$ corresponds to the dotted horizontal circle, and is identified with the aligned $\widetilde{y} := \widetilde{y}^{\mathrm{Rot}}$ in the half-plane of p; $[y]_{\mathrm{Rot}\times\mathrm{Scl}}$ is identified with the unit vector $\widetilde{y}^{\mathrm{Rot}\times\mathrm{Scl}} = \frac{\widetilde{y}}{\|\widetilde{y}\|}$ projecting $\widetilde{y}$ onto the hemisphere depicted by the vertical semicircle. Form and shape distances between $[p]$ and $[y]$ correspond to the length of the geodesics $c(\tau)$ (*thick lines*) on the plane and sphere, respectively. *Right:* Geodesic line $c(\tau)$ between $p = c(0)$ and $p' = c(1)$, Log-map projecting y to $\varepsilon \in T_p\mathcal{M}$, parallel transport $\mathrm{Transp}_{pp'}$ forwarding ε to $\varepsilon' \in T_{p'}\mathcal{M}$, 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 $\dot{c}(0) \cong \dot{c}(1)$.

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 $\varepsilon_{[\mu]} = \mathrm{Log}_{[\mu]}([Y])$ in $T_{[\mu]}\mathcal{Y}^*_{/G}$ to have mean $\mathbb{E}\left(\varepsilon_{[\mu]}\right) = \mathcal{0}$, which corresponds to $\mathbb{E}\left(\varepsilon_{[\mu]}(t)\right) = 0$ for (ν-almost) all $t \in \mathcal{T}$. Here, we assume $[Y]$ is sufficiently close to $[\mu]$ with probability 1 such that $\mathrm{Log}_{[\mu]}$ is well-defined, which is the case whenever $\langle \widetilde{Y}, \widetilde{\mu} \rangle \neq 0$ for centered shape/form representatives $\widetilde{Y}$ and $\widetilde{\mu}$, an un-restrictive 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 $\epsilon = \mathrm{Transp}_{[\mu],[p]}(\varepsilon_{[\mu]})$ 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 $\mathbb{E}\left(\epsilon\right) = \mathcal{0}$ for the transported residuals ϵ. $\mathrm{Exp}_{[p]}$ 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 $[p]$. 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 $\mathbb{E}\left(h(\mathbf{X})\right) = \mathcal{0}$, 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 et al. (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 $\partial_r \in T_{[p]}\mathcal{Y}^*_{/G},\ r = 1, \ldots, m$, and m_j basis functions $b_j^{(l)} : \mathcal{X} \to \mathbb{R},\ l = 1, \ldots, m_j$, for the j-th covariate effect depending on one or more covariates. The real coefficients can be arranged as a matrix $\{\theta_j^{(r,l)}\}_{r,l} = \mathbf{\Theta}_j \in \mathbb{R}^{m \times m_j}$. Also for infinite-dimensional $T_{[p]}\mathcal{Y}^*_{/G}$ and a general

non-linear 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 $j = 1, \ldots, J$, 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 $b_0^{(l)} : \mathcal{T} \to \mathbb{R}$, $l = 1, \ldots, m_0$ 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 $\partial_r = \sum_{l=1}^{m_0} \left(z_p^{(l,r)} + z_p^{(m_0+l,r)} \sqrt{\text{-}1} \right) b_0^{(l)}$, employing the same basis for the 1- and $\sqrt{\text{-}1}$-dimension before transforming it with a basis transformation matrix $\mathbf{Z}_p = \{z_p^{(l,r)}\}_{l,r} \in \mathbb{R}^{2m_0 \times m}$ with $m < 2m_0$ implementing the linear tangent space constraints (Section 2). Practically, $\mathbf{Z}_p$ is obtained as null space basis matrix of the matrix $(\text{Re}(\mathbf{C}), \text{Im}(\mathbf{C}))$ with $\mathbf{C} = \{\langle b_0^{(l)}, \zeta^{(r)} \rangle\}_{r,l}$ (or with the empirical inner product on the product space of irregular curves instead) constructed from the normal vectors $\zeta^{(r)} \in \mathcal{Y}$, $r = 1, \ldots, 2m_0 - m$, to $T_{[p]}\mathcal{Y}^*_{/G}$. For closed curves, we additionally choose $\mathbf{Z}_p$ to enforce periodicity, i.e. $\partial_r(t) = \partial_r(t + t_0)$ for some $t_0 \in \mathbb{R}$ (compare Hofner et al., 2016).

Given the tangent space basis, we may now modularly specify the usual additive model basis functions $b_j^{(l)} : \mathcal{X} \to \mathbb{R}$, $l = 1, \ldots, m_j$, for the j-th covariate effect to obtain the full functional additive model "tool box" offered by, e.g., Brockhaus et al. (2015). A linear effect – linear in the tangent space – of the form $h_j(\mathbf{x}) = \beta z$ with a scalar (typically centered) covariate z in $\mathbf{x}$ 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, e.g., a B-spline basis (Asymptotic properties of penalized B-splines and connections to kernel estimators are discussed, e.g., by Wood et al. (2016); Li and Ruppert (2008)). For a categorical covariate effect of the form $h_j(\mathbf{x}) : \{1, \ldots, K\} \to T_{[p]}\mathcal{Y}^*_{/G}$, $\kappa \mapsto \beta_\kappa$, the basis $\mathbf{b}_j(\mathbf{x}) : \kappa \mapsto \mathbf{e}_\kappa \in \mathbb{R}^{K-1}$ maps category κ to a usual contrast vector $\mathbf{e}_\kappa$ 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 et al., 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 $\sum_{i=1}^{n} h_j(\mathbf{x}_i) = \mathcal{0}$ for observed covariates $\mathbf{x}_i$ can be specified, and similar constraints can be used to distinguish linear from non-linear effects and interactions from their marginal effects (Kneib et al., 2009). Different quadratic penalties can be specified for the coefficients $\boldsymbol{\Theta}_j$, allowing to regularize high-dimensional effect bases and to balance effects of different complexity in the model fit (cf. 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 non-linear (interaction) effects, or when effect sizes are visually small. To solve this problem, we suggest to re-write estimated TP effects $\hat{h}_j$ with estimated coefficient matrix $\widehat{\boldsymbol{\Theta}}_j$ 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 $\hat{h}_j^{(r)} : \mathcal{X} \to \mathbb{R},\ r = 1, \ldots, m_j'$, in corresponding orthonormal directions $\xi_j^{(r)} \in T_{[p]}\mathcal{Y}_{/G}^*$ with $\langle \xi_j^{(r)}, \xi_j^{(l)} \rangle = \mathbb{1}(r = l)$, i.e. 1 if $r = l$ and 0 otherwise. Assuming $\mathbb{E}\left(b_j^{(l)}(\mathbf{X})^2\right) < \infty$, $l = 1, \ldots, m_j$, for the underlying effect basis, the $\hat{h}_j^{(r)}$ are specified to achieve decreasing component variances $v_j^{(1)} \geq \cdots \geq v_j^{(m_j')} \geq 0$ given by $v_j^{(r)} = \mathbb{E}\left(\hat{h}_j^{(r)}(\mathbf{X})^2\right)$. In practice, the expectation over the covariates $\mathbf{X}$ and the inner product $\langle ., . \rangle$ are replaced by empirical analogs (compare Supplement Corollary 3). Due to orthonormality of the $\xi_j^{(r)}$, the component variances add up to the total predictor variance $\sum_{r=1}^{m_j'} v_j^{(r)} = v_j = \mathbb{E}\left(\langle \hat{h}_j(\mathbf{X}), \hat{h}_j(\mathbf{X}) \rangle\right)$. 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 $\hat{h}_*^{(r)} : \mathcal{X} \to \mathbb{R}$, such that $\mathbb{E}\left(\|\hat{h}_j(\mathbf{X}) - \sum_{r=1}^{l} \xi_*^{(r)} \hat{h}_*^{(r)}(\mathbf{X})\|^2\right) < \mathbb{E}\left(\|h_j(\mathbf{X}) - \sum_{r=1}^{l} \xi_j^{(r)} \hat{h}_j^{(r)}(\mathbf{X})\|^2\right)$, i.e. 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 $\widehat{\boldsymbol{\Theta}}_j$ and the fact that it is well-

defined is a variant of the Eckart-Young-Mirsky theorem (proof in Supplement S.2). 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

$$h(\mathbf{x}) = \sum_{r=1}^{m'} \xi^{(r)}\hat{h}^{(r)}(\mathbf{x}) = \sum_{r=1}^{m'} \xi^{(r)} \left(\hat{h}_1^{(r)}(\mathbf{x}) + \cdots + \hat{h}_J^{(r)}(\mathbf{x})\right), \quad m' = \min(\sum_j m_j, m),$$

with 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 $c(\tau) = \text{Exp}_{[p]}\left(\xi^{(1)}\tau\right)$ 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 re-fitting 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 $-\nabla_{[\mu]} \frac{1}{2} d^2([y], [\mu]) = \mathrm{Log}_{[\mu]}([y]) = \varepsilon_{[\mu]}$ (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

$$\sigma^2(h) = \mathbb{E}\left(d^2\left([Y], \mathrm{Exp}_{[p]}\left(h(\mathbf{X})\right)\right)\right)$$

with the point-wise minimizer $h^\star(\mathbf{x}) = \underset{h:\mathcal{X} \to T_{[p]}\mathcal{Y}^*_{/G}}{\mathrm{argmin}} \mathbb{E}\left(d^2\left([Y], \mathrm{Exp}_{[p]}\left(h(\mathbf{X})\right)\right) \mid \mathbf{X} = \mathbf{x}\right)$ minimizing the conditional expected squared distance. Fixing a covariate constellation $\mathbf{x} \in \mathcal{X}$, the prediction $[\mu] = \mathrm{Exp}_{[p]}\left(h^\star(\mathbf{x})\right)$ corresponds to the Fréchet mean (Karcher, 1977) of $[Y]$ conditional on $\mathbf{X} = \mathbf{x}$. In a finite-dimensional context, Pennec (2006) show that $\mathbb{E}\left(\varepsilon_{[\mu]}\right) = \mathcal{0}$ 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 $\hat{\sigma}^2(h) = \frac{1}{n}\sum_{i=1}^n d_i^2\left([y_i], \mathrm{Exp}_{[p]}\left(h(\mathbf{x}_i)\right)\right)$, 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 $h_j(\mathbf{x}) = \sum_{r,l} \theta_j^{(r,l)} b_j^{(l)}(\mathbf{x})\, \partial_r$, $\boldsymbol{\Theta}_j = \{\theta_j^{(r,l)}\}_{r,l}$, which is repeatedly fit to the transported residuals $\epsilon_1, \ldots, \epsilon_n$ by penalized least-squares (PLS) minimizing $\sum_{i=1}^n \|\epsilon_i - h_j(\mathbf{x}_i)\|_i^2 + \lambda_j \operatorname{tr}\left(\boldsymbol{\Theta}_j \mathbf{P}_j \boldsymbol{\Theta}_j^\top\right) + \lambda \operatorname{tr}\left(\boldsymbol{\Theta}^\top \mathbf{P} \boldsymbol{\Theta}\right)$. 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 $\operatorname{vec}(\boldsymbol{\Theta}_j) = (\theta_j^{(1,1)}, \ldots, \theta_j^{(m,1)}, \ldots, \theta^{(m,m_j)})^\top$, where $\otimes$ denotes the Kronecker product. The standard PLS estimator is then given by $\operatorname{vec}(\widehat{\boldsymbol{\Theta}}_j) = (\boldsymbol{\Psi}_j + \mathbf{R}_j)^{-1} \boldsymbol{\psi}_j$ with $\boldsymbol{\Psi}_j = \sum_{i=1}^n \left\{ \operatorname{Re}\Big(\langle b_j^{(l)}(\mathbf{x}_i)\partial_r, b_j^{(l')}(\mathbf{x}_i)\partial_{r'} \rangle_i \Big) \right\}_{\substack{(r,l)=(1,1),\ldots,(m,1),\ldots,(m,m_j) \\ (r',l')=(1,1),\ldots,(m,1),\ldots,(m,m_j)}} \in \mathbb{R}^{m\, m_j \times m\, m_j}$ and $\boldsymbol{\psi}_j = \sum_{i=1}^n \left\{ \operatorname{Re}\Big(\langle b_j^{(l)}(\mathbf{x}_i)\partial_r, \epsilon_i \rangle_i \Big) \right\}_{(r,l)=(1,1),\ldots,(m,1),\ldots,(m,m_j)} \in \mathbb{R}^{m\, m_j}$. In a regular design, using the functional linear array model (Brockhaus et al., 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, e.g., by re-sampling 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 $[p] = \operatorname*{argmin}_{q \in \mathcal{Y}^*} \mathbb{E}\left(d^2([Y], [q])\right)$ 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

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 $\mathbf{\Theta}_j = \mathbf{0}$

for $j = 1$ **to** J **do** `# Prepare penalized least-squares (PLS)`

  `# set up` $m\,m_j \times m\,m_j$ `matrix:`

  $\mathbf{\Psi}_j \leftarrow \sum_{i=1}^n \left\{ \mathrm{Re}\Big(\langle b_j^{(l)}(\mathbf{x}_i)\partial_r, b_j^{(l')}(\mathbf{x}_i)\partial_{r'} \rangle_i \Big) \right\}_{\substack{(r,l)=(1,1),\ldots,(m,1),\ldots,(m,m_j) \\ (r',l')=(1,1),\ldots,(m,1),\ldots,(m,m_j)}}$

end

repeat `# boosting steps`

  for $i = 1, \ldots, n$ **do** `# Compute current transported residuals`

    $[\mu_i] \leftarrow \mathrm{Exp}_{[p]}(h(\mathbf{x}_i))$

    $\varepsilon_{[\mu_i]} \leftarrow \mathrm{Log}_{[\mu_i]}([y_i])$

    $\epsilon_i \leftarrow \mathrm{Transp}_{[\mu_i],[p]}(\varepsilon_{[\mu_i]})$

  end

  for $j = 1, \ldots, J$ **do** `# PLS fit to residuals`

    `#` $m\,m_j$ `vector:`

    $\boldsymbol{\psi}_j \leftarrow \sum_{i=1}^n \left\{ \mathrm{Re}\Big(\langle b_j^{(l)}(\mathbf{x}_i)\partial_r, \epsilon_i \rangle_i \Big) \right\}_{(r,l)=(1,1),\ldots,(m,1),\ldots,(m,m_j)}$

    $\widehat{\mathbf{\Theta}}_j = \{\hat{\theta}_j^{(r,l)}\}_{r,l} \leftarrow$ `Solve(` $(\mathbf{\Psi}_j + \mathbf{R}_j)\,\mathrm{vec}(\mathbf{\Theta}) = \boldsymbol{\psi}_j$`)`

  end

  $\hat{\jmath} \leftarrow \underset{j \in \{1,\ldots,J\}}{\mathrm{argmin}} \sum_{i=1}^n \|\epsilon_i - \sum_{r,l} \hat{\theta}_j^{(r,l)} b_j^{(l)}(\mathbf{x})\partial_r\|_i^2$; `# Select base-learner`

  $\mathbf{\Theta}_{\hat{\jmath}} \leftarrow \mathbf{\Theta}_{\hat{\jmath}} + \eta\,\widehat{\mathbf{\Theta}}_{\hat{\jmath}}$; `# Update selected model coefficients`

until *Stopping criterion (e.g. minimal cross-validation error)*

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 mean of a reasonably well aligned sample $y_1, \ldots, y_n \in \mathcal{Y}$ 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 et al., 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 et al. (2019) investigate shapes of sheep astragali (ankle bones) 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 et al. (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 Supplement S.3 shows the distribution of available covariates within the populations. Each sheep astragalus shape, $i = 1, \ldots, n$, 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 $\mathbf{y}_i^{\text{c1}} \in \mathbb{C}^{14}$ and $\mathbf{y}_i^{\text{c2}} \in \mathbb{C}^{18}$ evaluated along two outline curve segments, marked on a 2D image of the bone (dorsal view). Several example configurations are displayed in Supplement Figure S1. In general, we could separately specify smooth

function bases for the outline segments y_i^{c1} and y_i^{c2}, 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 $\mathbf{y}_i = \left(\mathbf{y}_i^{\text{lm}\top}, \mathbf{y}_i^{\text{c1}\top}, \mathbf{y}_i^{\text{c2}\top}\right)^\top \in \mathbb{C}^{43} = \mathcal{Y}$, 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 $\left(\mathbf{1}_{11}^\top, \frac{3}{14}\mathbf{1}_{14}^\top, \frac{3}{18}\mathbf{1}_{18}^\top\right)^\top$ 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

$$[\boldsymbol{\mu}_i] = \text{Exp}_{[\mathbf{p}]}\left(\boldsymbol{\beta}_{\text{status}_i} + \boldsymbol{\beta}_{\text{pop}_i} + \boldsymbol{\beta}_{\text{age}_i} + \boldsymbol{\beta}_{\text{sex}_i} + \boldsymbol{\beta}_{\text{mobility}_i}\right)$$

with the pole $[\mathbf{p}] \in \mathcal{Y}^*_{/G}$ 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 $x_{ij} \mapsto \boldsymbol{\beta}_{x_{ij}} \in T_{[\mathbf{p}]}\mathcal{Y}^*_{/\,\text{Trl}\times\text{Rot}\times\text{Scl}}$. 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 et al., 2015) for saving computation time and memory, which lead to 8 seconds of initial model fit followed by 47 seconds 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 et al. (2019), we find little influence of age, sex and mobility on the astragalus shape. Yet, all covariates were selected by the boosting algorithm.

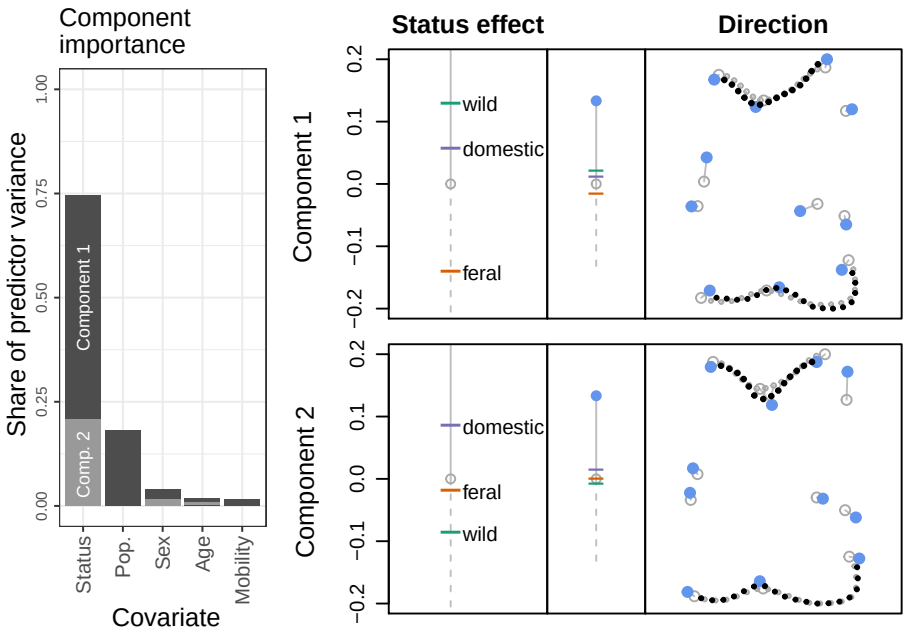

Figure 2: *Left*: Shares of different factorized covariate effects in the total predictor variance. *Right*: Factorized effect plots showing the two components of the status effect (*rows*): in the *right column*, the two first directions $\boldsymbol{\xi}_1^{(1)}, \boldsymbol{\xi}_1^{(2)} \in T_{[\mathbf{p}]}\mathcal{Y}^*_{/\,\mathrm{Trl}+\mathrm{Rot}+\mathrm{Scl}}$ 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.

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 $y_i : [0, 1] \to \mathbb{C}$, clockwisely 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 Volkmann 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 Supplement Figure S4. 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 $[y_i] \in \mathcal{Y}^*_{/\,\mathrm{Trl}+\mathrm{Rot}}$. Their estimated overall form mean $[p]$ serves as pole in the additive model

$$[\mu_i] = \mathrm{Exp}_{[p]}\big(h(\mathbf{x}_i)\big) = \mathrm{Exp}_{[p]}\big(\sum_j \beta_j x_{ij} + \sum_j f_j(x_{ij}) + \sum_{j \neq j} f_{jj}(x_{ij}, x_{ij})\big)$$

where the conditional form mean $[\mu_i]$ is modeled in dependence on tangent-space linear effects with coefficients $\beta_j \in T_{[p]}\mathcal{Y}_{/\,\mathrm{Trl}\times\mathrm{Rot}}$ and non-linear smooth effects f_j for covariate

$j = 1, \ldots, 4$, as well as smooth interaction effects $f_{j\ddot{j}}$ for each pair of covariates $j \neq \ddot{j}$. 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 $f_{j\ddot{j}}(x_j, x_{\ddot{j}})$ are centered around their marginal effects $f_j(x_j), f_{\ddot{j}}(x_{\ddot{j}})$, which are in turn centered around the linear effects $\beta_j x_j$ and $\beta_{\ddot{j}} x_{\ddot{j}}$, 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 hours 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 (Supplement Fig. S3), 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 (Supplement Fig. S2). Around 80% of its effect are explained by the linear term reflecting gradual shrinkage at the side of the cells with increasing bulk stiffness.

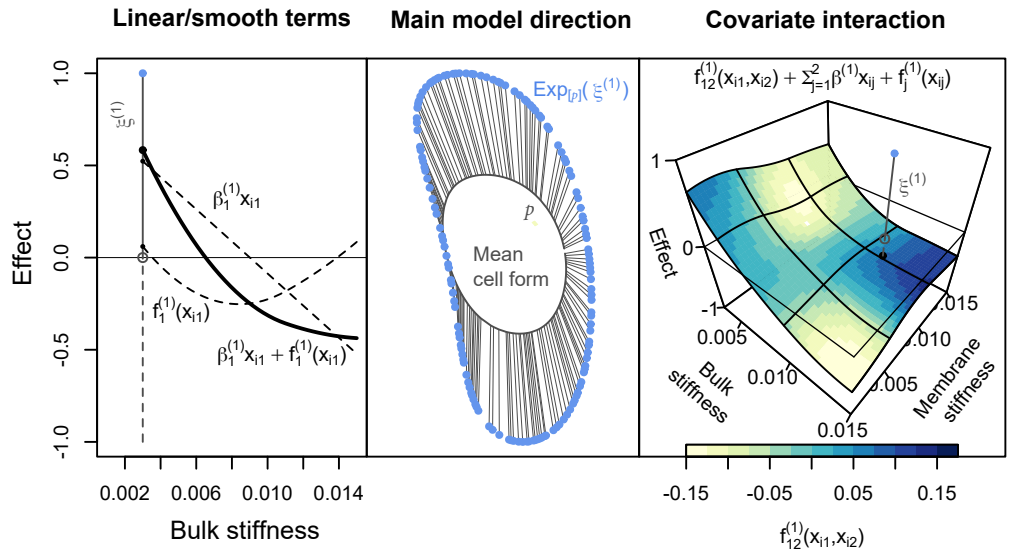

Figure 3: *Center:* the main direction $\xi^{(1)}$ of the model illustrated as vectors pointing from the overall mean cell form $[p]$ (*grey curve*) to the form $\text{Exp}_{[p]}(\xi^{(1)})$ (*blue dots*), which are both oriented as cells migrating rightwards. *Left:* Effects of the bulk stiffness x_{i1} 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 asses 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.

5.3 Realistic shape and form simulation studies

To evaluate the proposed approach, we conduct simulation studies for both form and shape 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 $n = 720$, i.e. 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 Supplement S.5.

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 $T_{[p]}\mathcal{Y}^*_{/G}$, placing them irregularly at 1/27-quantiles of unit-speed parameterization time-points of the curves. Cubic B-splines with 4 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. 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 towards 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 $[y_1], \ldots, [y_n]$ via residual re-sampling (Supplement S.5) to preserve realistic autocorrelation. Subsequently, we randomly translate, rotate and scale $y_1, \ldots, y_n \in \mathcal{Y}$ somewhat around the aligned form/shape 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 βz_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 (Fig. 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

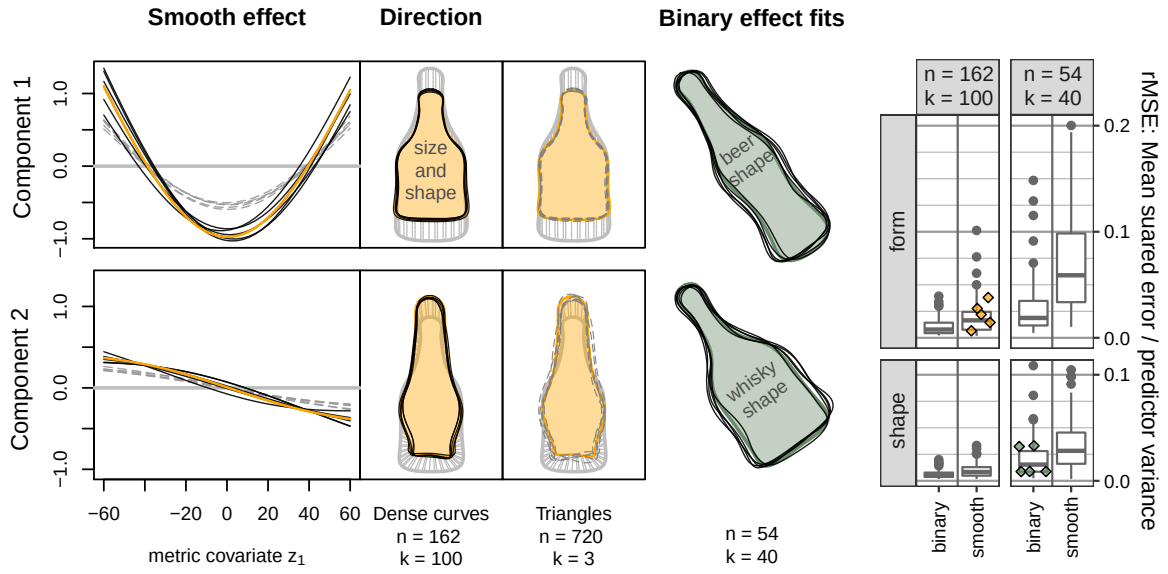

Figure 4: *Left:* First *(row 1)* and second *(row 2)* main components of the smooth effect $f_1(z_1)$ in the form scenario obtained from TP factorization. Normalized component directions are visualized as bottle outlines after transporting them to the true pole *(gray solid outline)*. Underlying truth *(orange solid lines / 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 *(olive-green areas)*. *Right:* rMSE of shown example estimates *(jittered colored diamonds)* contextualized with boxplots of rMSE distributions observed in respective simulation scenarios.

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 (Fig. 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 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, e.g., 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 factor-

ization 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 non-parametric approach of Jeon and Park (2020) in Supplement S.8.

Instead of operating on the original evaluations $\mathbf{y}_i \in \mathbb{C}^{k_i}$ of response curves y_i as in all applications above, another frequently used approach expands y_i, $i = 1, \ldots, n$, 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 (Supplement S.6).

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 non-zero 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.a., 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.

Acknowledgement

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. Moreover, we gratefully acknowledge funding by grant GR 3793/3-1 from the German research foundation (DFG).

SUPPLEMENTARY MATERIAL

Supplementary material with further details is provided in an online supplement.

References

Adams, D., F. Rohlf, and D. Slice (2013). A field comes of age: geometric morphometrics in the 21st century. *Hystrix, the Italian Journal of Mammalogy 24*(1), 7–14.

Backenroth, D., J. Goldsmith, M. D. Harran, J. C. Cortes, J. W. Krakauer, and T. Kitago (2018). Modeling motor learning using heteroscedastic functional principal components analysis. *Journal of the American Statistical Association 113*(523), 1003–1015.

Baranyi, P., Y. Yam, and P. Várlaki (2013). *Tensor product model transformation in polytopic model-based control*. CRC press.

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., M. Melcher, F. Leisch, and S. Greven (2017). Boosting flexible functional regression models with a high number of functional historical effects. *Statistics and Computing 27*(4), 913–926.

Brockhaus, S., D. Rügamer, and S. Greven (2020). Boosting functional regression models with FDboost. *Journal of Statistical Software 94*(10), 1–50.

Brockhaus, S., F. Scheipl, and S. Greven (2015). The Functional Linear Array Model. *Statistical Modelling 15*(3), 279–300.

Bühlmann, P. and B. Yu (2003). Boosting with the L2 loss: regression and classification. *Journal of the American Statistical Association 98*(462), 324–339.

Clutton-Brock, J., K. Dennis-Bryan, P. L. Armitage, and P. A. Jewell (1990). Osteology of the Soay sheep. *Bull. Br. Mus. Nat. Hist. 56*(1), 1–56.

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.

Davis, B. C., P. T. Fletcher, E. Bullitt, and S. Joshi (2010). Population shape regression from random design data. *International journal of computer vision 90*(2), 255–266.

Dryden, I. L. and K. V. Mardia (2016). *Statistical Shape Analysis: With Applications in R*. John Wiley & Sons.

Ferraty, F., A. Goia, E. Salinelli, and P. Vieu (2011). Recent advances on functional additive regression. *Recent Advances in Functional Data Analysis and Related Topics*, 97–102.

Fletcher, P. T. (2013). Geodesic regression and the theory of least squares on Riemannian manifolds. *International Journal of Computer Vision 105*(2), 171–185.

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.

Happ, C. and S. Greven (2018). Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association 113*(522), 649–659.

Hinkle, J., P. T. Fletcher, and S. Joshi (2014). Intrinsic polynomials for regression on Riemannian manifolds. *Journal of Mathematical Imaging and Vision 50*(1), 32–52.

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*(1-2), 1–14.

Hong, Y., N. Singh, R. Kwitt, and M. Niethammer (2014). Time-warped geodesic regression. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 105–112. Springer.

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.

Jeon, J. M., Y. K. Lee, E. Mammen, and B. U. Park (2022). Locally polynomial hilbertian additive regression. *Bernoulli 28*(3), 2034–2066.

Jeon, J. M. and B. U. Park (2020). Additive regression with hilbertian responses. *The Annals of Statistics 48*(5), 2671–2697.

Jeon, J. M., B. U. Park, and I. Van Keilegom (2021). Additive regression for non-euclidean responses and predictors. *The Annals of Statistics 49*(5), 2611–2641.

Jupp, P. E. and J. T. Kent (1987). Fitting smooth paths to spherical data. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 36*(1), 34–46.

Karcher, H. (1977). Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics 30*(5), 509–541.

Kendall, D. G., D. Barden, T. K. Carne, and H. Le (1999). *Shape and shape theory*, Volume 500. John Wiley & Sons, LTD.

Kent, J. T., K. V. Mardia, R. J. Morris, and R. G. Aykroyd (2001). Functional models of growth for landmark data. *Proceedings in Functional and Spatial Data Analysis 109115*.

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 (mglm) 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.

Kim, H. J., N. Adluru, H. Suri, B. C. Vemuri, S. C. Johnson, and V. Singh (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.

Klingenberg, W. (1995). *Riemannian geometry*. de Gruyter.

Kneib, T., T. Hothorn, and G. Tutz (2009). Variable selection and model choice in geoadditive regression models. *Biometrics 65*(2), 626–634.

Kume, A., I. L. Dryden, and H. Le (2007). Shape-space smoothing splines for planar landmark data. *Biometrika 94*(3), 513–528.

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.

Li, Y. and D. Ruppert (2008). On the asymptotics of penalized splines. *Biometrika 95*(2), 415–436.

Lin, L., B. St. Thomas, H. Zhu, and D. B. Dunson (2017). Extrinsic local regression on manifold-valued data. *Journal of the American Statistical Association 112*(519), 1261–1273.

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.

Mallasto, A. and A. Feragen (2018). Wrapped gaussian process regression on riemannian manifolds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5580–5588.

Mayr, A., H. Binder, O. Gefeller, and M. Schmid (2014). The evolution of boosting algorithms. *Methods of information in medicine 53*(06), 419–427.

Meyer, M. J., B. A. Coull, F. Versace, P. Cinciripini, and J. S. Morris (2015). Bayesian function-on-function regression for multilevel functional data. *Biometrics 71*(3), 563–574.

Morris, J. S. (2015). Functional Regression. *Annual Review of Statistics and its Applications 2*, 321–359.

Morris, J. S. and R. J. Carroll (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society, Series B 68*(2), 179–199.

Müller, H.-G. and F. Yao (2008). Functional additive models. *Journal of the American Statistical Association 103*(484), 1534–1544.

Muralidharan, P. and P. T. Fletcher (2012). Sasaki metrics for analysis of longitudinal data on manifolds. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 1027–1034. IEEE.

Olsen, N. L., B. Markussen, and L. L. Raket (2018). Simultaneous inference for misaligned multivariate functional data. *Journal of the Royal Statistical Society: Series C 67*(5), 1147–1176.

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.

Pigoli, D., A. Menafoglio, and P. Secchi (2016). Kriging prediction for manifold-valued random fields. *Journal of Multivariate Analysis 145*, 117–131.

Pöllath, N., R. Schafberg, and J. Peters (2019). Astragalar morphology: Approaching the cultural trajectories of wild and domestic sheep applying geometric morphometrics. *Journal of Archaeological Science: Reports 23*, 810–821.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis*. Springer New York.

Rosen, O. and W. K. Thompson (2009). A Bayesian regression model for multivariate functional data. *Computational statistics & data analysis 53*(11), 3773–3786.

Schafberg, R. and J. Wussow (2010). Julius Kühn. Das Lebenswerk eines agrarwissenschaftlichen Visionärs. *Züchtungskunde 82*(6), 468–484.

Schaffer, S. A. (2021). *Cytoskeletal dynamics in confined cell migration: experiment and modelling*. PhD thesis, LMU Munich. DOI: 10.5282/edoc.28480.

Scheipl, F., A.-M. Staicu, and S. Greven (2015). Functional additive mixed models. *Journal of Computational and Graphical Statistics 24*(2), 477–501.

Schiratti, J.-B., S. Allassonnière, O. Colliot, and S. Durrleman (2017). A bayesian mixed-effects model to learn trajectories of changes from repeated manifold-valued observations. *The Journal of Machine Learning Research 18*(1), 4840–4872.

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. and E. P. Klassen (2016). *Functional and Shape Data Analysis*. Springer-Verlag.

Stöcker, A., S. Brockhaus, S. A. Schaffer, B. v. Bronk, M. Opitz, and S. Greven (2021). Boosting functional response models for location, scale and shape with an application to bacterial competition. *Statistical Modelling 21*(5), 385–404.

Stöcker, A., M. Pfeuffer, L. Steyer, and S. Greven (2022). Elastic full Procrustes analysis of plane curves via Hermitian covariance smoothing.

Thüroff, F., A. Goychuk, M. Reiter, and E. Frey (2019, dec). Bridging the gap between single-cell migration and collective dynamics. *eLife 8*, e46842.

Volkmann, A., A. Stöcker, F. Scheipl, and S. Greven (2021). Multivariate functional additive mixed models. *Statistical Modelling*.

Wood, S. N., N. Pya, and B. Säfken (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association 111*(516), 1548–1563.

Yao, F., H. Müller, and J. Wang (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association 100*(470), 577–590.

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.

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.

Zhu, H., R. Li, and L. Kong (2012). Multivariate varying coefficient model for functional responses. *Annals of statistics 40*(5), 2634–2666.

Zhu, H., J. S. Morris, F. Wei, and D. D. Cox (2017). Multivariate functional response regression, with application to fluorescence spectroscopy in a cervical pre-cancer study. *Computational Statistics and Data Analysis 111*, 88–101.

6. Elastic Analysis of Irregularly and Sparsely Sampled Curves

Complementary to Chapter 5, we consider curves with fixed orientation and size in this contribution – but as equivalence classes modulo re-parameterization (warping). Based on the square-root-velocity (SRV) framework, we develop methods for estimating Fréchet means of irregularly/sparsely sampled curves using spline-representations and show identifiability statements for these. Using the "elastic" metric, proper distances can be defined via optimal warping alignment (registration) of parameterized curves (up to translation). Moreover, we illustrate the use of elastic distances for clustering and classification in irregularly sampled curves in data on undocumented walking paths in Berlin Tempelhofer Feld and a spiral test used for diagnosis of Parkinson's disease.

Contributing article:
Steyer, L., Stöcker, A., and Greven, S. (2022). Elastic analysis of irregularly or sparsely sampled curves. *Biometrics*. DOI: 10.1111/biom.13706.

Declaration on personal contributions:
Major parts of this project were conducted by Lisa Steyer. The author of this thesis was involved in the development of central research questions of the project and was passively/supportingly involved throughout the process taking a consulting role (including joint prototyping of implementations together with Lisa Steyer).

Received: 6 August 2021 | Accepted: 1 June 2022

DOI: 10.1111/biom.13706

BIOMETRIC METHODOLOGY

Biometrics WILEY

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 $\boldsymbol{\beta} : [0,1] \to \mathbb{R}^d$, $d \in \mathbb{N}$. 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

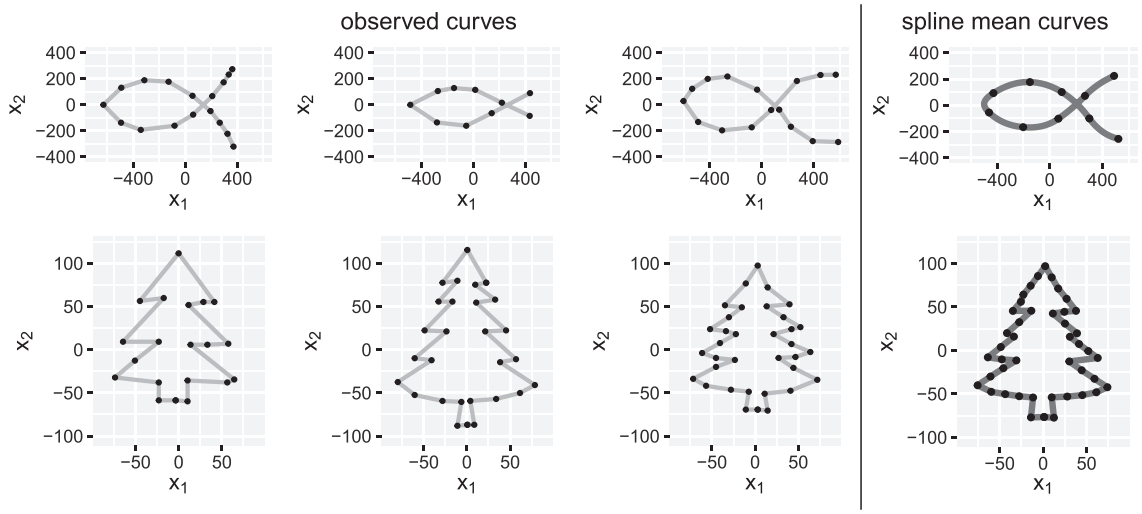

FIGURE 1 Two toy examples of sparsely and irregularly observed curves in $\mathbb{R}^2$ with observed points indicated as black dots and linear interpolation (first three columns). Ideally, the analysis should yield an optimal alignment of different curves to allow comparison of corresponding points 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 $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, 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 $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 : [0,1] \to \mathbb{R}^d$ be absolutely continuous and $[\boldsymbol{\beta}_1]$ and $[\boldsymbol{\beta}_2]$ their respective equivalence classes modulo parameterization and translation. Then the elastic distance between $[\boldsymbol{\beta}_1]$ and $[\boldsymbol{\beta}_2]$ is

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

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

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

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

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

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

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

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

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

where $\langle \cdot, \cdot \rangle_+$ denotes the positive part of the scalar product in $\mathbb{R}^d$. For a maximizer $\boldsymbol{t}$ of (3), there is a $\gamma : [0,1] \to [0,1]$ 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 $\gamma : [0,1] \to [0,1]$, onto and with $\dot{\gamma} \ge 0$ 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 $\boldsymbol{p}$. If $\boldsymbol{p}$ 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 constrains (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 $\boldsymbol{q}$ to a curve with SRV transformation $\boldsymbol{p}$, 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 $\boldsymbol{p}$. If $\boldsymbol{p}$ 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 $\boldsymbol{p}, \boldsymbol{q}$; convergence tolerance $\epsilon > 0$;

starting values $0 \leqslant t_1^{(0)} \leqslant \cdots \leqslant t_{m-1}^{(0)} \leqslant 1$ `// e.g. relative arc length`

for $k \in \mathbb{N}$ **do**

  for $j = 1, \ldots, m-1$ **do**

    if $j - k$ *even* **then**

      $t_j^{(k)} = \operatorname{argmax}_{t_j \in \left[t_{j-1}^{(k-1)}, t_{j+1}^{(k-1)}\right]} \Phi \mid_{\{t_{j'} = t_{j'}^{(k-1)}, j' \neq j\}}$

    else if $j - k$ *odd* **then**

      $t_j^{(k)} = t_j^{(k-1)}$

  if $\|\boldsymbol{t}^{(k)} - \boldsymbol{t}^{(k-2)}\| < \epsilon$ *and* $\|\boldsymbol{t}^{(k-1)} - \boldsymbol{t}^{(k-3)}\| < \epsilon$ **then**

    return $\boldsymbol{t}^{(k)} = (t_1^{(k)}, \ldots, t_{m-1}^{(k)})$

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 $\boldsymbol{q}$, 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 "elasdics." 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* $(\boldsymbol{t}^{(k)})_{k\in\mathbb{N}} = (t_1^{(k)}, \ldots, t_{m-1}^{(k)})_{k\in\mathbb{N}}$ *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 $(\boldsymbol{t}^{(k)})_{k\in\mathbb{N}}$ has more than one accumulation point, they all give the same value $\Phi(\boldsymbol{t})$. 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 $\boldsymbol{t}$ to obtain a locally optimal warping function. As we cannot guarantee this $\boldsymbol{t}$ 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 $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_d)^T : [0,1] \to \mathbb{R}^d$ with $d \in \mathbb{N}$ a d-dimensional spline curve of degree $l \in \mathbb{N}_0$ if all its components $\xi_1, \ldots, \xi_d : [0,1] \to \mathbb{R}$ are spline curves of degree l with a common knot set $0 = \kappa_0 < \kappa_1 < \cdots < \kappa_{K-1} < \kappa_K = 1$ for some $K \geq 2$. That means that $\xi_1, \ldots, \xi_d$ are piecewise polynomial of degree l between the knots $\kappa_0, \ldots, \kappa_K$, as well as continuous and $(l-1)$-times continuously differentiable on the whole domain [0,1] for $l \geq 1$. Denote by $\mathcal{S}^l_{K;\kappa_0,\ldots,\kappa_K}$ the set of all such spline curves.

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

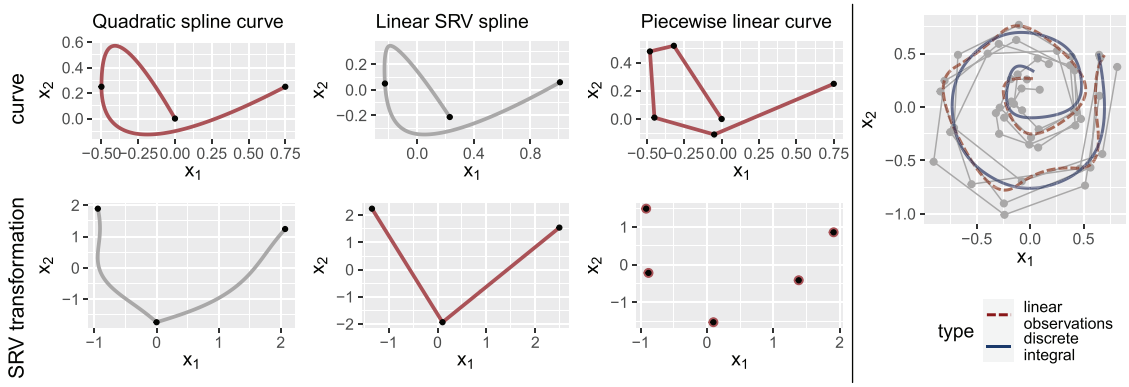

FIGURE 2 Left: Two-dimensional curves and corresponding SRV transformations. Spline curves are plotted as red curves with their values at knots marked as black dots; other curves are gray. Note that the SRV-curve in the sixth panel is piecewise constant in t and t is not visible in the image. Right: 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 $\boldsymbol{\beta}$ if $\boldsymbol{p}$ is a spline of degree $l \geq 1$. Only if $\boldsymbol{\beta}$ is piecewise linear ($l = 1$), then both $\boldsymbol{\beta}$ 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 $\boldsymbol{p}$ is a continuous SRV transformation of $\boldsymbol{\beta}$, the backtransform $\boldsymbol{\beta}(t) = \boldsymbol{\beta}(0) + \int_0^t \boldsymbol{p}(s)\|\boldsymbol{p}(s)\|ds$ 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 $(\Omega, \mathcal{F}, P)$ be a probability space and $\mathcal{X}$ a metric space with distance function d, equipped with the Borel-σ-Algebra. For a random variable $X : \Omega \to \mathcal{X}$, we call every element in $\operatorname{arginf}_{A\in\mathcal{X}} \boldsymbol{E}_P(d(X, A)^2)$ an expected element of X. For a set of observations $x_1, \ldots, x_n \in \mathcal{X}$, 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

$X_1, \dots, X_n : \Omega \to \mathcal{X}$, 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 $\boldsymbol{q}_1, \dots, \boldsymbol{q}_n$ and for a given degree $l \in \{0, 1\}$ and a given set of knots $\kappa_0, \dots, \kappa_K$, we thus define

$$\bar{\boldsymbol{p}} \in \operatorname*{arginf}_{\boldsymbol{p} \in S^l_{K;\kappa_0,\dots,\kappa_K}} \sum_{i=1}^{n} \inf_{\gamma_i} \left\| \boldsymbol{p} - (\boldsymbol{q}_i \circ \gamma_i)\sqrt{\dot{\gamma}_i} \right\|^2_{L_2} \tag{4}$$

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 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).

Algorithm 2: Elastic spline Fréchet mean

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

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

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

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

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

$\bar{\boldsymbol{p}}_{new} = \operatorname{arginf}_{\bar{\boldsymbol{p}}} \sum_{i=1}^{n} \left\| \bar{\boldsymbol{p}} - (\boldsymbol{q}_i \circ \gamma_i)\sqrt{\dot{\gamma}_i} \right\|^2_{L_2}$ `// L2 spline fit via (weighted) least-squares`

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

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 $\boldsymbol{q}_i$, $i = 1, \dots n$, as in Lemma 1. We tackle the remaining maximization problem (3) using a gradient descent algorithm as discussed before if $\bar{\boldsymbol{p}}$ is piecewise linear and Algorithm 1 if $\bar{\boldsymbol{p}}$ is piecewise constant. In the L_2 spline fitting step, the integrals $\|\bar{\boldsymbol{p}} - (\boldsymbol{q}_i \circ \gamma_i)\sqrt{\dot{\gamma}_i}\|^2_{L_2}$ in the sum need to be approximated, because the curves $\boldsymbol{\beta}_i$ are only observed on a finite grid $0 = s_{i,0} \le s_{i,1} \le \dots \le s_{i,m_i} = 1$, and the SRV-curves $\boldsymbol{q}_1, \dots, \boldsymbol{q}_n$ are thus unobserved. One option is to assume that the SRVs $\boldsymbol{q}_i$ of the observed curves are piecewise constant as in the warping step. As $\bar{\boldsymbol{p}}$ is piecewise linear, $(\boldsymbol{q}_i \circ \gamma_i)\sqrt{\dot{\gamma}_i}$ 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 $\bar{\boldsymbol{p}}$. (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 $P : [0, 1] \to \mathbb{R}$, 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 $\gamma_a'(x) \ge 0$ 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 $\boldsymbol{Q}, \boldsymbol{P} : [0,1] \to \mathbb{R}^d$ be quadratic or cubic spline curves, where $\boldsymbol{Q}$ has a nonlinear image between each of its knots. Moreover, let $\gamma : [0,1] \to [0,1]$ be monotonically increasing and onto. Then $\boldsymbol{P} = \boldsymbol{Q} \circ \gamma \Rightarrow \gamma = id$.*

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 $\boldsymbol{Q}$ and $\boldsymbol{P}$. That means there is almost always a unique representative modulo warping in $\bigcup_{K,\kappa_0,\ldots,\kappa_K} S^l_{K;\kappa_0,\ldots,\kappa_K}$ 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 $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 : [0,1] \to \mathbb{R}^d$ with SRV functions $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$, respectively. If $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$ are nowhere constant linear splines and $\boldsymbol{q}_2(t) = \boldsymbol{q}_1(\gamma(t))\sqrt{\dot{\gamma}(t)}$, then $\boldsymbol{q}_1 = \boldsymbol{q}_2$.*

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 $S^1_{K;\kappa_0,\ldots,\kappa_K}$ 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 $K \geq 2$, $0 = \kappa_0 < \cdots < \kappa_K = 1$: (i) a subset of $S^l_{K;\kappa_0,\ldots,\kappa_K}$, $l = 2, 3$, 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 $S^1_{K;\kappa_0,\ldots,\kappa_K}$ from Corollary 1; or (iii) the set of curves with piecewise constant SRV-curves in $S^0_{K;\kappa_0,\ldots,\kappa_K}$ from Remark 1.

Lemma 2 (Topological embedding). *Let $f : (\Xi, \|\cdot\|) \to (\mathcal{A}, d)$ 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 $\boldsymbol{\beta}$ with known B-spline coefficients $\vartheta_1, \dots, \vartheta_B$, we generate a sample of observed curves $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_n$ 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 $\boldsymbol{\beta}_i$ 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 $n \to \infty$, although we do not expect them to precisely converge to the coefficients of the template in this simulation setup, not even if $m_i \to \infty$ for all $i = 1, \dots, n$. This is because we draw the sample curves $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_n$ such that $\boldsymbol{\beta}$ is the mean w.r.t. the L_2 distance on SRV level, but this does in general not imply that $\boldsymbol{\beta}$ 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.

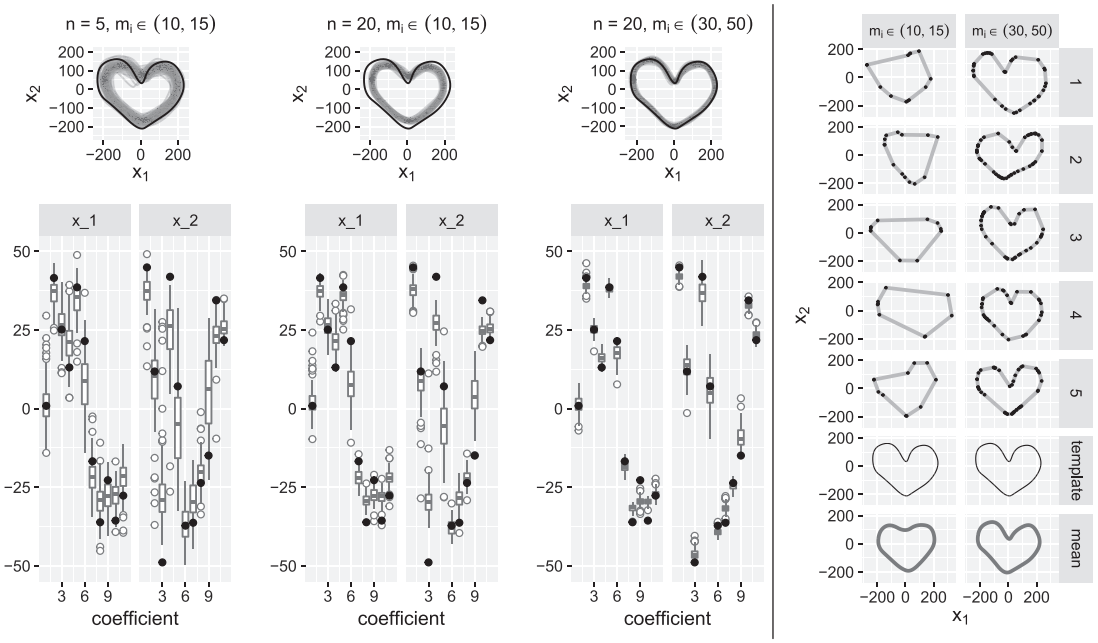

FIGURE 3 Top left: Smooth means (in gray) computed for a set of n simulated curves drawn from the heart-shaped template curve (in black) via sampling its B-spline coefficients from a normal distribution with standard deviation $\sigma = 4$ and $m_i, i = 1, \ldots, n$ points observed per curve. The means 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 $i = 1, \ldots 5$ 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 "fdasrvf" 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

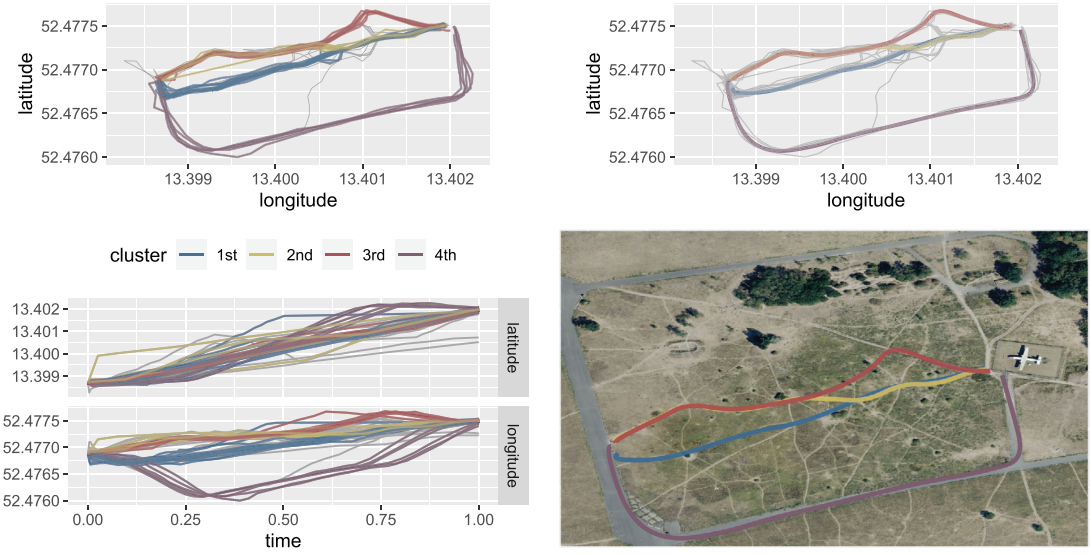

FIGURE 4 Top left: The observed trajectories with elements of the four largest clusters indicated by color. Bottom left: Longitude and latitude for the trajectories (with the four largest clusters indicated by the same colors) over relative time. Top right: Smooth means modeled as linear SRV-curves with 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., Chkrebtii, 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 mi, i=1, . . . , 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 σ=0.4 and mi, i=1,..., 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 mi ϵ30, 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. (2022). Elastic analysis of irregularly or sparsely sampled curves. *Biometrics*, 1–13. https://doi.org/10.1111/biom.13706

7. Elastic Full Procrustes Analysis of Plane Curves via Hermitian Covariance Smoothing

Generalized Procrustes analysis (GPA) presents a popular procedure in statistical shape analysis, superimposing landmark configurations by optimal translation, rotation and scaling alignment to their full Procrustes mean. In this contribution, we generalize GPA to elastic full Procrustes analysis of shapes of plane curves. The full Procrustes shape distance used here differs from the Riemannian distance used in Chapters 5 and 8 but offers an important connection to covariance estimation as conducted in Chapter 4. This lets us estimate elastic full Procrustes means of plane curves modulo translation, rotation, scale and warping based on their complex covariance structure. To this end, we develop Hermitian covariance smoothing as a generalization of symmetric covariance smoothing used in Chapter 4 to complex-valued stochastic processes. In preparation of the proposed method, we present two results that shed light on the role of the complex covariance operator and complex principal component analysis of rotation-invariant data, and provide a theorem on feasible SRV-representation of irregularly/sparsely sampled curves, which are also of independent interest. We illustrate the performance of our method in familiar everyday shapes and a phonetic analysis of tongue shape 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.*
DOI: 10.48550/ARXIV.2203.10522.
Supplementary material provided in Appendix C.

Declaration on personal contributions:
After the author of this thesis prototyped an inelastic version of this approach (originally intended for alternative overall mean estimation in Chapter 5 above), he developed the idea for the elastic extension with Lisa Steyer and supervised Manuel Pfeuffer's master's thesis implementing and refining it together with her and Sonja Greven. He wrote wide parts of the manuscript including proposal and proofs of theoretical results. Lisa Steyer substantially contributed to the proofs of Proposition 1 and Theorem 3. Software (also for the applications) was implemented by Manuel Pfeuffer, who also wrote single parts of the manuscript concerning the "fda" simulation study. All authors played an advisory role in all parts of the project.

Elastic Full Procrustes Analysis of Plane Curves via Hermitian Covariance Smoothing

BY A. STÖCKER, M. PFEUFFER, L. STEYER AND S. GREVEN

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

almond.stoecker@hu-berlin.de manuel.pfeuffer@esmt.org lisa.steyer@hu-berlin.de sonja.greven@hu-berlin.de

SUMMARY

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 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 framework. Identifying the real plane with the complex numbers, we establish a connection to covariance estimation in irregular/sparse functional data analysis and propose Hermitian covariance smoothing for (in)elastic full Procrustes mean estimation. 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.

Some key words: Complex Gaussian process; Functional data; Phonetic tongue shape; Principal component analysis; Shape analyis; 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 $\beta : [0, 1] \to \mathbb{R}^2$, 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 & Mardia, 2016) with invariances a) to this functional (curve) case, implying the alignment of the recorded data with respect to all involved invariances a) and b).

For landmark shapes (i.e. a)), 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 (Dryden & Mardia, 2016) 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, invariance with respect to re-parameterization (warping) b) poses an additional challenge in the analysis of curves, which is highly related to the registration problem in function data analysis (FDA, Ramsay & Silverman, 2005). In this context, Srivastava et al. (2011) propose an *elastic* re-parameterization invariant metric on curves, 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, their framework also allows incorporation of shape invariances a) along the lines of statistical shape analysis. This lead to a rapidly growing literature on *functional* shape analysis of curves in the SRV-framework (see e.g., Srivastava & 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. We will compare both approaches in our simulations.

Moreover, except for Steyer et al. (2021) considering only reparameterization invariance b), analysis of sparsely/irregularly observed curves in the SRV-framework has not yet been considered. 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 (Ramsay & Silverman, 2005), 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 involve an assumption of (small) measurement errors on the discrete curve evaluations as common in practice (Greven & Scheipl, 2017).

Focusing on shape analysis of sparsely/irregularly measured curves combining a) and b), 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 & Xu (2020) discuss (symmetric) tensor-product spline smoothing for this purpose, considering univariate functional data. Happ & Greven (2018) generalize univariate approaches to conduct functional principal component analysis also for multivariate sparse/irregular data. Here, our contributions are to 1. propose Hermitian covariance smoothing for complex functions and 2. use it as tool in the estimation of the 3. (elastic) full Procrustes means we propose to 4. handle sparsely/irregularly measured plane curves.

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 as our first contribution. This lays the groundwork for the second part and second contribution 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 sparsely/irregularly observing SRVs of curves (i.e., curve derivatives) 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. 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 & Eubank, 2015), underlying functional analytic statements cover Hilbert spaces over $\mathbb{C}$ as well (e.g., Rynne & 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 & 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 $\kappa : \mathbb{C} \to \mathbb{R}^2$, $z \mapsto \mathbf{z} = (\Re(z), \Im(z))^\top$ 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 $\exp(\mathbf{i}\,\omega)\, z \in \mathbb{C}$.

Let Y be a complex-valued stochastic process with realizations $y : \mathcal{T} \to \mathbb{C}$ 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 $\Sigma : \mathbb{H} \to \mathbb{H}$ of a random element Y are defined via $\langle \mu, x\rangle = \mathbb{E}\left(\langle Y, x\rangle\right)$ and $\langle \Sigma(x), y\rangle = \mathbb{E}\left(\langle x, Y - \mu\rangle\langle Y - \mu, y\rangle\right)$ 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 $Y = \mu + \sum_{k\geq 1} \langle e_k, Y\rangle e_k$ with probability one. The scores $Z_k = \langle e_k, Y\rangle$, $k \geq 1$, are complex random variables with mean zero and covariance $\mathrm{Cov}\left(Z_k, Z_{k'}\right) = \mathbb{E}\left(\langle Y, e_k\rangle\langle e_{k'}, Y\rangle\right) = \lambda_k 1_{\{k'\}}(k)$, 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 $\langle \mathbf{x}, \mathbf{y}\rangle = \int \Re\left(x(t)\right) \Re\left(y(t)\right)\, d\nu(t) + \int \Im\left(x(t)\right) \Im\left(y(t)\right)\, d\nu(t) = \Re\left(\langle x, y\rangle\right)$.

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

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

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 & Massey, 1993;

Picinbono, 1996; Kent, 1994) and the simplification by taking a complex approach becomes evident:

THEOREM 2. *A stochastic process Y 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 $Z_k = \langle e_k, Y\rangle$ 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 $\mathbf{\Sigma}$ 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 eigen decomposition $\mathbf{\Sigma}(\mathbf{f}) = \sum_{k\neq 0} \lambda_k \langle \mathbf{e}_k, \mathbf{f}\rangle \mathbf{e}_k$ of $\mathbf{\Sigma}$. 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 $\operatorname{var}(\mathbf{Z}_k) = \lambda_k$ and $\kappa(Z_k) = (\mathbf{Z}_k, \mathbf{Z}_{-k})^\top$.

While rotation invariance of $\mathfrak{L}(\mathbf{Y})$ leads to even multiplicities in the eigenvalues of the bivariate covariance operator $\mathbf{\Sigma}$, it does not pose a constraint on the complex eigenvalues and eigenfunctions of Σ, which would complicate the eigendecomposition. Here, rotation invariance of $\mathfrak{L}(\mathbf{Y})$ 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 $C(s,t) = \mathbb{E}\left(Y^\dagger(s)Y(t)\right) = C(t,s)^\dagger$. 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 eigen decomposition. In fact, under standard assumptions, such as continuity of $\mu^*(t)$ and $C(s,t)$, Fubini allows switching integrals such that the point-wise mean $\mu^* = \mu$ coincides with the mean element and the operator $\Sigma^* = \Sigma$ with the covariance operator. In this case, the eigen decomposition 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 $y_1, \ldots, y_n : \mathcal{T} \to \mathbb{C}$ (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 $y_i(t_{i1}) = y_{i1}, \ldots, y_i(t_{in_i}) = y_{in_i}$ are available for $i = 1, \ldots, n$ 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 $\mathbb{E}\left(Y^\dagger(s)Y(t)\right) = C(s,t)$ with a (smooth) regression estimator $\hat{C}(s,t)$ fitted to response products $y_{ij}^\dagger y_{i\ddot{j}}$ at respective tuples $(t_{ij}, t_{i\ddot{j}}) \in \mathcal{T}^2$, for $j, \ddot{j} = 1, \ldots, n_i$ and $i = 1, \ldots, n$. 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

$\widetilde{C}(s,t) = C(s,t) + \tau^2(t)\, 1_{\{s\}}(t)$ with $\tau^2(t) = \mathrm{var}(\varepsilon(t))$ 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

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

with real-valued basis functions $f_k : \mathcal{T} \to \mathbb{R},\ k = 1, \ldots, m$, stacked to a vector $\mathbf{f}(t) = (f_1(t), \ldots, f_m(t))^\top$, and a Hermitian coefficient matrix $\boldsymbol{\Xi} = \{\xi_{kk'}\}_{kk'} = \boldsymbol{\Xi}^\dagger \in \mathbb{C}^{m\times m}$ 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 $\Re(\boldsymbol{\Xi}) = \Re(\boldsymbol{\Xi})^\top$ and the anti-symmetry of the imaginary part $\Im(\boldsymbol{\Xi}) = -\Im(\boldsymbol{\Xi})^\top$ 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 $(\mathbf{f}\otimes\mathbf{f})(s,t) = (f_1(s)\mathbf{f}^\top(t), \ldots, f_m(s)\mathbf{f}^\top(t))^\top$ 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 $\tau^2(t) \approx \boldsymbol{\xi}_\tau^\top \mathbf{f}_\tau(t)$ 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 $\mathrm{vec}\left(\widehat{\boldsymbol{\Xi}}\right) = \mathbf{D}_\Re\hat{\boldsymbol{\xi}}_\Re + \mathbf{i}\,\mathbf{D}_\Im\hat{\boldsymbol{\xi}}_\Im$ of the covariance estimator $\hat{C}(s,t)$ minimize the penalized least-squares criterion

$$\textsc{pls}(\boldsymbol{\Xi}, \boldsymbol{\xi}_\tau) = \sum_{i,j,\ddot{\jmath}} \left| \mathbf{f}^\top(t_{ij})\, \boldsymbol{\Xi}\, \mathbf{f}(t_{ij}) + \boldsymbol{\xi}_\tau^\top \mathbf{f}_\tau(t_{ij})\, 1_{\{\ddot{\jmath}\}}(j) - y_{ij}^\dagger y_{i\ddot{\jmath}} \right|^2 + \textsc{pen}(\boldsymbol{\Xi}, \boldsymbol{\xi}_\tau)$$

with quadratic penalty term PEN. They are seperately obtained for the real and imaginary part of the covariance using $\textsc{pls} = \textsc{pls}_\Re + \textsc{pls}_\Im$ via the well-known linear estimators $\hat{\boldsymbol{\xi}}_\Re \in \mathbb{R}^{(m^2+m)/2}$, $\hat{\boldsymbol{\xi}}_\tau \in \mathbb{R}^{m_\tau}$ minimizing $\textsc{pls}_\Re = \sum_{i,j,\ddot{\jmath}} (\boldsymbol{\xi}_\Re^\top \mathbf{D}_\Re\, (\mathbf{f}\otimes\mathbf{f})(t_{ij}, t_{i\ddot{\jmath}}) + \boldsymbol{\xi}_\tau^\top \mathbf{f}(t_{ij})\, 1_{\{\ddot{\jmath}\}}(j) - \Re(y_{ij}^\dagger y_{i\ddot{\jmath}}))^2 + \eta_\Re\, \boldsymbol{\xi}_\Re^\top \mathbf{D}_\Re \mathbf{P}_\otimes \mathbf{D}_\Re^\top \boldsymbol{\xi}_\Re + \eta_\tau\, \boldsymbol{\xi}_\tau^\top \mathbf{P}_\tau\, \boldsymbol{\xi}_\tau$, and $\hat{\boldsymbol{\xi}}_\Im \in \mathbb{R}^{(m^2-m)/2}$ minimizing $\textsc{pls}_\Im = \sum_{i,j,\ddot{\jmath}} (\boldsymbol{\xi}_\Im^\top \mathbf{D}_\Im\, (\mathbf{f}\otimes\mathbf{f})(t_{ij}, t_{i\ddot{\jmath}}) - \Im(y_{ij}^\dagger y_{i\ddot{\jmath}}))^2 + \eta_\Im\, \boldsymbol{\xi}_\Im^\top \mathbf{D}_\Im \mathbf{P}_\otimes \mathbf{D}_\Im^\top \boldsymbol{\xi}_\Im$. Smoothing parameters $\eta_\tau, \eta_\Re, \eta_\Im > 0$ control the penalty induced by the matrices $\mathbf{P}_\tau$ 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 $\mathbf{P}_\tau$ should typically penalize deviations from the constant. Based on a working normality assumption, $\eta_\Re, \eta_\tau$ and $\eta_\Im$ 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 eigen decomposition $\hat{C}(s,t) = \sum_{k\geq 1} \hat{\lambda}_k \hat{e}_k^\dagger(s) \hat{e}_k(t)$ of the respective covariance operator $\hat{\Sigma}$. Based on $\hat{\boldsymbol{\Xi}}$ and the Gram matrix $\mathbf{G} = \{\langle f_k, f_{k'}\rangle\}_{k,k'=1}^m$, the right eigenvalues of the matrix $\mathbf{G}^{-1}\hat{\boldsymbol{\Xi}}$ yield the eigenvalues $\hat{\lambda}_k$ of $\hat{\Sigma}$. The corresponding eigen-

vectors $\hat{\boldsymbol{\theta}}_k$ yield the eigenfunctions $\hat{e}_k(t) = \hat{\boldsymbol{\theta}}_k^\top \mathbf{f}(t)$ of $\hat{\Sigma}$ for $k = 1, \ldots, m$. 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 $\beta : [0,1] \to \mathbb{C}$, 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 $[0,1] \to \mathbb{C}$ 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 $\gamma : [0,1] \to [0,1]$ 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, rotation, re-scaling 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

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

for $\beta_1, \beta_2 \in \mathcal{AC}^*([0,1],\mathbb{C})$, with a pre-shape map $\Psi : \mathcal{AC}^*([0,1],\mathbb{C}) \to \mathbb{L}^2([0,1],\mathbb{C}),\ \beta \mapsto q$ 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 & 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 & 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 $q : [0,1] \to \mathbb{C}$ 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 $\tilde{\Psi}\left(u\, \beta \circ \gamma + v\right)(t) = (u/|u|^{1/2})\, q \circ \gamma(t) \dot{\gamma}(t)^{1/2}$, warping

and rotation act by isometries with $\|\tilde{\Psi}(a\exp(\mathbf{i}\,\omega)\,\beta_1\circ\gamma+v)-\tilde{\Psi}(a\exp(\mathbf{i}\,\omega)\,\beta_2\circ\gamma+v)\| = a^{1/2}\|\tilde{\Psi}(\beta_1)-\tilde{\Psi}(\beta_2)\|$ 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 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.

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

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

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 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. Then 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*

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

In contrast to the shape invariances, we have no closed form solution for the optimization over $\gamma\in\Gamma$ available. This makes it convenient to also define an *inelastic* full Procrustes mean of shapes of plane curves with fixed parameterization. It 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 $d_\mathcal{G}((\beta_1),(\beta_2)) = \inf_{u\in\mathbb{C}}\|q_1-uq_2\|$ with $\|q_i\|=1$ for $\Psi(\beta_i)=q_i$, $i=1,2$, defines a metric on the shape space $\tilde{\mathfrak{B}}=\{(\beta):\mathcal{AC}^*([0,1],\mathbb{C})\}$ of parameterized plane curves and can be expressed as $d_\mathcal{G}^2((\beta_1),(\beta_2))=1-|\langle q_1,q_2\rangle|^2$;*

ii) multiplication by $\langle q_1,q_2\rangle^\dagger/|\langle q_1,q_2\rangle| = \operatorname{argmin}_{u:|u|=1}\|q_1-uq_2\|$ 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 eigen decomposition 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 $\sigma^2_{\mathcal{G}} = \mathbb{E}(d^2_{\mathcal{G}}((\mu_{\mathcal{G}}),(B))) = 1 - \lambda_1$.

To allow estimation of the (in)elastic full Procrustes means from a sample of plane curves in practice, we have to address potentially sparse and/or irregular sampling points of such curves. In the following, we first answer the question of how it is still possible to work with derivative-based SRV-curves even in the sparsely observed setting. We then propose to use Hermitian covariance smoothing as introduced in Section 2 to deal with sparsity in estimation of (in)elastic full Procrustes means.

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 $\beta^* : [0,1] \to \mathbb{C}$ 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 $\beta^*\left((t_{j-1}^*, t_j^*)\right) \subset \mathbb{C}$ between the observed points in $\mathbf{b}$ have no edges or loops:

THEOREM 3 (FEASIBLE SAMPLING). *If β^* is continuous and $\beta^* : (t_{j-1}^*, t_j^*) \to \mathbb{C}$ 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*

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

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. However, if the curve has edges, they must be contained in $\mathbf{b}$, as well as a point inside of each loop. 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 $\hat q(t) = \sum_{j=1}^{n_0} q_j\, 1_{[s_{j-1}, s_j)}(t)$ 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 $s_j = \sum_{l=1}^{j} |\Delta_j| / L(\hat\beta)$ 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 Subsection 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 $\psi(t) = \sum_{k=1}^m \theta_k f_k(t) = \boldsymbol{\theta}^\top \mathbf{f}(t)$ in a basis $\mathbf{f}(t) = (f_1(t), \cdots f_m(t))^\top$ 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., 2021) 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 & Mardia, 2016, p. 139) and intrinsic elastic mean curve shapes (Srivastava & Klassen, 2016, p. 319). We follow a similar strategy to find an estimator $\hat{\psi}(t) = \hat{\boldsymbol{\theta}}^\top \mathbf{f}(t)$ 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. (2021), which has also proven suitable for irregularly/sparsely sampled curves. Involved 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 $n = 1, 2, \dots$:

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 $\hat{C}^{[h]}(s,t) = \mathbf{f}(s)^\top \hat{\boldsymbol{\Xi}}^{[h]} \mathbf{f}(t)$ with coefficient matrix $\hat{\boldsymbol{\Xi}}^{[h]} \in \mathbb{C}^{m \times m}$. While for dense sampling, an estimate can be directly obtained from the covariance of the $\langle q_i^{[h]}, f_k \rangle$ (Supplement), we propose Hermitian covariance smoothing as described in Section 2 for sparse/irregular data. This also 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 $\hat{\psi}^{[h]}(t) = \hat{e}_1^{[h]}(t) = \hat{\boldsymbol{\theta}}_1^{[h]\top} \mathbf{f}(t)$ to the leading eigenfunction of $\hat{\Sigma}^{[h]}$ obtained from the leading right eigenvector $\hat{\boldsymbol{\theta}}_1^{[h]}$ of $\mathbf{G}^{-1} \hat{\boldsymbol{\Xi}}^{[h]}$ 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 $u_i^{[h]} = \left(z_{i1}^{[h]} / |z_{i1}^{[h]}|\right)^\dagger (L^{[h]}(\beta_i))^{-1/2}$ with $z_{i1}^{[h]} = \langle \hat{e}_1^{[h]}, q_i \rangle$, $u_i^{[h]} q_i^{[h]}$ 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 $\hat{z}_{i1}^{[h]} = \mathbb{E}\big(\langle \hat{e}_1^{[h]}, Q \rangle \mid Q(t_{ij}) + \varepsilon(t_{ij}) = q_{ij}^{[h]},\ j = 1, \dots, n_i\big)$ and $\hat{L}^{[h]}(\beta_i) = \mathbb{E}\big(\|Q\|^2 \mid$

$Q(t_{ij}) + \varepsilon(t_{ij}) = q_{ij}^{[h]},\ j = 1, \ldots, n_i\big)$ 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 ith curve is (approximately) warping aligned to $\hat{\mu}^{[h]}$ using the approach of Steyer et al. (2021), where SRV-transforms are approximated as piece-wise constant functions $\hat{q}_i^{[h]}(t) \approx q_i^{[h]}(t)$ to find the infimum of $\|\hat{\mu}^{[h]} - \hat{q}_i^{[h]} \circ \gamma\, \dot{\gamma}^{1/2}\|$ over $\gamma \in \Gamma$. This yields new parameterization timepoints $t_{ij}^{[h+1]}$, $j = 1, \ldots, n_i$, 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 samples 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 and 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 & Mardia (2016) 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 $n_i \in \{17, \ldots, 22\}$ sampling points per spiral or with $n_i \in \{4, \ldots, 7\}$ 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 & Silverman (2005) 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 "*fda*" writings can be found in Figure S1 in the Online Supplement.

Based on `digit3.dat`, we compare our elastic full Procrustes mean estimator $\hat{\mu}_{\mathcal{E}}$ with its inelastic analog $\hat{\mu}_{\mathcal{G}}$ and with an elastic curve mean estimator $\hat{\mu}_{\mathcal{C}}$ taking shape invariances not into account (fitted with R package `elasdics`). Moreover, we investigate fitting performance of $\hat{\mu}_{\mathcal{E}}$ 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 $\hat{\mu}_{\mathcal{E}}$ and $\hat{\mu}_{\mathcal{G}}$. No penalty is available for $\hat{\mu}_{\mathcal{C}}$. Figure 1 shows the estimates fitted on the first $n = 4$ digits in the dataset. Without warping alignment, $\hat{\mu}_{\mathcal{G}}$ does not capture the pronounced central nose in the digit "*3*" as distinctly as $\hat{\mu}_{\mathcal{E}}$. 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, $\hat{\mu}_{\mathcal{C}}$ is very close to $\hat{\mu}_{\mathcal{E}}$ when fitting on all digits. When fitting only on the first $n = 4$ digits in the data, however, $\hat{\mu}_{\mathcal{C}}$ 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) $\hat{\mu}_{\mathcal{C}}$ being more affected by the one outlying "*3*" (top-left) than $\hat{\mu}_{\mathcal{E}}$, 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

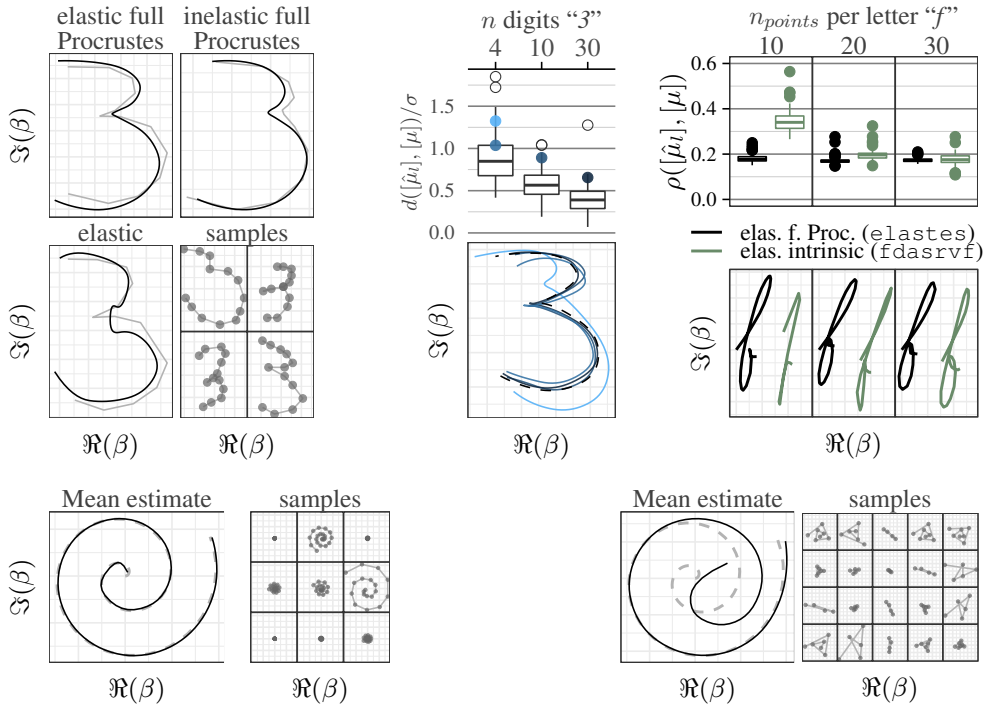

Fig. 1: *Top left:* Different digit "*3*" mean curves (*black*: order 1, *grey*: order 0 B-splines on SRV-level) estimated on the first $n = 4$ sample polygons in `digit3.dat` shown in the bottom-right. *Top center:* Simulation results from 101-fold bootstrap samples of different sample sizes on `digit3.dat`. 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 boostrap 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 polygons displayed to their right, in front of the original spiral (*grey, dashed line*).

mean estimates $\hat{\mu}_l$, $l = 1, \ldots, 101$, 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 $d_l = d_{\mathcal{E}}([\hat{\mu}_l], [\mu])$, $l = 1, \ldots, 101$. 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, however, 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 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 actually 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 sur-

face 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 ($\mathcal{S} = \{1, \ldots, 6\}$) repeating the same set of fictitious words, such as "`pada`", "`pidi`", "`pala`" or "`pili`". The words implement different combinations of two flanking vowels in $\mathcal{V} = \{\mathtt{a}*\mathtt{a}, \mathtt{i}*\mathtt{i}\}$ around a consonant in $\mathcal{C} = \{\mathtt{d}, \mathtt{l}, \mathtt{n}, \mathtt{s}\}$. 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 $\mathbf{b}_i \in \mathbb{C}^{n_i}$, $i = 1, \ldots, n$, 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 ith 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 $c \in \{\mathtt{d}, \mathtt{s}\}$ 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 $\sigma^2_{\mathcal{A}} = \mathbb{E}(d^2_{\mathcal{E}}([B], [\mu_{\mathcal{A}}]) \mid X \in \mathcal{A})$ with X constrained on a subset $\mathcal{A} \subset \mathcal{X}$, which we estimate by $\hat{\sigma}^2_{\mathcal{A}} = 1 - \hat{\lambda}_{\mathcal{A},1}(\sum_{k=1}^{m} \hat{\lambda}_{\mathcal{A},k})^{-1}$ with $\hat{\lambda}_{\mathcal{A},1}, \ldots, \hat{\lambda}_{\mathcal{A},m}$ 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 $\check{\sigma}^2_{\mathcal{A}} = 1 - \hat{\lambda}_{\mathcal{A},1}$ 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 $\hat{\sigma}^2_{\mathcal{A}}$ 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 $\mathrm{R}^2_{\mathcal{A}_1} = 1 - (|\mathcal{X}|\,\hat{\sigma}^2_{\mathcal{X}})^{-1}|\mathcal{A}_2| \sum_{a\in\mathcal{A}_1} \hat{\sigma}^2_{\{a\}\times\mathcal{A}_2}$ 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 ($\mathrm{R}^2_{\mathcal{V}} = 0.68$, $\mathrm{R}^2_{\mathcal{C}} = 0.11$), which increases only to $\mathrm{R}^2_{\mathcal{V}\times\mathcal{C}} = 0.73$ when also distinguishing consonants. Comparing the different vowel contexts, we observe nearly double variation for $\mathtt{a}*\mathtt{a}$ than for $\mathtt{i}*\mathtt{i}$ with $\hat{\sigma}^2_{\{\mathtt{a}*\mathtt{a}\}\times\mathcal{C}\times\mathcal{S}}/\hat{\sigma}^2_{\{\mathtt{i}*\mathtt{i}\}\times\mathcal{C}\times\mathcal{S}} = 1.95$, which might potentially relate to different pronun-

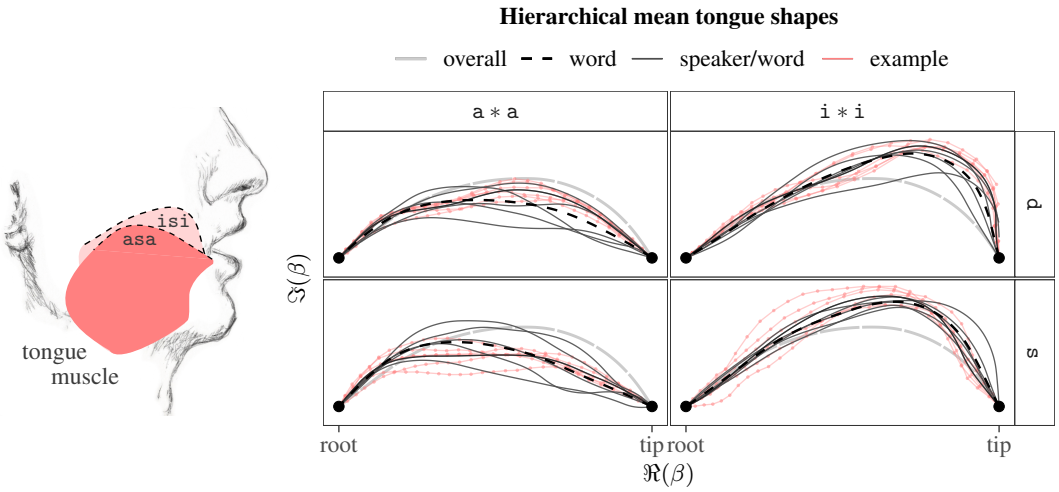

Fig. 2: *Left*: schematic illustrating the tongue muscle modulation when pronouncing "isi" and "asa". Dashed lines correspond to the respective mean shapes in the right plot. With its multiple and multi-directional fibers, the tongue muscle almost fills the entire oral cavity and can flexibly adjust 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. 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"). Tongue shapes are depicted in Bookstein coordinates, i.e. with the tongue roots at $\beta(0) = 0$ and the tongue tips at $\beta(1) = 1$.

ciations of "a" in German dialects. When considering single word articulation of a speaker ($R^2_{\mathcal{V}\times\mathcal{C}\times\mathcal{S}} = 0.93$) 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.

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 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.

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.

ACKNOWLEDGEMENT

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. & 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. & WOOD, A. T. (2014). Mean shapes, projections and intrinsic limiting distributions. *Journal of Statistical Planning and Inference* **145**, 25–32.

DRYDEN, I. L. & MARDIA, K. V. (2016). *Statistical Shape Analysis: With Applications in R*. John Wiley & Sons.

GREVEN, S. & SCHEIPL, F. (2017). A general framework for functional regression modelling (with discussion and rejoinder). *Statistical Modelling* **17**, 1–35 and 100–115.

HAPP, C. & 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. & 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. & 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. & SILVERMAN, B. W. (2005). *Functional Data Analysis*. Springer New York.

REISS, P. T. & XU, M. (2020). Tensor product splines and functional principal components. *Journal of Statistical Planning and Inference* **208**, 1–12.

RYNNE, B. & YOUNGSON, M. A. (2007). *Linear functional analysis*. Springer Science & Business Media.

SRIVASTAVA, A., KLASSEN, E., JOSHI, S. H. & 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. & KLASSEN, E. P. (2016). *Functional and Shape Data Analysis*. Springer-Verlag.

STEYER, L., STÖCKER, A. & GREVEN, S. (2021). Elastic analysis of irregularly or sparsely sampled curves. *arXiv preprint arXiv:2104.11039* .

STONE, M., DAVIS, E. P., DOUGLAS, A. S., AIVER, M. N., GULLAPALLI, R., LEVINE, W. S. & 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 ed.

WOOD, S. N., PYA, N. & 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. & 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*. Springer.

8. Elastic Shape Regression for Plane Curves

This contribution combines previous work in Chapters 5 and 6 to obtain elastic shape regression for plane curve responses based on the SRV-framework, allowing to model and systematically interpret various types of (non-)linear effect types of scalar covariates. The proposed approach generally covers open curves, but also closed curves when modeling axial symmetric curves as often required in practice. In the second case, which is illustrated in an analysis of bottle design, implemented symmetry constraints promote practically satisfactory results despite ignoring non-linear closedness constraints for computational simplification.

Contributing article:
Stöcker, A., Steyer, L., and Greven, S. (2022). Elastic Shape Regression for Plane Curves. *Unpublished manuscript.*

Declaration on personal contributions:
Major parts of this project were conducted by the author of this thesis with important advise and discussion from Sonja Greven and Lisa Steyer. Lisa Steyer in particular also provided helpful discussion on closure of symmetric curves and the proofs asociated therewith.

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 down 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 $y : \mathcal{I} \to \mathbb{C}$ 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* $\gamma : \mathcal{I} \to \mathcal{I}$ 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 $[y]_\mathrm{w}$ for the oriented curve described by a parameterized curve y for clarity. Mapping into an arbitrary coordinate system, the shape $[y]_\mathrm{s}$ of y is defined as its equivalence class $[y]_\mathrm{s} = \{\lambda \exp(\sqrt{-1}\omega)y + z \mid \lambda > 0, \omega \in \mathbb{R}, z \in \mathbb{C}\}$ over re-scaling by λ, rotation by ω radian, and translation by z. The definition directly carries over to the shape of $[y]_\mathrm{w}$ as union over its representatives $[y] = \bigcup_{y \in [y]_\mathrm{w}} [y]_\mathrm{s}$ 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 an surjective map from $\mathcal{AC}^*(\mathcal{I})$ to $\mathbb{L}^2_\mathbb{C}(\mathcal{I})$, or briefly $\mathbb{L}^2_\mathbb{C}$, 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 $[y]_\mathrm{s}$ with $[q]_\mathrm{s} = \{\lambda^2 \exp(\sqrt{-1}\omega)q \mid \lambda > 0, \omega \in \mathbb{R}\}$ on SRV-level. The quotient space of such $[q]_\mathrm{s}$ with $q \neq 0$ corresponds to the complex projective space $\mathrm{P}\mathbb{L}^2_\mathbb{C}$ 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

$$g(\mu) = f(\boldsymbol{x}) = f_1(\boldsymbol{x}) + \cdots + f_J(\boldsymbol{x})$$

consists of three components: a target parameter μ of the distribution of Y depending on covariate values $\boldsymbol{x}$, an additive predictor $f(\boldsymbol{x}) = \sum_{j=1}^{J} f_j(\boldsymbol{x})$, 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 $\boldsymbol{X}$ are assumed a random vector of scalar covariates $\boldsymbol{x} \in \mathcal{X}$ 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

$$\mathbb{E}\left(d^2(\mu, Y) \mid \boldsymbol{X} = \boldsymbol{x}\right) = \sigma^2_{\boldsymbol{x}} = \inf_{\mu' \in \mathcal{M}} \mathbb{E}\left(d^2(\mu', Y) \mid \boldsymbol{X} = \boldsymbol{x}\right)$$

assuming finite variance(s) $\sigma^2_{\boldsymbol{x}} < \infty$ 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 $f : \mathcal{X} \to \mathcal{V}$ 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 $v_1, \ldots, v_K \in \mathcal{V}$. This lets us follow an analogous approach to Brockhaus et al. (2015); Scheipl et al. (2016) for functional data, modeling covariate effect functions $f_j(\boldsymbol{x})$ as

$$f_j(\boldsymbol{x}) = \sum_{k=1}^{K} \sum_{h=1}^{H} \theta_{jhk} b_{jh}(\boldsymbol{x}) v_k$$

expanded in a finite tensor-product basis of the basis $\{v_k\}_k$ and some effect basis $b_{jh} : \mathcal{X} \to \mathbb{R},\ h = 1, \ldots, H$. Estimating $f_j(\boldsymbol{x})$ then reduces to estimating the $H \times K$ coefficient matrix $\boldsymbol{\Theta}_j = \{\theta_{jhk}\}_{h,k}$. 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 $\boldsymbol{x}$ include linear effects $f_j(\boldsymbol{x}) = \beta x_1$ (specifying $b_{j1}(\boldsymbol{x}) = x_1$, $H = 1$) and smooth spline effects with $\{b_{jh}\}_h$, say, a B-spline basis, where coefficient β like all $f_j(\boldsymbol{x})$ here is an element of $\mathcal{V}$. Effects of a categorical covariate $x_2 \in \{1, \ldots, L\}$ are implemented by mapping the lth 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 $g^{-1} : \mathcal{V} \to \mathcal{M}$ 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 $\mathrm{Exp}_p : T_p\mathcal{Y} \to \mathcal{Y}$ 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 $\mu = \mathrm{Exp}_p(\beta x_1)$ 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 $y_1, \ldots, y_n \in \mathcal{AC}^*(\mathcal{I})$ recorded together with vectors of scalar covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. We model the conditional Fréchet mean $[\mu_i]$ of their shapes $[y_i], i = 1, \ldots, n$, considering the $([y_i], \boldsymbol{x}_i)$ 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

$$d([y_1], [y_2]) = \inf_{\gamma \in \Gamma} d_{\mathrm{P}\mathbb{L}^2_{\mathbb{C}}}([q_1]_{\mathrm{s}}, [q_2 \circ \gamma \sqrt{\dot{\gamma}}]_{\mathrm{s}}) = \inf_{\gamma \in \Gamma, \omega \in \mathbb{R}} d_{\mathbb{S}}(q_1, \exp(\omega\sqrt{-1}) q_2 \circ \gamma \sqrt{\dot{\gamma}})$$

by the geodesic distance $d_{\mathrm{P}\mathbb{L}^2_{\mathbb{C}}}$ on the complex projective space of the $[q_i]_{\mathrm{s}}$, 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 $\gamma : \mathcal{I} \to \mathcal{I}$. 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 $\mathrm{P}\mathbb{L}^2_{\mathbb{C}}$ is in turn induced from the submanifold geometry of the Hilbert sphere $\mathbb{S} = \{q \in \mathbb{L}^2_{\mathbb{C}} \mid \|q\| = 1\} \subset \mathbb{L}^2_{\mathbb{C}}$, where $\|q\| = (\int_{\mathcal{I}} |q(t)|^2 \, dt)^{1/2}$ denotes the standard norm on $\mathbb{L}^2_{\mathbb{C}}$. 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 $[y_i]_{\mathrm{w}}$ to unit-length. Due to the SRV-representation, not only rotation by ω radian but also re-parameterization by $\gamma \in \Gamma$ acts by isometries, i.e. for common actions $\exp(\omega\sqrt{-1})\, y_i \circ \gamma$, $i = 1, 2$, the $\mathbb{L}^2_{\mathbb{C}}$ inner product $\langle q_1, q_2 \rangle = \langle \exp(\omega\sqrt{-1})\, q_1 \circ \gamma \sqrt{\dot{\gamma}}, \exp(\omega\sqrt{-1})\, q_2 \circ \gamma \sqrt{\dot{\gamma}} \rangle$ 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 ith 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

$$[\mu_i] = g_{[\psi]}^{-1}\big(f(\boldsymbol{x}_i)\big) = g_{[\psi]}^{-1}\Big(\sum_{j=1}^{J} f_j(\boldsymbol{x}_i)\Big)$$

induced by the Riemannian (inelastic) functional additive model

$$m_i = \mathrm{Exp}_p\big(f(\boldsymbol{x}_i)\big)$$

on SRV-level: we choose the Riemannian exponential $\mathrm{Exp}_p(\beta) = \cos(\|\beta\|)p + \sin(\|\beta\|)\beta/\|\beta\|$ on $\mathbb{S}$ as response function mapping the additive predictor $f(\boldsymbol{x})$ 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 $\mathrm{P}\mathbb{L}^2_{\mathbb{C}}$ and lets us identify $T_{[p]_\mathrm{s}}\mathrm{P}\mathbb{L}^2_{\mathbb{C}}$ with the subspace $\mathcal{V}_p = \{q \in \mathbb{L}^2_{\mathbb{C}} \mid \langle q, p\rangle = 0\}$ orthogonal to $p \in \mathbb{S}$ (compare, e.g., Stöcker et al., 2022; Dryden and Mardia, 2016; Klingenberg, 1995). Thus, common basis functions $\tilde{v}_k : \mathcal{I} \to \mathbb{R}$, $k = 1, \ldots, K+1$, used for functional additive models (Scheipl et al., 2015), such as polynomial splines, can be utilized for constructing tensor-product effects $f_j(\boldsymbol{x})$ after linear transformation to a constrained basis v_k , $k = 1, \ldots, K$, 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 $f_1, \ldots f_J$ 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 $g_{[\psi]}([\mu]) = \mathrm{Log}_p(m)$ 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 $[y_1], \ldots, [y_n]$. Correspondingly, effects $f_j(\boldsymbol{x})$ are typically constrained to be centered to zero mean $\sum_{i=1}^n f_j(\boldsymbol{x}_i) = 0$. 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 $f : [-1, 1] \to \mathbb{C}$ 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 $v_k^{\Re} : [-1, 1] \to \mathbb{R}$ for its real part and odd basis functions $v_k^{\Im} : [-1, 1] \to \mathbb{R}$ 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 $[\mu]_\mathrm{w}$ is closed if any and hence all $\mu \in [\mu]_\mathrm{w}$ 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 $\mathrm{P}\mathbb{L}^2_{\mathbb{C}}$ 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 $i = 1, \ldots, n$ analogously to the Riemannian case. After initialization, the proposed boosting algorithm (Algorithm 1) repeatedly adds to the model predictor $\hat{f}(\boldsymbol{x})$ by iteratively A) computing warping-aligned transported residuals, B) fitting them with the base-learners corresponding to predictor components $f_j(\boldsymbol{x})$, 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 $\hat{f}(\boldsymbol{x}) = 0$;
repeat
  A) Computing residuals:
  foreach $i = 1, \ldots, n$ **do**
    Predict mean shape representative $\hat{\mu}_i$ based on current predictor $\hat{f}(\boldsymbol{x}_i)$;
    Warping-align $y_i \overset{\text{align to } \hat{\mu}_i}{\longmapsto} \tilde{y}_i$;
    Map $\tilde{y}_i \overset{\text{SRV-trafo}}{\longmapsto} \tilde{q}_i \overset{\text{Log}}{\longmapsto} \tilde{\varepsilon}_i \overset{\text{Transp}}{\longmapsto} \varepsilon_i$ to transported residual $\varepsilon_i \in T_{[p]_\text{s}} \text{P}\mathbb{L}^2_{\mathbb{C}}$;
  end
  B) Fitting baselearners:
  foreach $j = 1, \ldots, J$ **do**
    Fit jth base-learner to residuals ε_i, $i = 1 \ldots, n$, to obtain $\check{f}_j(\boldsymbol{x})$;
    Determine insample performance ;
  end
  C) Updating the predictor:
  Set $\hat{f}(\boldsymbol{x}) \leftarrow \hat{f}(\boldsymbol{x}) + \eta \check{f}_j(\boldsymbol{x})$ 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 $\mathbb{L}^2_{\mathbb{C}}$-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 prameterized curves $[y_i]_\text{s}$ predicted as $[\hat{\mu}_i]_\text{s}$ 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 $\tilde{\epsilon}_i \in T_{[\hat{m}_i]_\text{s}} \text{P}\mathbb{L}^2_{\mathbb{C}}$ in the (linear) tangent space is obtained as $\tilde{\epsilon}_i = \text{Log}_{[\hat{m}_i]_\text{s}}([q_i]_\text{s})$ from the SRV-transform q_i of y_i. Due to the geometry of $\text{P}\mathbb{L}^2_{\mathbb{C}}$, 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 $\tilde{q}_i \in [q_i]_\text{s}$ and $\hat{m}_i$ are rotation-aligned (compare, e.g., Huckemann et al., 2010). The local residuals reflect the distance $\|\tilde{\epsilon}_i\| = d([\hat{m}]_\text{s}, [q_i]_\text{s})$ and correspond to the negative gradient $\tilde{\epsilon}_i = -\nabla_{[\hat{m}]_\text{s}} d^2([\hat{m}]_\text{s}, [q_i]_\text{s})$ pointing into the direction of loss-reduction (Pennec, 2006). However, for $i = 1, \ldots, n$, they are elements of different spaces. Parallel-transport $\text{Transp}_{[\hat{m}_i]_\text{s},[p]_\text{s}} : T_{[\hat{m}_i]_\text{s}} \text{P}\mathbb{L}^2_{\mathbb{C}} \to T_{[p]_\text{s}} \text{P}\mathbb{L}^2_{\mathbb{C}}$ isometrically maps the local residuals to transported residuals $\varepsilon_i = \text{Transp}_{[\hat{m}_i]_\text{s},[p]_\text{s}}(\tilde{\varepsilon}_i)$ in the space $\mathcal{V}_p \cong T_{[p]_\text{s}} \text{P}\mathbb{L}^2_{\mathbb{C}}$ 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 $[\tilde{y}_i]_\text{s}$ 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 $\| \text{Log}_p(\hat{m}_i) - \epsilon_i \|^2 = d^2_{\text{P}\mathbb{L}^2_{\mathbb{C}}}([\hat{m}_i]_\text{s}, [\tilde{q}_i]_\text{s}) \approx d^2([\hat{\mu}_i], [y_i])$ 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 $f_j(\boldsymbol{x})$, $j = 1, \ldots, J$, 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 $\boldsymbol{x}_i \in \mathcal{X}$, $i = 1, \ldots, n$. As elements of $\mathcal{V}_p$, they are fitted with respect to the penalized least-squares criterion to obtain $\check{f}_j = \arg\min_{f_j} \sum_{i=1}^n \|f_j(\boldsymbol{x}_i) - \epsilon_i\|^2 + \text{pen}_j(f_j)$. Using tensor-product effects $f_j(\boldsymbol{x}) = \sum_{k=1}^K \sum_{h=1}^H \theta_{jhk} b_{jh}(\boldsymbol{x}) v_k$ 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 $j = 1, \ldots, J$. The in-sample performance of the jth base-learner is then measured in terms of its residual sum of squares $\text{RSS}_j = \sum_{i=1}^n \|\check{f}_j(\boldsymbol{x}_i) - \epsilon_i\|^2$.

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, $i = 1, \ldots, n$, are recorded at discrete sampling points and computations involving $\mathbb{L}^2_{\mathbb{C}}$ 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, $i = 1, \ldots, n$, 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 $\boldsymbol{x}_i = (x_{\texttt{type},i}, x_{\texttt{size},i})^\top$) as

$$[\mu_i] = g_{[p]}^{-1}(\alpha_{\texttt{type},i} + \beta x_{\texttt{size},i} + \beta_{\texttt{type}} x_{\texttt{size},i} + f(x_{\texttt{size},i}) + f_{\texttt{type}}(x_{\texttt{size}_i}))$$

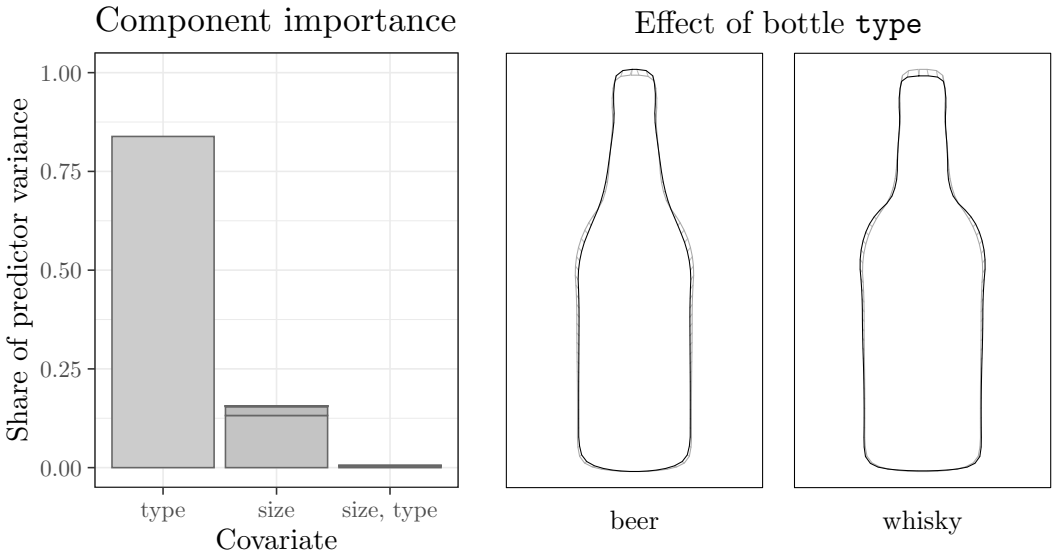

Figure 1: *Left:* Shares $\sum_{i=1}^{n} \left(\hat{f}_j^{[k]}(\boldsymbol{x}_i)\right)^2 / \sum_{i=1}^{n} \sum_{j=1}^{J} \|\hat{f}_j(\boldsymbol{x}_i)\|^2$ of the variance of each (centered) factorized effect component $\hat{f}_j^{[k]}(\boldsymbol{x})$ selected into the model in overall predictor variance. Bars for its factorization components are stacked for each base-learner. *Right:* Estimated elastic mean shape of beer and whisky bottles setting `size`-effects $\hat{f}(x_{\texttt{size}}) + \hat{f}_{\texttt{type}}(x_{\texttt{size}}) = 0$. 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

$$m_i = \text{Exp}_p(\alpha_{\texttt{type},i} + \beta x_{\texttt{size},i} + \beta_{\texttt{type}} x_{\texttt{size},i} + f(x_{\texttt{size},i}) + f_{\texttt{type}}(x_{\texttt{size},i}))$$

of μ_i with an effect-coded binary effect $x_{\texttt{type}} \mapsto \alpha_{\texttt{type}} \in \mathcal{V}_p$ and, for `size`, a linear effect with coefficient β and a smooth effect $f(x_{\texttt{size}})$ centered around the linear effect, as well as their interactions with `type`. The effect functions f and $f_{\texttt{type}}$ 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 $\hat{f}(x_{\texttt{type}}, x_{\texttt{size}}) = \hat{\alpha}_{\texttt{type}} + \hat{f}(x_{\texttt{size}}) + \hat{f}_{\texttt{type}}(x_{\texttt{size}})$ 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 $\hat{f}(x_{\texttt{size}}) = \sum_{k=1}^{K'} \hat{v}^{[k]} \hat{f}^{[k]}(x_{\texttt{size}})$, with K' the minimum of marginal basis

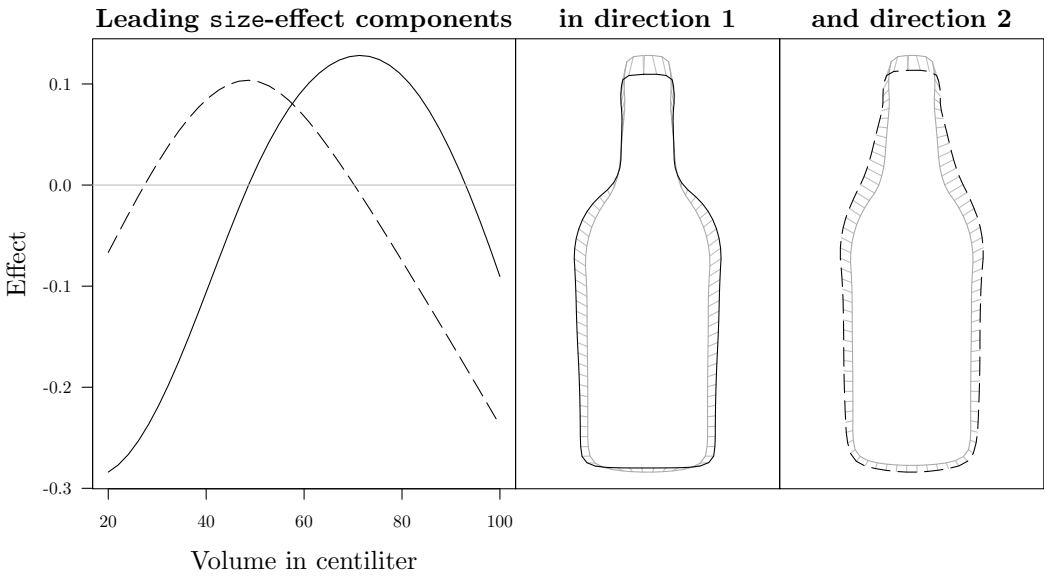

Figure 2: The two leading size-effect components $\hat{f}^{[1]}(x_{\texttt{size}})$ (*black solid lines*) and $\hat{f}^{[2]}(x_{\texttt{size}})$ (*black dashed lines*), explaining around 13.2% and 2.2% of the predictor variance respectively, depicted together with their respective directions $\hat{v}^{[1]}$ 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 $\hat{f}^{[k]}(x_{\texttt{size}}) = 1$, $k = 1, 2$.

dimensions, into independent effect components $\hat{f}^{[k]} : [0, 100] \to \mathbb{R}$ presenting scalar effects into orthogonal effect directions $\hat{v}^{[k]} \in T_{[p]}\text{P}\mathbb{L}^2_{\mathbb{C}}$ sorted with decreasing effect variance $\frac{1}{n}\sum_{i=1}^{n}(\hat{f}^{[k]}(x_{\texttt{size}}))^2$ 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 $\hat{f}^{[1]}(x_{\texttt{size}}) > 0$ or < 0, respectively. A positive or negative second component $\hat{f}^{[2]}(x_{\texttt{size}})$ 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 ($x_{\texttt{size}} \in [25, 75]$, average $\bar{x}_{\texttt{size}} \approx 42$ centiliter) and whisky bottles ($x_{\texttt{size}} \in [70, 100]$, $\bar{x}_{\texttt{size}} \approx 73$) 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 capturs 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 $y_1, \ldots, y_n$. 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 orginal 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 extend 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 $d^2([\hat{\psi}], [\psi]) << \sigma_0^2$ 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

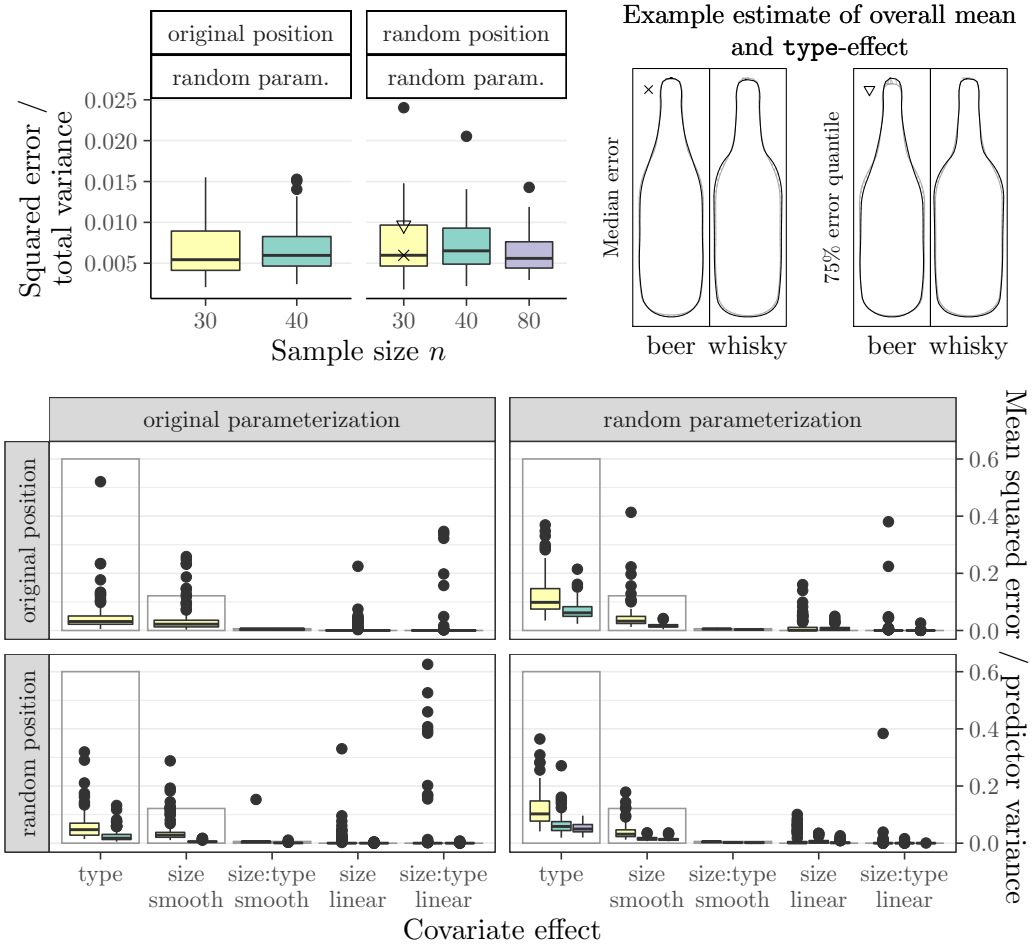

Figure 3: *Top left:* Distributions of intercept (overall mean) estimation accuracy (squared errors $d^2([\hat{\psi}], [\psi])/\sigma_0^2$ relative to total variance) for simulation scenarios with random warping of bottle outline representatives $y_1, \ldots, y_n$. For concise display, relative errors 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 $\triangle$ in the plot on the left. *Bottom:* MSE distributions for covariate effects in the model relative to the overall variance $\frac{1}{n}\sum_{i=1}^{n} \|f(\boldsymbol{x}_i)\|^2$ 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 $\frac{1}{n}\sum_{i=1}^{n} \|f_j(\boldsymbol{x}_i)\|^2$, $j = 1, \ldots, n$, 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 $\mathbb{L}^2_{\mathbb{C}}$-average of the SRV-transforms $q_1, \ldots, q_n$ 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 $y_1, \ldots, y_n$ 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 differentiably 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 $\mu\big|_{[0,1]}$ and $\mu\big|_{[-1,0]}$ 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 $\mu^* : [-1 - \frac{\delta}{2\dot{\mu}(1)}, 1 + \frac{\delta}{2\dot{\mu}(1)}] \to \mathbb{C}$ 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 $\mu\big|_{[0,1]}$ and $\mu\big|_{[-1,0]}$ of the symmetric curve can be shifted to close the curve at -1/1 while opening it at 0. Then, we may differentiably extend them at 0 to obtain a closed curve $\mu^* : [-1 + \frac{\delta}{2\dot{\mu}(0)}, 1 - \frac{\delta}{2\dot{\mu}(0)}] \to \mathbb{C}$ 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 re-parameterization 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 differentiably 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 $y_1, \ldots, y_n$ 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 $m : [-1,1] \to \mathbb{C}$ 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 $m(1)^\dagger \overset{\text{even}}{=} m(-1) \overset{\text{closed}}{=} m(1)$. □

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 $y_i : [0,1] \to \mathbb{C}$, given by the sample polygon of the ith 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 $0 = \gamma_i(t_1) < \cdots < \gamma_i(t_{100}) = t_{100}$ are randomly drawn independently for the ith curve in the simulated data corresponding to the ith curve in the original dataset. In scenarios with random positioning, we draw

$$\begin{aligned}&\lambda_i \sim \text{Gamma}(100, 100) \quad \text{(given with shape and rate parameter)} \\ &\omega_i \sim \text{N}(0, \frac{\pi^2}{400}), \quad z_i^{\Re} \sim \text{N}(0, \sigma_{\Re}^2), \quad z_i^{\Im} \sim \text{N}(0, \sigma_{\Im}^2)\end{aligned}$$

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 $\sigma_{\Re}^2$ and $\sigma_{\Im}^2$ 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 $\text{sd}(\Delta_i) = \frac{1}{3}$. 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.

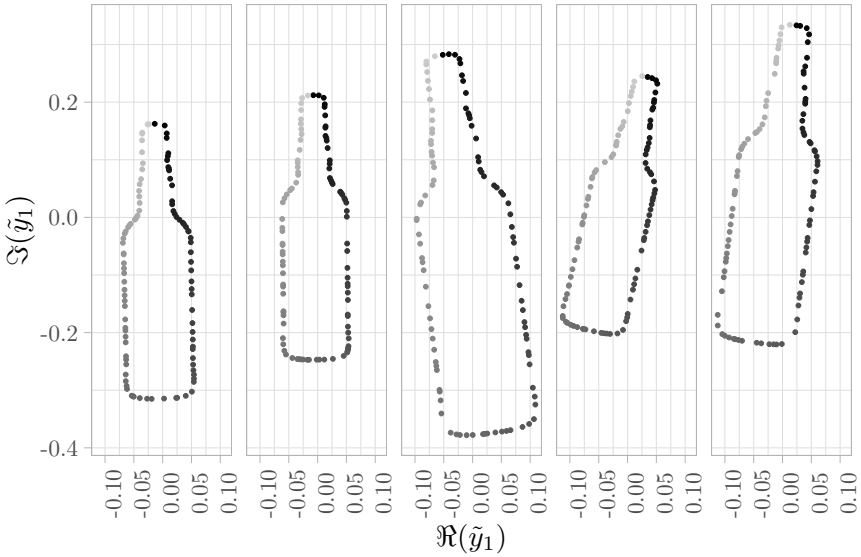

Figure 4: Five example draws of the first curve in the original dataset applying random warping and positioning. The parameterization of the 100 sample points per curve is always initialized with a regular grid starting and ending at the top of the bottle.

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 (mglm) 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. Chkrebtii, 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.

Appendices for Selected Contributions

A. Supplementary material for Chapter 2 "Boosting Functional Response Models for Location, Scale and Shape with an Application to Bacterial Competition"

Online supplement for the contribution:
Stöcker, A., Brockhaus, S., Schaffer, S., 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.
DOI: 10.1177/1471082X20917586.

Web-based supporting materials for Boosting Functional Response Models for Location, Scale and Shape with an Application to Bacterial Competition

Almond Stöcker [1], Sarah Brockhaus [2], Sophia Anna Schaffer [3], Benedikt von Bronk [3], Madeleine Opitz [3] and Sonja Greven [1]

[1] School of Business and Economics, HU Berlin, Germany

[2] Department of Statistics, LMU Munich, Germany

[3] Faculty of Physics, LMU Munich, Germany

Address for correspondence: Almond Stöcker, School of Business and Economics, Chair of Statistics, Humboldt University of Berlin, Unter den Linden 6, D–100 99 Berlin, Germany.

E-mail: almond.stoecker@hu-berlin.de.

Phone: (+49) 30 2093 99554.

Fax: (+49) 30 2093 99591.

Abstract: This Online Supplement contains supplementary material in four different respects: Appendix A provides more details concerning the construction of partial effect functions from tensor product bases, as discussed in Section 2 of the main manuscript; Appendix B introduces the technical background necessary to draw appropriate smooth random response curves and random effect functions in the GAMLSS-scenario for both simulation studies presented in Section 4; Appendix C gives more detailed insights into the setup of the simulation studies, the used measures of estimation quality, and the obtained results; and Appendix E contains additional information concerning the analysis of bacterial interaction in Section 3, i.e. data details, a comparison to other growth models in the literature and further model results.

Key words: Bacterial Growth, Distributional Regression, Functional data, Functional Regression, GAMLSS

A Basis representations and orthogonalization

A.1 Tensor product bases

For both fitting and prediction, it is necessary to evaluate effect functions h_j, as defined in Section 2.1, at all data points simultaneously. Following Scheipl et al.

(2015), this can be written in a simple closed form expression using row tensor products.

Definiton Consider an $n \times m$ matrix $\mathbf{A}$ and an $r \times s$ matrix $\mathbf{B}$. The *Kronecker product* '$\otimes$' is defined as

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{1,1}B & \dots & a_{1,m}B \\ \vdots & \ddots & \vdots \\ a_{n,1}B & \dots & a_{n,m}B \end{bmatrix}$$

If $\mathbf{A}$ and $\mathbf{B}$ have the same number of rows $n = r$, the *row tensor-product* '$\odot$' is defined by

$$\mathbf{A} \odot \mathbf{B} = \begin{bmatrix} \mathbf{a}_1 \otimes \mathbf{b}_1 \\ \vdots \\ \mathbf{a}_n \otimes \mathbf{b}_n \end{bmatrix},$$

where $\mathbf{a}_i$ and $\mathbf{b}_i$ are the i-th rows of the matrices $\mathbf{A}$ and $\mathbf{B}$.

Consider two variables $t \in T$ and $\mathbf{x} \in \mathcal{X}$. Let $h_j(\mathbf{x}, t) = (\mathbf{b}_{Xj}(\mathbf{x}, t) \otimes \mathbf{b}_{Yj}(t))^\top \boldsymbol{\theta}_j$ with parameter vector $\boldsymbol{\theta}_j$ and function basis vectors $\mathbf{b}_{Xj}(\mathbf{x}, t)$ and $\mathbf{b}_{Yj}(t)$ as defined in Section 2.1. For a data set with N observations and G measurements per response curve, let covariates $\mathbf{X} = \left(\mathbf{x}_{1,1}, \dots, \mathbf{x}_{1,G}, \mathbf{x}_{2,1}, \dots, \mathbf{x}_{N,G}\right)^\top$ and respective

time points $\mathbf{t} = \left(t_{1,1}, \ldots, t_{1,G}, t_{2,1}, \ldots, t_{N,G}\right)^\top$. Define

$$\mathbf{B}_{Xj}(\mathbf{X}, \mathbf{t}) = \begin{bmatrix} \mathbf{b}_{Xj}^\top(\mathbf{x}_{1,1}, t_{1,1}) \\ \vdots \\ \mathbf{b}_{Xj}^\top(\mathbf{x}_{N,G}, t_{N,G}) \end{bmatrix}$$

the $NG \times K_j$ design matrix. Analogously, define the vector $\mathbf{H}_j(\mathbf{X}, \mathbf{t})$ of length NG to entail the evaluations of h_j and the $NG \times K_Y$ design matrix $\mathbf{B}_{Yj}(\mathbf{t})$ with the evaluations of $\mathbf{b}_{Yj}$. Then, we can write the joint evaluation as equation

$$\mathbf{H}_j(\mathbf{X}, \mathbf{t}) = \left(\mathbf{B}_{Xj}(\mathbf{X}, \mathbf{t}) \odot \mathbf{B}_{Yj}(\mathbf{t})\right)^\top \boldsymbol{\theta}_j \tag{A.1}$$

in terms of the row tensor product. In the same way, covariate effect bases of two different covariates can be combined to an interaction effect. Note that a common number of measurements per curve G is only assumed for ease of notation.

If all response curves are observed on a common grid $\mathbf{t}_0 = \left(t_1, ..., t_G\right)^\top$ and covariates and their effect bases are not time dependent, equation (A.1) can be re-written in terms of the Kronecker product. In this case, we can denote covariates as $\mathbf{X}_0 = \left(\mathbf{x}_1, \ldots, \mathbf{x}_N\right)^\top$. Then the covariate part simplifies to an $N \times K_j$ matrix $\mathbf{B}_{Xj}(\mathbf{X}_0) = \mathbf{B}_{Xj}(\mathbf{X}_0, \mathbf{t}_0)$ that is independent of $\mathbf{t}_0$ and also the $\mathbf{B}_{Yj}(\mathbf{t}_0)$ can be

arranged as a $G \times K_Y$ matrix. For these matrices, we obtain

$$\mathbf{H}_j(\mathbf{X}, \mathbf{t}) = \left(\mathbf{B}_{Xj}(\mathbf{X}_0) \otimes \mathbf{B}_{Yj}(\mathbf{t}_0)\right)^\top \boldsymbol{\theta}_j \,. \tag{A.2}$$

The Kronecker product has desirable mathematical properties, which simplify the implementation of linear constraints in Section A.2. In particular, it enables us to consider the model as Functional Linear Array Model (Brockhaus et al., 2015), which increases computational efficiency. In terms of Generalized Linear Array Models (Currie et al., 2006), the model can be fitted without actually computing the Kronecker product.

A.2 Othogonalization of effect functions

Consider two effect functions $h_1(\mathbf{x}, t) = \mathbf{b}_1^\top(\mathbf{x}, t)\boldsymbol{\theta}_1$ and $h_2(\mathbf{x}, t) = \mathbf{b}_2^\top(\mathbf{x}, t)\boldsymbol{\theta}_2$ with function bases $\mathbf{b}_1$ and $\mathbf{b}_2$ of dimension K_1 and K_2. Assume constant functions are entailed in $span(\mathbf{b}_2)$, i.e. $\exists \boldsymbol{\theta}_c : \mathbf{b}_2(\mathbf{x}, t)\boldsymbol{\theta}_c = 1 \; \forall \mathbf{x}, t$. Assume the same for $\mathbf{b}_1$. This property typically holds for, e.g., spline bases. We want to construct an interaction effect in terms of the Kronecker product basis $\mathbf{b}_1(\mathbf{x}, t) \otimes \mathbf{b}_2(\mathbf{x}, t)$. Applying $\boldsymbol{\theta}_c$ from above we get

$$h_1(\mathbf{x}, t) = \mathbf{b}_1^\top(\mathbf{x}, t)\boldsymbol{\theta}_1 = \mathbf{b}_1^\top(\mathbf{x}, t)\boldsymbol{\theta}_1 \otimes \mathbf{b}_2^\top(\mathbf{x}, t)\boldsymbol{\theta}_c = \left(\mathbf{b}_1(\mathbf{x}, t) \otimes \mathbf{b}_2(\mathbf{x}, t)\right)^\top (\boldsymbol{\theta}_1 \otimes \boldsymbol{\theta}_c) \,.$$

This is analogously obtained for h_2. Hence, $h_1, h_2 \in span(\mathbf{b}_1 \otimes \mathbf{b}_2)$. However, for separate model fitting, interpretation and automatic model selection, we want to

include marginal effects h_1, h_2 and their interaction h_{12} as distinct effects into the additive predictor. To do so, we have to construct linearly independent design matrices:

For observed covariates $\mathbf{X}$ and NG time points $\mathbf{t}$, let $\mathbf{B}_1 = \mathbf{B}_1(\mathbf{X}, \mathbf{t})$ and $\mathbf{B}_2 = \mathbf{B}_2(\mathbf{X}, \mathbf{t})$ be the $NG \times K_j$ design matrices as defined in Section A.1. The k-th column of $\mathbf{B}_1$ corresponds to the evaluations of the k-th basis function $b_{1,k}(\mathbf{x}, t)$. The same for $\mathbf{B}_2$. Let $\widetilde{\mathbf{B}} = \mathbf{B}_1 \odot \mathbf{B}_2$ denote the complete tensor product design matrix. Then, we obtain the design matrix $\mathbf{B}$ for the interaction effect via a linear transform $\mathbf{B} = \widetilde{\mathbf{B}}\mathbf{Z}$. The transformation matrix $\mathbf{Z}$ is specified as a matrix with NG columns and with a maximum number of orthogonal rows such that $\mathbf{B}^\top \mathbf{B}_1 = \mathbf{B}^\top \mathbf{B}_2 = \mathbf{0}$. This means, we construct $\mathbf{B}$ such that it is orthogonal to the design matrices of the marginal effects. Applying QR-decomposition to the matrix $\mathbf{C} = \widetilde{\mathbf{B}}^\top \left[\mathbf{B}_1 : \mathbf{B}_2\right]$ a suitable matrix $\mathbf{Z}$ is determined by

$$\mathbf{C} = \left[\mathbf{Q} : \mathbf{Z}\right] \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix} = \mathbf{QR}\,,$$

since $\mathbf{B}^\top \left[\mathbf{B}_1 : \mathbf{B}_2\right] = \mathbf{Z}^\top \widetilde{\mathbf{B}}^\top \left[\mathbf{B}_1 : \mathbf{B}_2\right] = \mathbf{Z}^\top \mathbf{QR} = \mathbf{0}$, i.e. $\mathbf{B}$ is orthogonal to $\mathbf{B}_1$ and $\mathbf{B}_2$ (Wood, 2006; Brockhaus et al., 2015). According to Brockhaus et al. (2015), we proceed just the same way in order to distinguish effect functions $h_j(\mathbf{x}, t) = (\mathbf{b}_j(\mathbf{x}, t) \otimes \mathbf{b}_Y(t))^\top \boldsymbol{\theta}_j$ from the functional linear intercept $h_0(\mathbf{x}, t) = \mathbf{b}_Y^\top(t)\boldsymbol{\theta}_0$, which is a special case of the above. Computation can be further simplified in

the linear array case.

B Simulation details I: generation of random smooth errors and effect functions

In order to sample appropriate data for a simulation study concerning functional GAMLSS, we have to draw appropriate smooth random response curves and, at the same time, we do not only have to flexibly control the mean, but also the variance and other distributional parameters over time. In the Gaussian response simulation study, we additionally sample the true underlying effect functions to increase representativeness. For both purposes, we have to control the smoothness of the randomly drawn curves in order to achieve realistic samples. This section introduces the technical framework used to implement these points.

B.1 Random spline generation

The simulation relies crucially on random spline generation. Given a closed interval $T \subset \mathbb{R}$, a random spline $r : T \to \mathbb{R}$, $t \mapsto \mathbf{b}^\top(t)\boldsymbol{\theta}$ is understood as the product of a fixed spline basis vector $\mathbf{b}(t) = \left(b_1(t), \ldots, b_K(t)\right)^\top$ and a random vector $\boldsymbol{\theta}$ of K coefficients. In the following, $\boldsymbol{\theta}$ is always $\mathcal{N}(\mathbf{0}, \mathbf{I}_K)$ distributed, with any

necessary transformation subsumed into the basis $\mathbf{b}(t)$, i.e. $\mathbf{b}(t)$ is obtained by transforming an underlying prototype basis $\tilde{\mathbf{b}}(t) = \left(\tilde{b}_1(t), \ldots, \tilde{b}_K(t)\right)^\top$. For $\tilde{\mathbf{b}}(t)$ we take a B-spline basis of degree l with $K - l + 1$ equally spaced knots. The random splines are scaled employing a scale parameter $\bar{\sigma}^2 \geq 0$. In correspondence to P-splines, a d-th order difference penalty analogue can be specified, which is controlled with a smoothing parameter $\lambda \in [0, 1]$. This is accomplished by specifying the random spline as $r(t) = \mathbf{b}^\top(t)\boldsymbol{\theta} = \tilde{\mathbf{b}}^\top(t)\boldsymbol{\Omega}\mathbf{W}\boldsymbol{\theta}$ with a suitable $K \times K$ orthogonal matrix $\boldsymbol{\Omega}$ and a diagonal weight matrix $\mathbf{W}$ depending on two parameters $\bar{\sigma}^2$ and λ, which will be further clarified below.

In order to adjust the degree of smoothness of randomly drawn splines, we follow the mixed model representation of P-splines (Fahrmeir et al., 2004). Let $\mathbf{P} = \mathbf{D}^\top\mathbf{D}$ be a d-th order difference penalty matrix inducing a quadratic penalty of the form $\tilde{\boldsymbol{\theta}}^\top\mathbf{P}\tilde{\boldsymbol{\theta}}$, with $d \leq l$. Then there is a quadratic orthogonal matrix $\boldsymbol{\Omega}$, such that for $\tilde{\boldsymbol{\theta}} = \boldsymbol{\Omega}\boldsymbol{\theta}$ we obtain $\tilde{\boldsymbol{\theta}}^\top\mathbf{P}\tilde{\boldsymbol{\theta}} = (\boldsymbol{\Omega}\boldsymbol{\theta})^\top\mathbf{P}\boldsymbol{\Omega}\boldsymbol{\theta} = \boldsymbol{\theta}^\top\mathbf{I}_K^{(d)}\boldsymbol{\theta}$, where $\mathbf{I}_K^{(d)}$ is the diagonal matrix with the first d diagonal entries zero and the rest one. Thus, by multiplying with $\boldsymbol{\Omega}^\top$ the basis $\tilde{\mathbf{b}}(t)$ is transformed such that $b_1(t), \ldots, b_d(t)$ represent the unpenalized part of the spline and $b_{d+1}(t), \ldots, b_K(t)$ are subject to a ridge penalty. A suitable transformation matrix is given by $\boldsymbol{\Omega} = \left[\mathbf{L} : \mathbf{D}^\top(\mathbf{D}\mathbf{D}^\top)^{-1}\right]$. The m-th column of the $K \times d$ matrix $\mathbf{L}$ is given by $\mathbf{L}_m = \left(p_{m-1}(1), \ldots, p_{m-1}(K)\right)^\top$ with orthogonal polynomials $p_0, ..., p_{d-1}$ of order $0, ..., d-1$, such that $\mathbf{L}^\top\mathbf{L} = \mathbf{I}_d$. In the mixed model estimation, the penalized coefficients $\theta_{d+1}, ..., \theta_K$ are considered random, whereas the unpenalized coefficients $\theta_d, ..., \theta_K$ are considered

fixed. However, for drawing the whole vector $\boldsymbol{\theta}$, this would lead to a improper distribution. So, we also add some variance to the unpenalized coefficients and control the trade-off between the variance for the penalized and unpenalized parts with the smoothing parameter λ, which corresponds to the part of the variance of the randomly generated spline curve, which is explained by the unpenalized smooth part. Moreover, we want to be able to control the scale of the randomly generated spline curves and, therefore, use the parameter $\bar{\sigma}^2$, which is given by the total variance of the random curves. These two parameters determine $\mathbf{W}$. More precisely, they are defined as follows:

Evaluating the prototype B-spline basis $\tilde{\mathbf{b}}(t)$ on a grid $\mathbf{t} \in T^G$ we obtain a 'design' matrix $\tilde{\mathbf{B}}$ with $\tilde{\mathbf{b}}^\top(t_1), \ldots, \tilde{\mathbf{b}}^\top(t_G)$ as its rows. For a given $\mathbf{W}$, $\mathbf{B} = \tilde{\mathbf{B}}\boldsymbol{\Omega}\mathbf{W}$ presents the corresponding matrix for the desired spline basis $\mathbf{b}(t)$ of the random spline. Drawing $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_K)$, the covariance matrix for the random spline evaluations $\mathbf{r} = \mathbf{B}\boldsymbol{\theta}$ on $\mathbf{t}$ is given by $\mathbf{Cov}(\mathbf{r}) = \mathbf{B}\mathbf{B}^\top$. For the scale parameter we employ the mean variance over time $\bar{\sigma}^2 = \overline{\mathbb{V}\mathrm{ar}}(\mathbf{r}) = \frac{1}{G}\operatorname{tr}(\mathbf{B}\mathbf{B}^\top) = \frac{1}{G}\operatorname{tr}(\mathbf{B}^\top\mathbf{B})$. To control the smoothness of the resulting random spline curves we use $\lambda = 1 - \frac{\operatorname{tr}(\mathbf{B}^\top\mathbf{B}\mathbf{I}_K^{(d)})}{\operatorname{tr}(\mathbf{B}^\top\mathbf{B})}$, which corresponds to the percentage of $\overline{\mathbb{V}\mathrm{ar}}(\mathbf{r})$ explained by the unpenalized part of the spline. It presents a natural parameter of smoothness: Similar to the smoothing parameter λ in penalized regression, a high value of λ corresponds to a high degree of smoothness, because, in this case, most of the variance in the randomly drawn spline curves stems from the smooth unpenalized part. However, λ and λ are not directly mathematically related.

Technically, the above definitions of λ and $\bar{\sigma}^2$ are implemented via specifying the weight matrix $\mathbf{W}^2 = \bar{\sigma}^2 \left(\frac{\lambda}{d\bar{\sigma}^2_{un}} (\mathbf{I}_K - \mathbf{I}_K^{(d)}) + \frac{(1-\lambda)}{(K-d)\bar{\sigma}^2_{pe}} (\mathbf{I}_K^{(d)}) \right)$ with $\bar{\sigma}^2_{un} = \frac{1}{G} \operatorname{tr}\left((\tilde{\mathbf{B}}\boldsymbol{\Omega})^\top \tilde{\mathbf{B}}\boldsymbol{\Omega} (\mathbf{I}_K - \mathbf{I}_K^{(d)}) \right)$ the mean variance of the unpenalized part and $\bar{\sigma}^2_{pe} = \frac{1}{G} \operatorname{tr}\left((\tilde{\mathbf{B}}\boldsymbol{\Omega})^\top \tilde{\mathbf{B}}\boldsymbol{\Omega}\mathbf{I}_K^{(d)} \right)$ the mean variance of the penalized part.

Tensor product random splines depending on multiple variables are generated and orthogonalized as described in Sections A.1 and A.2. For tensor product random splines the total mean variance is adjusted in a further step.

B.2 Sampling smooth response curves

In order to sample response curves with the desired smoothness, i.e. the desired in-curve dependency structure, we sample splines with random coefficients, and then transform them, such that they follow the desired point-wise distribution of the response.

2.2.1 The Gaussian case

We start with the Gaussian case, where we may re-formulate the functional GAMLSS models as

$$y_i(t) = \mu(\mathbf{x}_i, t) + \gamma_i(t) + \varepsilon_{i,t} \tag{B.1}$$

with $t \in T$, covariates $\mathbf{x}_i$, a mean structure $\mu(\mathbf{x}_i, t)$, a smooth random error curve $\gamma_i(t)$ and independent errors $\varepsilon_{i,t} \sim \mathcal{N}(0, \sigma^2_\varepsilon(\mathbf{x}_i, t))$, with the joint error $\gamma_i(t) + \varepsilon_{i,t} \sim \mathcal{N}(0, \sigma^2(\mathbf{x}_i, t))$. In order to sample smooth error curves with the desired

properties, we represent them as $\gamma_i(t) = w(\mathbf{x}_i, t)r_i(t)$ with $w(\mathbf{x}_i, t)$ a suitable weight function and $r_i(t) = \mathbf{b}^\top(t)\boldsymbol{\zeta}_i$ a random spline with a vector of K_Y basis functions $\mathbf{b}(t)$ and $\boldsymbol{\zeta}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{K_Y})$ constructed as described above in Section B.1.

Hence, for obtaining the desired joint error variance it must hold that

$$\sigma^2(\mathbf{x}_i, t) = (w(\mathbf{x}_i, t))^2 \cdot \mathbf{b}^\top(t)\mathbf{b}(t) + \sigma_\varepsilon^2(\mathbf{x}_i, t)\,,$$

as the variance of the random spline is given by $\mathbb{V}\text{ar}(r_i(t)) = \mathbf{b}^\top(t)\mathbf{Cov}(\boldsymbol{\zeta}_i)\mathbf{b}(t) = \mathbf{b}^\top(t)\mathbf{I}_{\mathbf{K_Y}}\mathbf{b}(t)$. This is achieved by setting $w(\mathbf{x}_i, t) = \sqrt{\frac{\sigma^2(\mathbf{x},t) - \sigma_\varepsilon^2(\mathbf{x},t)}{\mathbf{b}^\top(t)\mathbf{b}(t)}}$.

In our simulating studies, we either restrict to the smooth error $\gamma_i(t)$ for response curves with in-curve dependency or to $\varepsilon_{i,t}$ for the independent case.

2.2.2 *The general case*

Let the random spline $r_i(t) = \mathbf{b}^\top(t)\boldsymbol{\zeta}_i$ of degree l be defined as above. Consider a probability distribution with a cumulative distribution function (CDF) F, such that its quantile function F^{-1} is $l-1$ times continuously differentiable (of class C^{l-1}). Let F depend on Q parameters in a parameter space $\Theta \subseteq \mathbb{R}^Q$ and assume F^{-1} is also C^{l-1}-differentiable with respect to these parameters. Let $\boldsymbol{\vartheta}_i : T \mapsto \Theta$ be an C^{l-1}-differentiable parameter function and let Ψ denote the CDF of the standard normal distribution. Then,

$$y_i(t) = F^{-1}\left(\Psi\left(\frac{r_i(t)}{\sqrt{\mathbf{b}^\top(t)\mathbf{b}(t)}} \right) \middle| \boldsymbol{\vartheta}_i(t) \right)$$

is a C^{l-1}-differentiable curve and for every $t \in T$ the $y_i(t)$ is marginally distributed according to F with the current parameter setting $\vartheta_i(t)$. This is due to chain rule and inversion method (see, e.g., Devroye (1986)).

C Simulation details II: simulation studies

We perform two different simulation studies: the first is a large simulation study for the case of Gaussian functional response curves, where we model both mean and standard deviation in dependence of scalar and categorical covariates and can compare to a competitor in this special case. The second simulation study is motivated by the bacterial interaction model applied in Section 3. It includes functional covariates and response measurements following a zero adjusted gamma distribution. In both studies, we explore the fitting behavior with respect to different tuning parameters, sample sizes and in-curve autocorrelation structures.

C.1 Measures for evaluation

In accordance with the fitting aim formulated in Section 2.2, the general goodness-of-fit of a model is measured with respect to the Kullback-Leibler divergence to the true probability distributions: we use $\overline{\text{KLD}}(\hat{\mathbf{h}}) = \frac{1}{NG} \sum_{i=1}^{N} \sum_{t \in T_0} \text{KLD}\left[\mathcal{F}_{Y(t)|\mathbf{x}} : \widehat{\mathcal{F}}_{Y(t)|\mathbf{x}}\right]$, the mean Kullback-Leibler divergence over all point-wise evaluations for an estimated predictor $\hat{\mathbf{h}}$, where $\mathcal{F}_{Y(t)|\mathbf{x}}$ is the true distribution with the true parameter

functions $\boldsymbol{\theta}_i(t)$ and $\widehat{\mathcal{F}}_{Y(t)|\mathbf{X}}$ is the one with the estimated $\hat{\boldsymbol{\theta}}_i(t)$ for the i-th curve.

In order to evaluate and compare the goodness-of-fit of the individual effects or coefficient functions, we rely on the root mean square error (RMSE). In the application motivated simulation study, where the scale of the true covariate effects are not controlled, the RMSE is normalized by the range of the particular true effect to achieve comparability. In this case, the applied individual goodness-of-fit measure is given by

$$\begin{aligned} rel\text{RMSE}\left(\hat{f}_j^{(q)}\right) &= \text{RMSE}\left(\hat{f}_j^{(q)}\right) \Big/ \left(\max_t f_j^{(q)}(t) - \min_t f_j^{(q)}(t)\right) \quad \text{or} \\ rel\text{RMSE}\left(\hat{\beta}_j^{(q)}\right) &= \text{RMSE}\left(\hat{\beta}_j^{(q)}\right) \Big/ \left(\max_{s,t} \beta_j^{(q)}(s,t) - \min_{s,t} \beta_j^{(q)}(s,t)\right), \end{aligned}$$

for effect or coefficient functions, respectively.

C.2 Gaussian response model simulation

As in scalar regression, Gaussian functional response models play a prominent role in functional response regression. However, in contrast to previous regression frameworks, we consider the case where both mean and standard deviation depend on covariates. In this large scale simulation study, we consider two different models: one with scalar continuous covariates $z_1, z_2 \in [0, 1]$ and one with categorical covariates $g_1, g_2 \in \{1, ..., 4\}$. The covariates z_1 and g_1 influence both the mean $\mu(t)$ and the standard deviation $\sigma(t)$. z_2 and g_2 influence only the mean. The continuous covariate model includes a covariate interaction for $\mu(t)$. Precise

Table 1: Simulated models with scalar continuous and categorical covariates

Models:	Distribution: $Y(t) \mid \mathbf{x} \sim \mathcal{N}\left(\mu(t), \sigma^2(t)\right)$
1. Continuous	$\mu(t) = f_0^\mu(t) + f_1^\mu(z_1, t) + f_2^\mu(z_2, t) + f_3^\mu(z_1, z_2, t)$
	$\log \sigma(t) = f_0^\sigma(t) + f_1^\sigma(z_1, t)$
2. Categorical	$\mu(t) = \beta_0^\mu(t) + \beta_{g_1}^\mu(t) + \beta_{g_2}^\mu(t)$
	$\log \sigma(t) = \beta_0^\sigma(t) + \beta_{g_1}^\sigma(t)$

model formulations are presented in Table 1.

3.2.1 Sampling

The simulations follow a two-stage sampling approach: in each run, n_{eff} true effect function sets and n_{cov} covariate and random error sets are sampled. Evaluating the true effect functions on the covariates yields a total of $n_{eff} n_{cov}$ simulated data sets.

For N observations, continuous or categorical covariates are drawn independently from a uniform distribution on the unit interval or on $\{1, 2, 3, 4\}$, respectively.

We simulate smooth covariate effect functions $f_j^{(q)}, \beta_j^{(q)}$ generating them as cubic B-splines with random coefficients. Note that the smooth interaction effect $f_3^\mu(z_1, z_2, t)$ is fairly complex as it involves a double tensor product spline basis, and we distinguish this interaction effect from the marginal effects $f_1^\mu(z_1, t)$ and $f_2^\mu(z_2, t)$ using basis-orthogonalization.

All effect functions are drawn as random splines. The scale of mean and variance effects is specified via scale parameters $\bar{\sigma}_\mu^2,\ \bar{\sigma}_\sigma^2 > 0$. $\bar{\sigma}_\mu^2$ corresponds to the

overall mean variance over time of the effect functions, i.e. a weighted sum of the individual random spline scale parameters. All effect functions except for the interaction effect are weighted equally. The interaction effect is weighted with 1/8 as for this type of effect more extreme values occur. For $\bar{\sigma}^2_\sigma$, the overall mean variance of the effect functions is transformed for comparability, as a log-link is applied for the standard deviation: The overall mean variance $\tau^2 = \tau^2(\bar{\sigma}^2_\sigma)$ is specified such that for a random variable $Z \sim \mathcal{N}(0, \tau^2)$ we obtain $\mathbb{V}\text{ar}(\exp(Z)) = \bar{\sigma}^2_\sigma$. Via the properties of the log-normal distribution, τ^2 is determined as $\tau(\bar{\sigma}^2_\sigma) = -\log(2) + \log(\sqrt{4\bar{\sigma}^2_\sigma + 1} + 1)$. If not otherwise specified, we set $\bar{\sigma}^2_\mu = \bar{\sigma}^2_\sigma = 1$. All smooth effect function bases are chosen as cubic B-splines with 2nd order difference penalty using the smoothing parameter $\lambda = 0.8$, which results in sensibly smooth true effect functions. For functional intercepts we use B-spline bases with $k = 8$ basis functions. The tensor product bases have a total of $k = 3 \cdot 8 = 24$ basis functions for categorical effect functions (4 categories minus reference times the size of the intercept basis), $k = 6 \cdot 5 = 30$ for smooth marginal continuous effect functions (6 for time and covariate, respectively, minus 6 for orthogonalization constraint) and $k = 6 \cdot 6 \cdot 8 - 2 \cdot 30 - 8 = 220$ for the smooth interaction of continuous covariates (6 for both covariates, 8 for the intercept, minus sizes of the other bases). All effects are centered with respect to the particular functional intercept and the continuous interaction effect is orthogonalized with respect to its marginal effect functions.

Given the covariate values and the true underlying effect functions, smooth

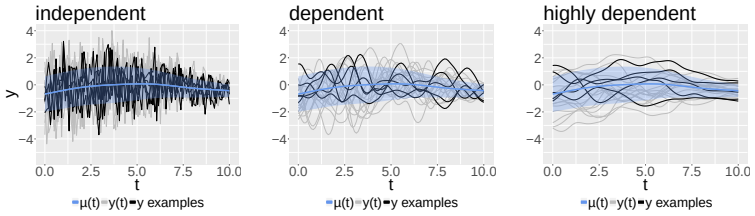

Figure 1: Sampled response curves for different in-curve dependency levels. A sample of N = 20 response curves (four highlighted in black) sampled around the mean curve *(blue)* over time for fixed covariate values. The blue area plotted around the mean corresponds to ± the standard deviation. Three levels of in-curve dependency are presented in increasing order. Compare Onl. App. B.

response curves are drawn randomly with the respective mean and standard deviation given by the model. Each curve is evaluated on G grid points of the time interval $T = [0, 10]$. Point-wise, the response curves are normally distributed. As described in Section B.2 smoothness is induced with random error splines. For obtaining different levels of in-curve dependency, different random splines are employed. For level *independent*, we do not use any random splines, but sample directly from the marginal distribution of $y_i(t)$. For level *dependent*, we use a cubic random spline with no smoothing penalty and $k = 20$ basis functions. We use a first order difference penalty with $\lambda = 0.5$ for level *high_dependency*.

3.2.2 *Model specification*

Corresponding to sampled true effect functions, all scalar function bases used to construct the effect functions occurring in the fitted model are specified as cubic P-splines with second order difference penalty. All base learners are set up such

that they have the same total degrees of freedom df. The default is $df = 13$. To obtain this, an additional Ridge penalty is applied to the categorical effects. The P-spline bases for the functional intercepts contain $k = 20$ basis splines, which is more than for the true intercepts such that the knowledge of the true knots is not used in the model fit. For other effects, the same basis dimension k is used as for the corresponding true effect. Again, all effects are centered around the functional intercept and marginal scalar effects are distinguished from the interaction.

For the hyper-parameters a default of step-size $\nu = 0.2$ is employed and by default the optimal stopping iteration m_{stop} is estimated by 10-fold curve-wise bootstrap. Furthermore, unless otherwise stated, the model is fitted with the GAMLSS boosting method described in Section 2.2, which corresponds to the 'noncyclic' boosting method in the R package `gamboostLSS`.

C.3 Application-motivated simulation study

In this simulation study, we adopt model and covariates entirely from the analysis of bacterial interaction in Section 3. Estimated effect functions from the original model fitted to the data present the true model structure in the simulation. Response curves are generated on the basis of random splines as described in Section B.2 and with the same specifications as discussed in 3.2.1 for the generic simulation study. Thus, in contrast to the above simulation study, we have one set of true effect functions and one set of covariates, only. In each simulation

run new response curves are sampled. We evaluate the model on 120 data sets per dependency level and compare *independent*, *dependent* and *high_dependency* response curves.

Figure 2 shows simulated response curves for each level of in-curve dependency, as well as corresponding original growth curves for comparison.

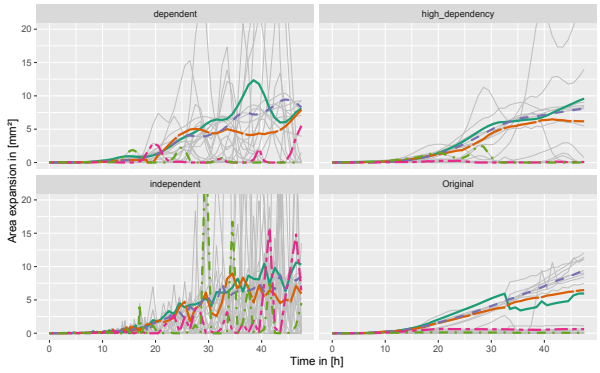

Figure 2: 24 simulated example curves for the applied dependency levels *independent*, *dependent* and *high_dependency* and original growth curves for comparison. Five randomly selected curves are highlighted in colors. Corresponding curves have the same covariate setting.

Visual comparison of the simulated curves and the original growth curves suggests that *high_dependency* might be roughly comparable to the in-curve dependency in the data. However, as we do not model the dependency structure in the presented model, the correspondence is limited.

D Simulation details III: results

In order to achieve comprehensive results, we consider a series of different simulation scenarios, where we vary the following parameters: the model scenario, i.e. continuous or categorical covariates; the sample size N and the number of measurements per curve G; the amount of in-curve dependency in the three categories described above; the boosting and re-sampling method; the effective degrees of freedom of the base-learners and the step length ν; and the ratio of the variance of the randomly generated true predictors for the mean and for the standard deviation. In addition, we compare the performance with the penalized likelihood approach of Greven and Scheipl (2017). An overview can be found in Table 2.

Table 2: Overview over simulations

	Model	n_{eff}	n_{cov}	N	G	*Dependency*	*df*	ν	*CV type*	*CV folds*	*Method*	$\bar{\sigma}^2_\mu$
Default	continuous	40	5	100	100	In, Dep, Hi	13	0.2	boot	10	noncyc.	1
Fig. 3, 5, 8	continuous, categorical			50, 100, 334, 500								
Fig. 3, 5				50, 100, 334, 500								2
Fig. 6					20, 50, 100, 150							
Fig. 7						In, Hi			kfold, boot, subs	10, 25, 50		
Fig. 9		6	3			In, Dep, Hi	13, 15, 17	0.05, 0.1, 0.2, 0.3				
Fig. 10, 11 Tab. 3		10	3								noncyc, cyclic	
Fig. 4, main man. 3	continuous	40	5	100, 334	104	In, Dep, Hi	–	–	–	–	refund	1
Fig. 12, 13	application	1	100	334	104	In, Dep, Hi	15	0.1	boot	10	noncyc.	–

Each row in the table corresponds to a simulation run, where for each of the n_{eff} true effect sets and n_{cov} covariate sets all combinations of the named simulation parameter specifications are compared. If nothing is specified explicitly, the default is applied. Besides the non-cyclic and cyclic boosting method one simulation run is also performed with penalized maximum likelihood, which is indicated by method 'refund'. *CV type* and *CV folds* describe the type of resampling method applied and the respective folds. 'kfold' denotes k-fold cross-validation, 'boot' bootstrapping, and 'subs' sub-sampling, each performed on curve level. In each fold of sub-sampling, the data is randomly devided into 50% training and 50% test data. $\bar{\sigma}^2_\sigma$ is always set to 1.

D.1 Figures

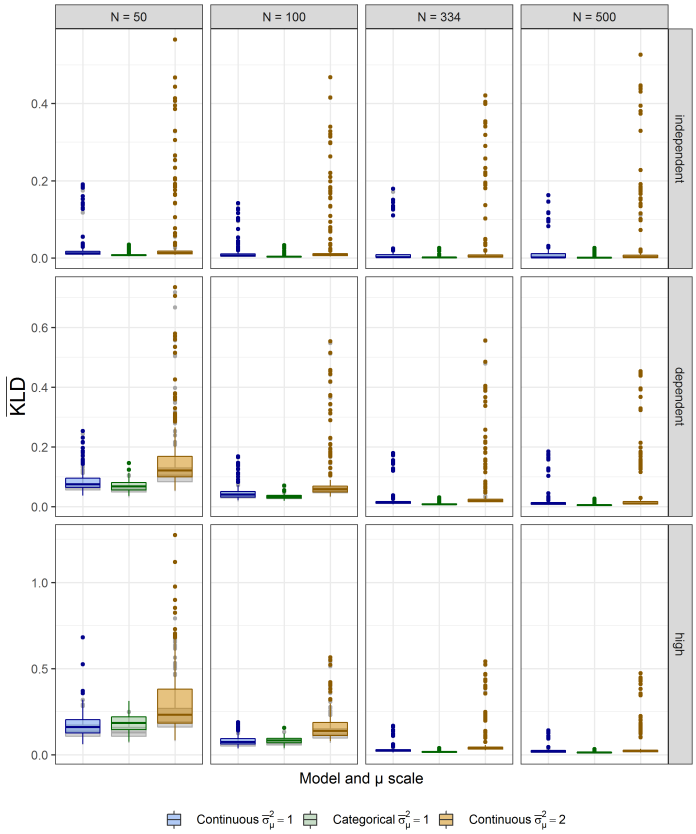

Figure 3: $\overline{\text{KLD}}$ for continuous and categorical covariate model: Comparison of $\overline{\text{KLD}}$ in the continuous models with $\bar{\sigma}^2_\mu = 1, 2$ and for the categorical model with $\bar{\sigma}^2_\mu = 1$ for different sample sizes and dependency levels. Grey box-plots in the background depict the distribution for the $\overline{\text{KLD}}$-optimal stopping iteration, instead of the one chosen by bootstrapping. Note that the y-axis limits change for the different dependency levels to enhance readability for the small values in the independent case.

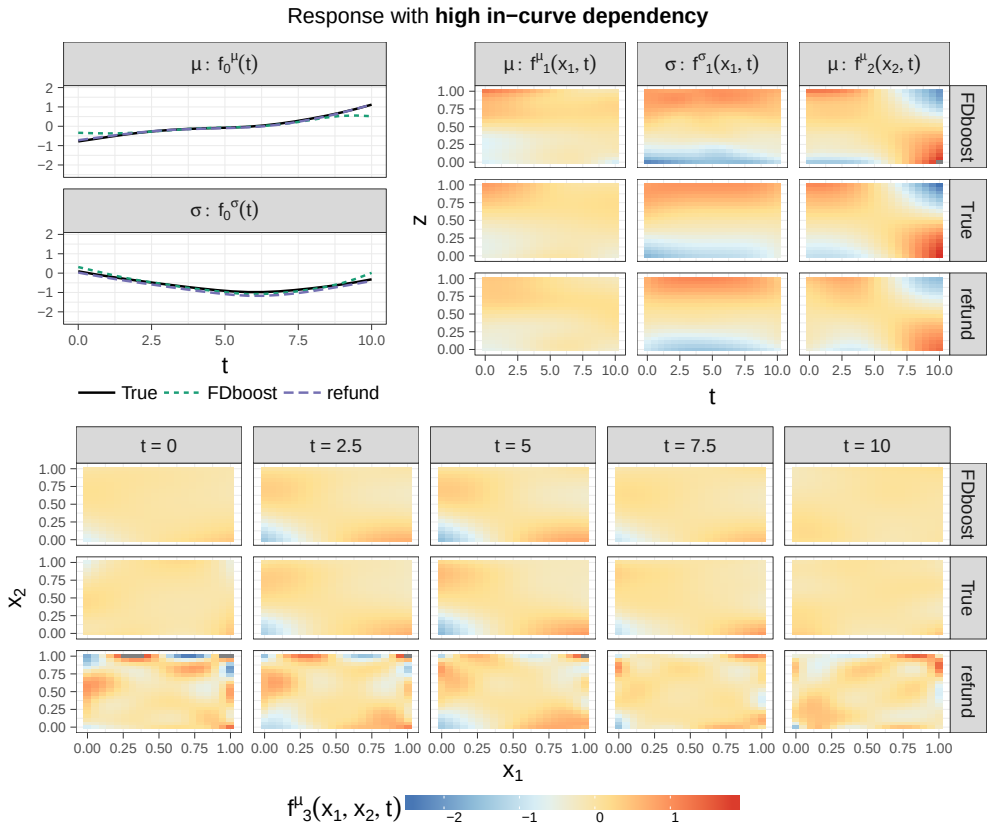

Figure 4: Example fit for high response in-curve dependency. The figure shows an example of true effect functions and estimates with the gradient boosting (`FDboost`) and penalized likelihood (`refund`) methods for the simulated continuous model scenario. The fit is based on a moderate number of $N = 334$ response-curves with $G = 100$ highly dependent measurements per curve and respective covariate samples. As the covariate interaction depends on z_1, z_2 and t, its effect functions are plotted for five fixed time points *(bottom)*. Especially for this complex effect, we see how the regularization via curve-wise bootstrapping for `FDboost` prevents the over-fitting that might occur with `refund`, if the estimated effects are too complex in relation to the given sample size of curves.

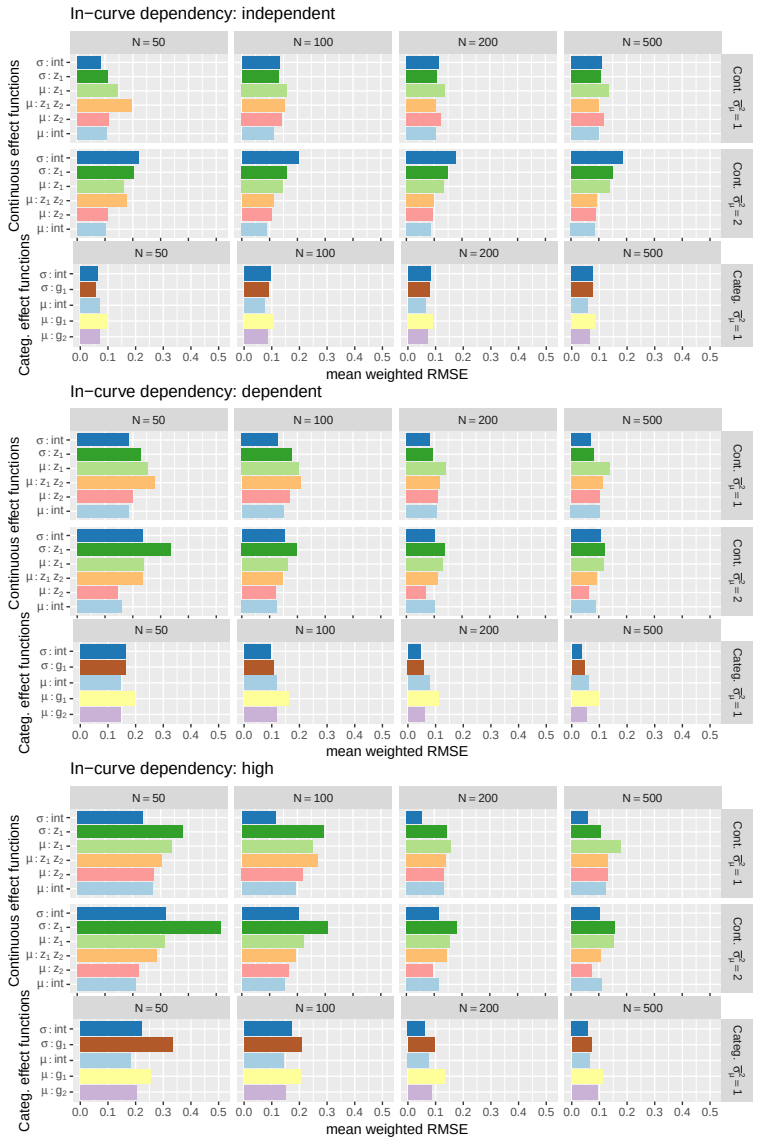

Figure 5: Weighted RMSE for individual effects in the continuous and categorical model: Comparison of goodness-of-fit for effect functions in the continuous models with $\bar{\sigma}^2_\mu = 1, 2$ and for the categorical model with $\bar{\sigma}^2_\mu = 1$ for different sample sizes and dependency levels. To account for their scaling, individual RMSE values are presented relative to the mean variance of simulated effect functions, i.e. relative to $\bar{\sigma}^2_\mu$ and $\tau(\bar{\sigma}^2_\sigma)$, respectively. For each effect function, the mean weighted RMSE over the simulations is depicted.

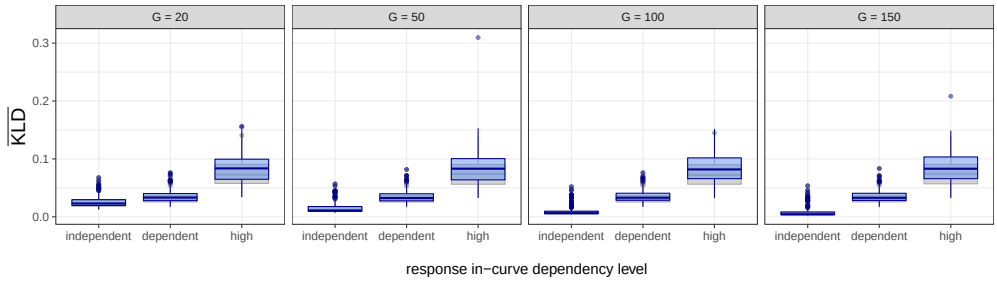

Figure 6: Comparison of different grid sizes G: Grid sizes G = 20, 50, 100, and 150 are compared for different dependency levels keeping $N = 100$ fixed. Box-plots visualize the distribution of $\overline{KLD}$ for fitted models *(blue, foreground)*. We only observe a considerable effect for independent response measurements. The $\overline{KLD}$ that would have been achieved with an optimal stopping iteration is depicted in the background *(gray)*.

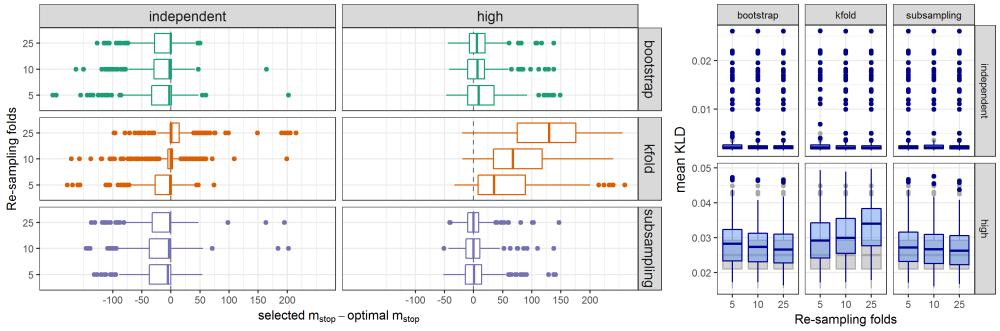

Figure 7: Comparison of re-sampling methods: Box-plots summarizing the deviation of selected stopping iterations m_{stop} from the KLD-optimal m_{stop} in the simulation. We compare bootstrapping, k-fold cross-validation, and sub-sampling for no and high in-curve dependency. Each method is applied with 5, 10, or 25 folds re-sampling. In sub-sampling, the data is randomly split into 50% training set, 50% test set in each fold. Dashed gray lines depict zero deviation. Since contribution of base learners decreases with boosting iterations, deviations are worse for high in-curve dependence.

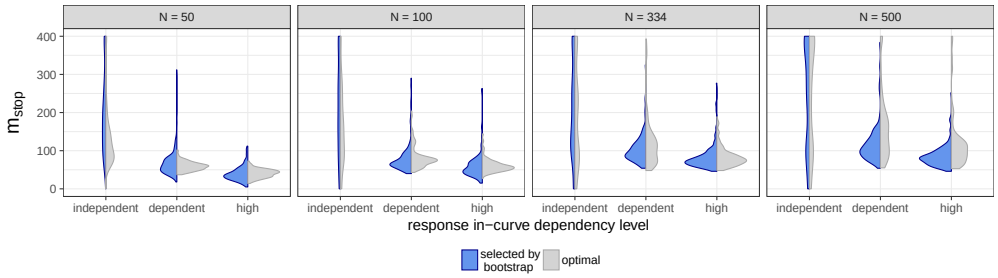

Figure 8: Violin-plots reflecting the empirical density of the stopping iterations m_{stop} selected via 10-fold bootstrap (*blue; left*) and for the $\overline{\text{KLD}}$-optimal m_{stop} (*gray; right*). Plots refer to the Gaussian model scenario with smooth covariate effects with 200 model fits per combination of sample size N and different in-curve dependency levels. Note that, as gradients decrease in absolute size with the number of boosting iterations, earlier iterations have a greater impact on the actual model fit than later iterations.

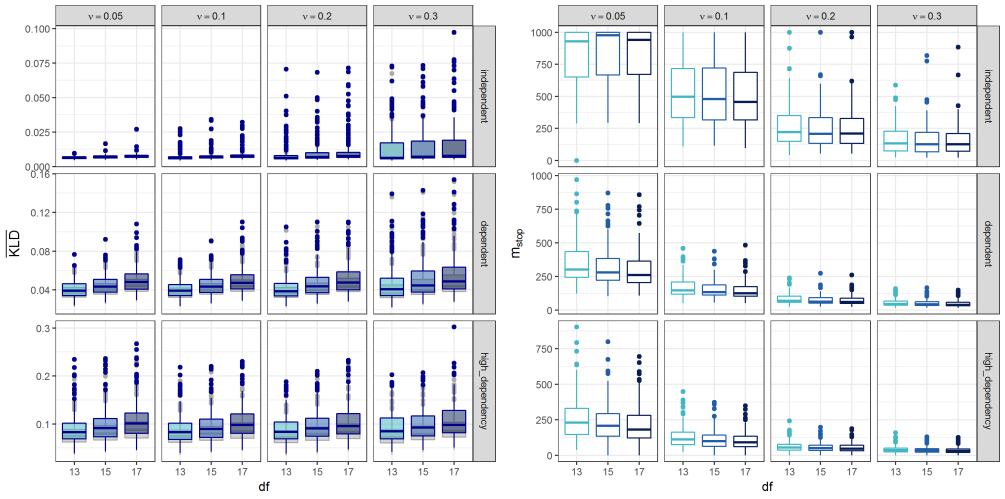

Figure 9: Comparison of different df and ν combinations: Base learner degrees of freedom df and step-length ν determine the flexibility of base learners in each boosting iteration. We compare different combinations for different dependency levels. To this end, distributions of $\overline{\text{KLD}}$ *(left)* and corresponding stopping iterations m_{stop} *(right)* in simulations are displayed with box-plots. For the $\overline{\text{KLD}}$, grey boxes in the background correspond to the $\overline{\text{KLD}}$ for the optimal m_{stop}. We identify $df = 13$ and $\nu = 0.2$ to be a suitable and yet fast combination. We do not consider $df < 13$, as our model effects are rather complex, and we do not want to apply df too close to the dimension of the penalty null space of the P-splines. E.g., for the base-learner reflecting the smooth interaction between z_1 and z_2 over t, we distribute the degrees of freedom such that we obtain $df = 3$ for the P-spline in direction of t (with penalty null space dimension 2) and $df = \sqrt{13/3} \approx 2.1$ in direction of z_1 and z_2, respectively.

Table 3: Mean $\overline{\text{KLD}}$ for available gradient boosting methods for GAMLSS

	cyclic	non-cyclic
independent	0.1764	0.0067
dependent	0.2410	0.0353
high dependency	0.3078	0.1010

Mean $\overline{\text{KLD}}$ for available boosting methods and for different levels of in-curve dependency. Method 'cyclic' was first proposed by Mayr et al. (2012). Later, non-cyclic methods were developed by Thomas et al. (2018). For all dependency levels, we observe that method 'non-cyclic' exhibits a better $\overline{\text{KLD}}$-performance than 'cyclic'.

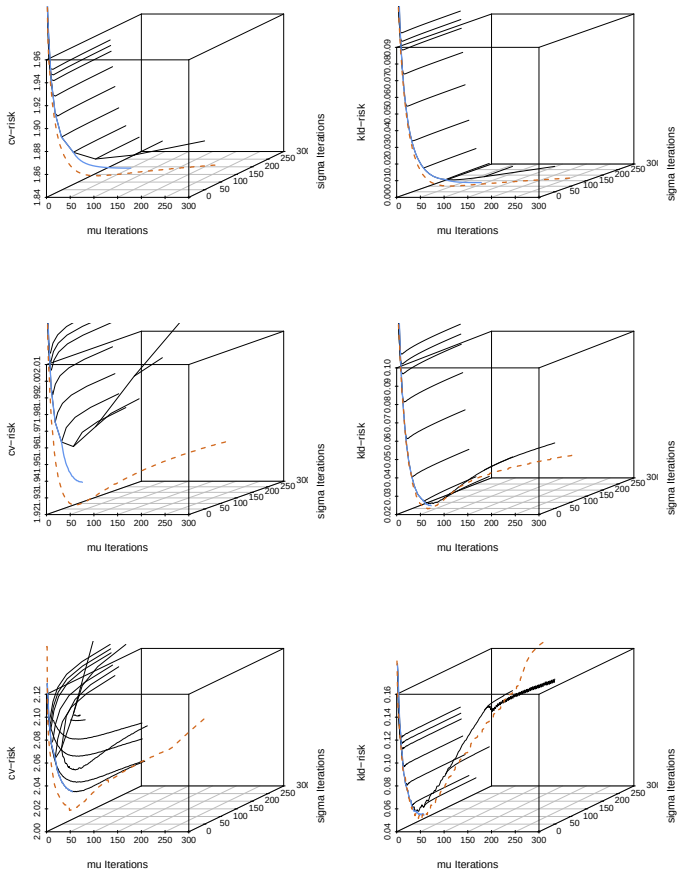

Figure 10: Cross-validation and $\overline{\text{KLD}}$ paths for the two different GAMLSS boosting methods: Exemplary cross-validation and $\overline{\text{KLD}}$ paths for both available gradient boosting methods for GAMLSS in an example scenario with $N = 100$ functional observations. x- and y-axes correspond to boosting steps in direction of μ and σ. The z-axis corresponds to resulting cross-validation error or $\overline{\text{KLD}}$, respectively. We compare cross-validation and $\overline{\text{KLD}}$ paths for in-curve dependency level *independent (top)*, for *dependent (middle)*, and for *high dependency (bottom)*. In method 'cyclic', μ- and σ-base learners are either updated alternately, or, from a certain point, the learners for one parameter are updated only *(black; blue indicates selected path)*. In the 'noncyclic' method *(chocolate, dashed)* free paths can be chosen. We observe, that the cross-validation error nicely approximates the structure of the $\overline{\text{KLD}}$. Especially for high in-curve dependencies, the $\overline{\text{KLD}}$ drops fast in the beginning, but rises afterwards when over-fitting occurs.

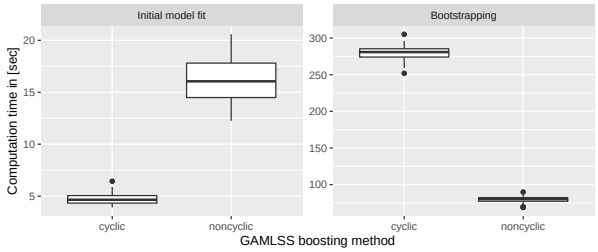

Figure 11: Computation time for available gradient boosting methods for GAMLSS: Computation time for fitting the continuous model with a maximum of $m_{stop} = 400$ boosting iterations. The cyclic and non-cyclic GAMLSS boosting methods are compared with respect to required time for a single model fit and for 10-fold bootstrapping. The models were fitted on a 64-bit linux server. While a single model fit in the cyclic method is faster, as the base-learners for the mean and standard deviation are fitted in an alternating way, re-sampling methods, like bootstrapping, take a lot longer than with the noncyclic method, since a seperate m_{stop} for each parameter has to be chosen, which demands for multiple model estimations per re-sampling fold.

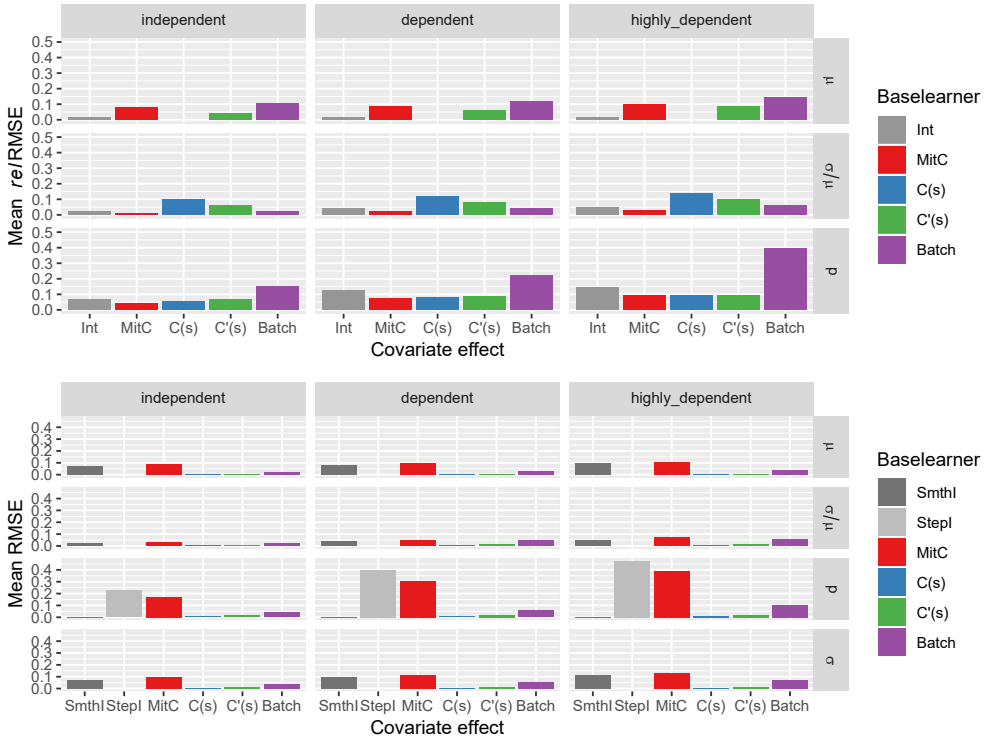

Figure 12: *Top:* *rel*RMSE per covariate effect: We compare the mean *rel*RMSE over 120 simulation runs per covariate effect in the model, and per in-curve dependency level. Effect functions are depicted for the mean μ and for the relative standard deviation σ/ν, as well as for the zero-probability p. The bars for the C-μ-effects are missing, as – after not being selected by the boosting algorithm in the original model – their relative error is not defined. *Bottom:* the corresponding absolute RMSE per covariate effect. For p both a a step-function intercept (StepI) and smooth intercept (SmthI) centered around StepI are included into the model formula and also distinguished in this plot. However, SmthI was never selected in the original model fit. In addition, the RMSE is here also depicted for σ as it results from the μ and σ/μ effects.

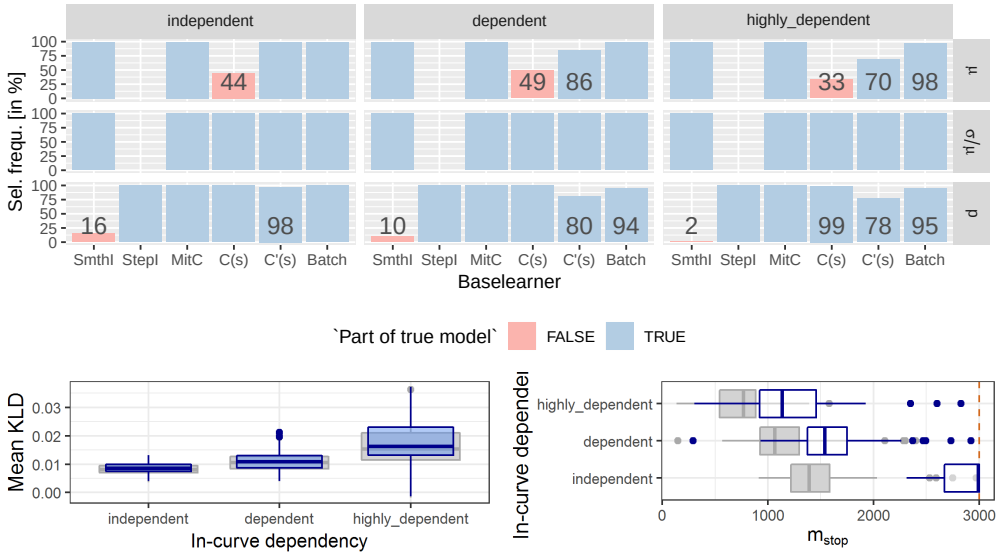

Figure 13: Diagnostic plots for the application-motivated simulation: *Top:* Bar-plots depicting the relative frequency of base-learners to be selected into the model (i.e. in at least one boosting iteration) in 120 simulation runs for different in-curve dependency levels (*columns*). Base-learners selected in less than 100% of the model fits are also labeled with their respective selection percentage. Base-learners which have not been selected in the original model fit and, thus, present nuisance effects in the simulation study are marked in red. *Bottom, left:* box-plots for $\overline{\text{KLD}}$ *(blue, foreground)* and the optimal $\overline{\text{KLD}}$ *(gray, background)*. *Bottom, right:* box-plots for stopping iteration m_{stop} determined by bootstrapping *(blue, foreground)* and the $\overline{\text{KLD}}$-optimal m_{stop} *(gray, background)*. While in this setting the optimal stopping iteration is typically overestimated by the bootstrap, this has only a slight effect on the fitting quality, as the $\overline{\text{KLD}}$ achieved by the estimation is very close to the optimal. This reflects the $\overline{\text{KLD}}$ paths in dependence of m_{stop} having a very flat minimum.

D.2 Discussion of results

4.2.1 *Preventing over-fitting with early stopping*

Especially for a small number N of sampled curves, high in-curve dependency promotes over-fitting. Flexible time-varying effect functions, as entailed in the present models, might be fitted to random patterns occurring in single curves. Early stopping in gradient boosting solves this problem. For dependent response measurements the optimal stopping iteration m_{stop}, with respect to the mean Kullback-Leibler Divergence ($\overline{\text{KLD}}$) over the domain, is much lower than for the independent case. At the same time, a larger sample size allows for more fitting iterations. As shown in Figure 3 in the main manuscript, this is captured well by estimating the $\overline{\text{KLD}}$-optimal m_{stop} with 10-fold curve-wise bootstrap. For high in-curve dependency bootstrapping or sub-sampling turn out to perform better than cross-validation, which tends to over-estimate the $\overline{\text{KLD}}$-optimal m_{stop} (Figure 7). As the information gain per measurement of a response curve is diminished by in-curve dependency, it is crucial that the available information is not over-estimated. In the case of in-curve dependency early stopping greatly improves the model fit, as otherwise over-fitting might easily occur (see also Figure 9).

This becomes even clearer when comparing the fitting performance with the GAMLSS-type regression for Gaussian response presented by Greven and Scheipl (2017), which is fit with the `R` package `refund`. In this alternative approach, the

model is fit via penalized maximum likelihood based on a working independence assumption for the response curve measurements. Usually, independence is assumed conditionally on a latent Gaussian random error process, which is included into the model to account for in-curve dependency. However, this would prohibit modeling the total variance of the response curves in a separate predictor, and, thus, no such process is included here. In contrast to the early stopping in our `FDboost` boosting approach, the current version of `refund` has, thus, no mechanism to account for in-curve dependency in the case of GAMLSS. The simulation results in Figure 3 of the main manuscript show that this may lead to severe over-fitting. While for the unrealistic case of independent response measurements the effect estimates are even slightly more accurate with `refund`, the fit is far better with `FDboost` for response curves with realistically dependent measurements. This is particularly visible for the smooth covariate interaction effect, the most complex effect in the model. The over-fitting in `refund` in this case is clearly visible also when comparing single example model fits (Figure 4).

4.2.2 Sample size

We compare sample sizes $N = 50, 100, 334, 500$ for a constant grid of $G = 100$ measurements per response curve. In the independent case, we obtain good estimations from $N = 50$ on (Figures 3 and 5). For dependent and highly dependent response measurements estimation is naturally harder due to the decrease in independent information provided. However, despite the high complexity

of the model, we observe that the structure of all effect functions is re-covered fairly well also in the highly dependent case (Figure 4). Here, the performance considerably improves with increasing sample size, such that we obtain good results for $N = 100$. In addition, the model fit gets more stable with higher sample sizes showing less variation of estimation quality across samples.

When comparing grid sizes $G = 20, 50, 100, 150$ for $N = 100$ response curves we observe no remarkable differences in estimation quality except for the independent case (Figure 6).The in-curve dependency structure reduces the 'effective number of measurements' provided by a fine grid.

Corresponding in some sense to the signal-to-noise ratio of simpler simulation scenarios, we control the scales of predictors for the mean and for the standard deviation by specifying the variances of the randomly generated true predictors. In most of the simulation scenarios both variances are chosen to be one. If the scale of the mean predictor is doubled, we observe that overall fitting gets worse with respect to the KLD. In this case, we observe an increased estimation quality for effects on the mean, but much worse estimation of the effects on the variance (Figures 3 and 5).

4.2.3 *Estimation of effects*

For the estimation quality of the individual effect functions, we obtain similar results as discussed above for the global estimation quality. As expected, the fitting error tends to increase with the effect complexity. In addition, we observe

that the covariate effects of variables which affect both mean and variance show an increased fitting error. However, neither mean nor variance effects show a clear general fitting advantage (see Figure 5).

4.2.4 *Application-motivated simulation*

Concerning the comparison of different dependency levels and the estimation of the optimal stopping iteration, we obtain similar results to the Gaussian case discussed above: accuracy and stability decrease with increasing dependency and selection of the best stopping iteration m_{stop} with cross-validation is sensitive to in-curve dependency. In this scenario, we observe a tendency of over-shooting the $\overline{\text{KLD}}$-optimal m_{stop} (Figure 13). Still, for all levels of in-curve dependency, the $\overline{\text{KLD}}$ for the optimal m_{stop} and the $\overline{\text{KLD}}$ achieved with the bootstrap-selected m_{stop} are quite similar, showing that the discrepancy in the stopping iteration has little impact on the fitting quality.

We observe that most of the RMSEs for the estimated covariate effects are lower than 10% of the effect range even in the highly dependent setting (Figure 12). However, the functional intercept in the predictor for $p(t)$ appears to be harder to estimate, being composed of a smooth functional intercept and a step function; the σ/μ-effect of $C(t)$ has a relatively large *rel*RMSE, due to its small effect size, while having a quite small absolute RMSE (Figure 12); the group-effects for the eight experimental batches show the largest *rel*RMSEs. They also have

comparably small effect sizes and the estimated effect for each batch may rely on the respective data only. Still, for the batch effects the interpretation of their exact shape is not of primary interest.

Overall, we observe the effects to be estimated quite well despite in-curve dependency and the high complexity of the model.

E Application details

E.1 Data

The data set contains $N = 334$ observed functional response curves $S_i(t)$ and covariate curves $C_i(t)$. Their value corresponds to area expansion in mm^2 of S- and C- bacterial strain, respectively. Original values in μm^2 were converted to avoid numerical instability. The curves are measured on a common grid of length $G = 105$ in the time interval $T = [0, 48\,h]$. For the first $18\,h\ 30\,min$, 75 measurements are taken every $15\,min$. The remaining 29 measurements are taken every hour. S- and C- strain are distinguished via red and green fluorescence. von Bronk et al. (2017) employ automatic segmentation of propagation areas from different color channels of recorded microscope pictures. In order to capture the full area over the whole time span, four different microscope zoom levels are applied. After $12\,h\ 15\,min$ the zoom level of the microscope is adjusted for the first time, again after $18\,h\ 30\,min$ and after $33\,h\ 30\,min$. The MitC concentration

is included as a categorical covariate with four levels. As the covariate $C'(t)$ is calculated from $C(t)$ by numerical differentiation, the last time point at $t = 48\,h$ is dropped for all other curves. For every observation a positive amount of *S*- and *C*-cells is present at the beginning. Due to automatic area segmentation some growth curves contain outliers marked by distinct jumps in the growth. Corresponding values of the cures were identified manually, deleted and replaced by spline interpolation. For each MitC concentration, two experimental batches were conducted. For MitC = 0 these include 34 and 41 bacterial growth spots; for MitC = 0.005 47 and 40 spots; for MitC = 0.01 45 and 40 spots; and for MitC = 0.1 46 and 41 spots. The number of spots per experimental batch varies as only spots with a positive number of S- and C-cells in the beginning were chosen. Figure 14 shows example microscope pictures of bacterial colonies and an overview over S- and C-growth curves.

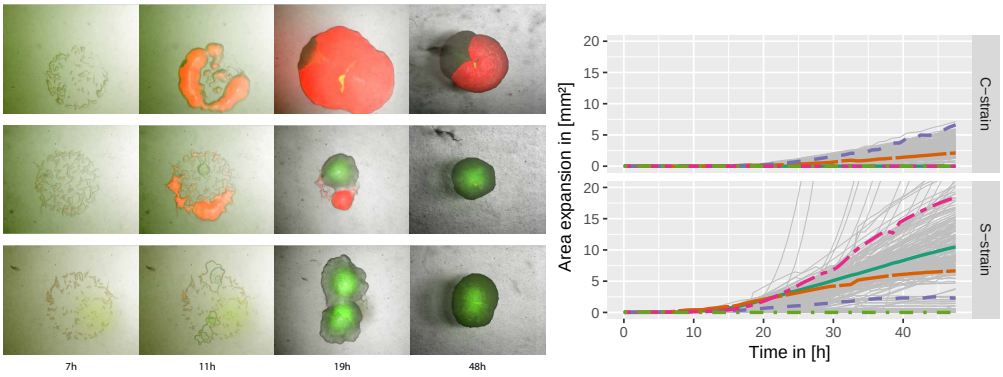

Figure 14: Bacterial growth data: *Left:* Overlay of bright-field, red and green fluorescence for three different bacterial spots after 7, 11, 19, and 48 hours. For the last two observation times the microscope zoom level was adjusted to cover the full bacterial area. *Right:* Bacterial growth curves of the C-strain *(top)* and S-strain *(bottom)*. Highlighted C- and S- example curves from the same spot are marked correspondingly.

E.2 Comparison with usual growth models

GAM(LSS) provide flexible means of modeling growth curves. We compare them to four parametric growths models. The four parameter Baranyi-Roberts model (Baranyi and Roberts, 1994) and the Gompertz model (Gibson et al., 1988) in the parametrization of Zwietering et al. (1990) present two popular approaches to modeling bacterial growth. Baty and Delignette-Muller (2014) formulate these in terms of common parameters y_0, y_∞, μ_{max} and L. The Baranyi-

Roberts model is given by

$$y(t) = y_0 + \log_{10}\left(\frac{-1 + \exp(\mu_{max}L) + exp(\mu_{max}t)}{exp(\mu_{max}t)} - 1 + 10^{y_\infty - y_0} exp(\mu_{max}L)\right).$$

The Gompertz model is given by

$$y(t) = y_0 + (y_\infty - y_0)\exp\left(-\exp\left(\frac{\mu_{max}e(L-t)}{(y_\infty - y_0)\log(10)} + 1\right)\right).$$

Weber et al. (2014) employ a five-parameter sigmoidal function describing effective radii of bacterial colonies in bacterial interaction. Depending on parameters $a, \ldots, d$ and y_0, their model takes the form

$$y(t) = a + \frac{y_0 - a}{(1 + (t/c)^b)^d}$$

We also add the standard three parameter logistic growth model of the form

$$y(t) = \frac{y_\infty y_0 \exp(rt)}{y_\infty + y_0\,(\exp(rt) - 1)}.$$

We fit two example S-strain growth curves from our data set with each of the models in order to obtain realistic parameter values. Fitting is done via least squares. In addition, we pick two alternative parameter settings. For comparison, we fit each of the generated parametric model curves with GAM. In analogy to our applied model, we fit the GAM by boosting using a gamma

distribution loss with a log-link and a functional intercept constructed with a B-spline. Resulting GAM approximation turns out to be nearly perfect for all parametric models (Fig. 15).

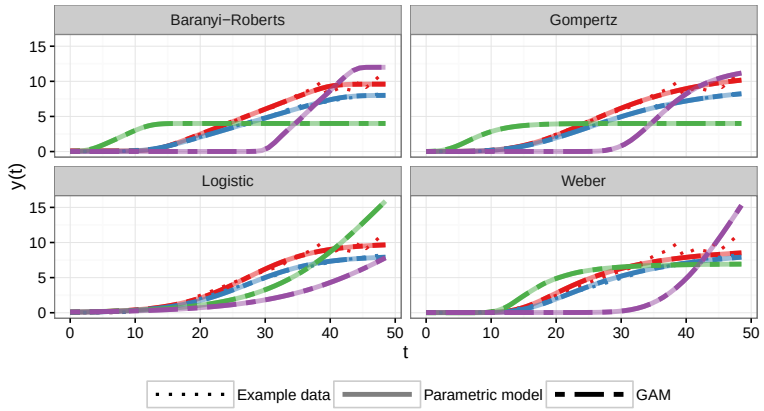

Figure 15: Comparison of growth models: GAM approximations to different curves of parametric growth models. For the red and blue curves the parametric models were fitted to example data. Green and violet curves display alternative growth curve shapes.

E.3 Estimated effect functions

This section contains supplementary figures showing all estimated effect C-effects (Figure 16) including those not shown in the main document. Figure 17 illustrates the effects of MitC on the overall mean and standard deviation, rather than the effect on the conditional mean and standard deviation given $S(t) > 0$.

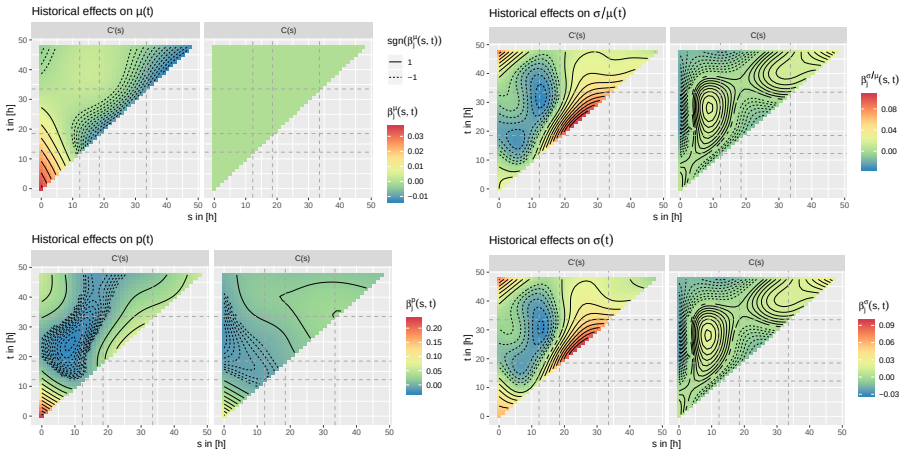

Figure 16: Historical effects of C-strain propagation area: Coefficient functions $\beta^{(q)}(s,t)$ for the historical effects of $C'(t)$ and $C(t)$ on the mean *(top, left)*, the scale parameter *(top, right)*, the zero area probability *(bottom, left)*, and the standard deviation *(bottom, right)* of the S-strain growth curves. The y-axis presents the time line for the S-strain curve, the x-axis the one for the C-strain. Grey dashed lines mark zoom breaks. Note that for in the early phase with $s, t \leq 10\ h$ there are almost no observations with $S_i(t) = 0$, so the corresponding effects on $p(t)$ should not be interpreted. Moreover, the $C - \mu$-effect was never selected by the boosting algorithm and is, thus, constantly zero.

As for $t \leq 10\,h$ the probability $P(S_i(t) = 0) \approx 0$ and there is almost no vanishing of S in the data, the estimated effects in this period are questionable and, thus, we refrain from interpreting them. For later growth periods we observe a mainly positive effect of $C_i(s)$ and $C_i'(s)$ on $p(t)$ corresponding to an increased probability for $S_i(t) = 0$ for an increased C strain growth (areal competition), consistent with

results for $\mu(t)$. Only the C'-μ-effect for $10\,h < s < 20\,h$ is estimated negative for later S-strain growth.

While we, thus, concentrate on the model part conditioning on $S_i(t) > 0$, we observed qualitatively similar effects of $C'(s)$ on the mean, when replacing zero response values by the smallest observed positive values in preliminary analyses. This indicates that $C'(s)$ shows a similar effect on the unconditional mean $\mathbb{E}[S_i(t)] = \mu_i(t)(1 - p_i(t))$.

Based on the estimates for $p(t)$ and for $\mu(t)$ and $\sigma/\mu(t)$ conditioning on $S(t) > 0$, we may also compute the unconditional mean and standard deviation of $S(t)$, e.g., for the different MitC concentrations (Figure 17).

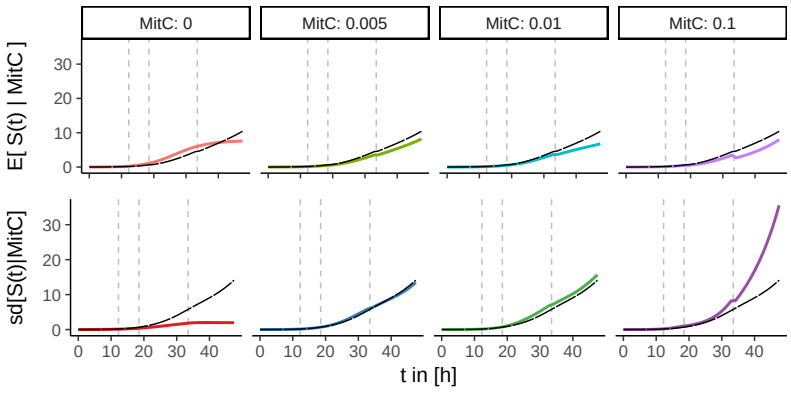

Figure 17: Overall MitC-effects: Estimated point-wise mean *(top)* and standard deviation *(bottom)* of the S-strain growth curves $S_i(t)$ without conditioning on $S_i(t) > 0$. Long-dashed black curves correspond to the functional intercept, dashed vertical lines to the zoom level change-points.

E.4 Bootstrap uncertainty bounds for effect functions on predictor level

We computed point-wise bootstrap uncertainty bounds for all effects in a basic bootstrap 95% confidence interval type procedure with 1066 bootstrap samples. Accordingly, the intervals are computed as $[\hat{f}_j^{(q)}(t) - \Delta^*_{0.975}(t), \hat{f}_j^{(q)}(t) - \Delta^*_{0.025}(t)]$ for group-specific functional effects and $[\hat{\beta}_j^{(q)}(s,t) - \Delta^*_{0.975}(s,t), \hat{\beta}_j^{(q)}(s,t) - \Delta^*_{0.025}(s,t)]$ for historical effect coefficient surfaces, where Δ^*_α denotes the bootstrap estimate of the point-wise α-quantiles of the distribution of $\hat{f}_j^{(q)} - f_j^{(q)}$ or $\hat{\beta}_j^{(q)} - \beta_j^{(q)}$, respectively. The bounds are meant to give indications on the estimation precision complementing the results of our simulation study, but due to the complexity of the matter we refrain from interpreting them as valid confidence bounds. As the estimators are subject to shrinkage bias introduced by early stopping of the boosting algorithm, we can not expect the bootstrap estimators Δ^*_α to be unbiased. In addition, we are limited to basic bootstrap confidence intervals as we do not have proper estimates for the estimators' variances and, thus, cannot compute studentized bootstrap or accelerated bias-corrected bootstrap intervals (compare, e.g., (Hall, 1988)) without tremendous computational burden. Visualizations of the uncertainty bounds can be found in Figure 18 for $f_j^{(q)}$ and in Figures 19, 20 and 21 for $\beta_j^{(q)}(s,t)$.

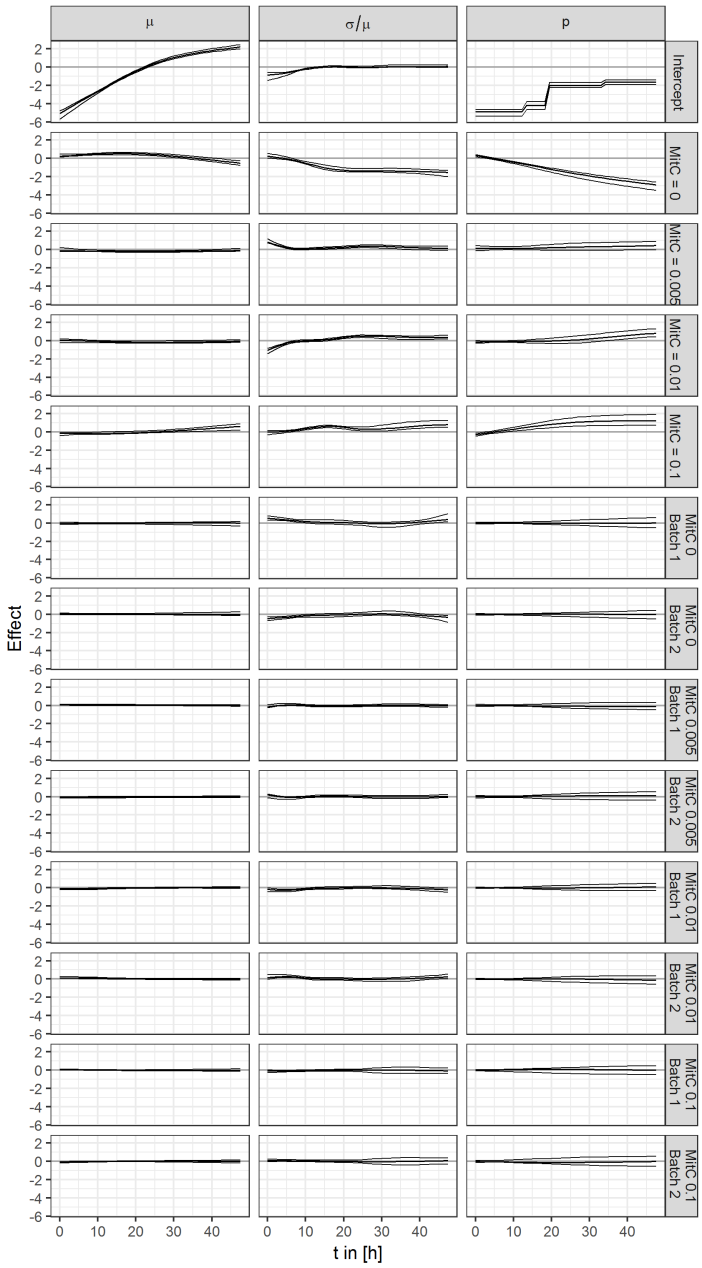

Figure 18: Estimated effect functions of the functional intercept, the MitC effects and the batch effects on μ, σ/μ and p with 95% bootstrap confidence interval type uncertainty bounds based on 1066 bootstrap samples.

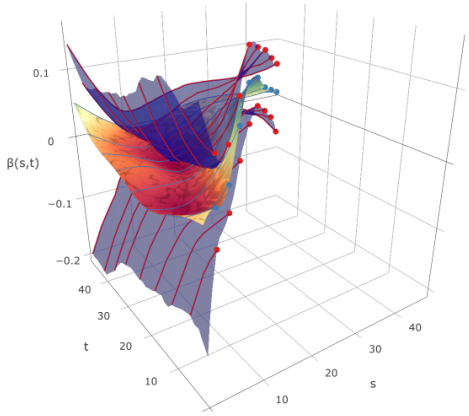

Figure 19: Example surface plot of C'-σ-effect with 95% bootstrap confidence interval type uncertainty bounds visualizing the line segments in the Figures 20 and 21.

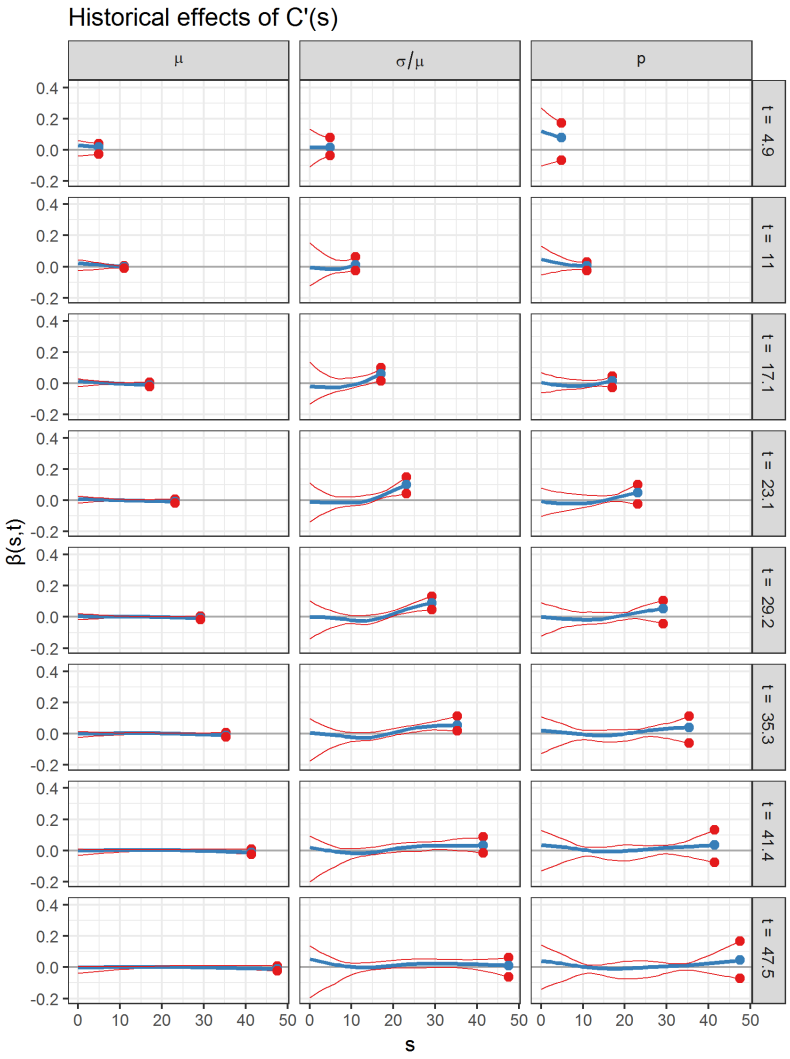

Figure 20: 95% bootstrap confidence interval type uncertainty bounds (*red*) for estimated line segments $s \mapsto \hat{\beta}(s,t)$ of the historical effect of $C'(s)$ for different fixed values of t (*blue*). The dots correspond to the values at $s = t$ and therefore serve for identification (compare Figure 19).

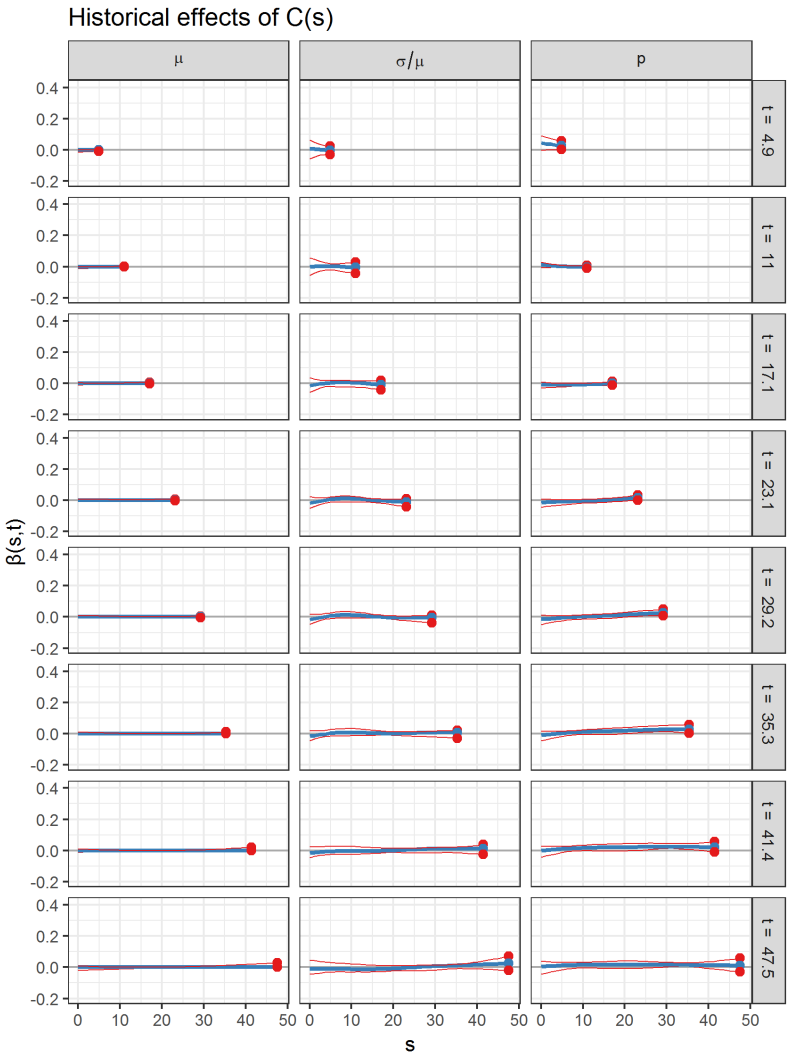

Figure 21: 95% bootstrap confidence interval type uncertainty bounds (*red*) for estimated line segments $s \mapsto \hat{\beta}(s,t)$ of the historical effect of $C(s)$ for different fixed values of t (*blue*). The dots correspond to the values at $s = t$ and therefore serve for identification (compare Figure 19).

References

Baranyi, J. and Roberts, T. A. (1994). A dynamic approach to predicting bacterial growth in food. *International Journal of Food Microbiology*, **23**(3-4), 277–94.

Baty, F. and Delignette-Muller, M.-L. (2014). *nlsMicrobio: data sets and nonlinear regression models dedicated to predictive microbiology*. R package version 0.0-1.

Brockhaus, S., Scheipl, F., and Greven, S. (2015). The Functional Linear Array Model. *Statistical Modelling*, **15**(3), 279–300.

Currie, I., Durban, M., and Eilers, P. (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**(2), 259–280.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer-Verlag.

Fahrmeir, L., Kneib, T., and Lang, S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica*, **14**, 731–761.

Gibson, A. M., Bratchell, N., and Roberts, T. A. (1988). Predicting microbial growth: growth responses of salmonellae in a laboratory medium as affected by ph, sodium chloride and storage temperature. *International Journal of Food Microbiology*, **6**(2), 155–178.

Greven, S. and Scheipl, F. (2017). A general framework for functional regression modelling (with discussion). *Statistical Modelling*, **17**(1-2), 1–35.

Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals. *The Annals of Statistics*, pages 927–953.

Mayr, A., Fenske, N., Hofner, B., Kneib, T., and Schmid, M. (2012). Generalized additive models for location, scale and shape for high dimensional data: a flexible approach based on boosting. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **61**(3), 403–427.

Scheipl, F., Staicu, A.-M., and Greven, S. (2015). Functional additive mixed models. *Journal of Computational and Graphical Statistics*, **24**(2), 477–501.

Thomas, J., Mayr, A., Bischl, B., Schmid, M., Smith, A., and Hofner, B. (2018). Gradient boosting for distributional regression: faster tuning and improved variable selection via noncyclical updates. *Statistics and Computing*, **28**(3), 673–687.

von Bronk, B., Schaffer, S. A., Götz, A., and Opitz, M. (2017). Effects of stochasticity and division of labor in toxin production on two-strain bacterial competition in *Escherichia coli*. *PLoS Biology*, **15**(5), e2001457.

Weber, M. F., Poxleitner, G., Hebisch, E., Frey, E., and Opitz, M. (2014). Chemical warfare and survival strategies in bacterial range expansions (online supplement). *Journal of The Royal Society Interface*, **11**(96), 20140172.

Wood, S. N. (2006). Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics*, **62**(4), 1025–1036.

Zwietering, M. H., Jongenburger, I., Rombouts, F. M., and van 't Riet, K. (1990). Modeling of the bacterial growth curve. *Applied and Environmental Microbiology*, **56**(6), 1875–1881.

B. Supplementary material for Chapter 5 "Functional Additive Regression on Manifolds of Planar Shapes and Forms"

Online supplement for the contribution:
Stöcker, A., Steyer, L., and Greven, S. (2022). Functional additive models on manifolds of planar shapes and forms. *arXiv pre-print*. DOI: 10.48550/ARXIV.2109.02624. *Tentatively accepted for publication in the Journal of Computational and Graphical Statistics.*

S Online Supplementary Material to

Functional additive models on manifolds of planar shapes and forms

by Almond Stöcker, Lisa Steyer and Sonja Greven

S.1 Geometry of functional forms and shapes

S.1.1 Translation, rotation and re-scaling as normal subgroups

We consider the following invariances of a response curve $y \in \mathcal{Y}$ with respect to the transformations $\mathcal{Y} \to \mathcal{Y}$ given by the group actions of translation $\text{Trl} = \{y \overset{\text{Trl}_\gamma}{\longmapsto} y + \gamma \mathbb{1} : \gamma \in \mathbb{C}\}$ with some $\mathbb{1} \in \mathcal{Y} \setminus \{0\}$ (for curves typically $\mathbb{1} : t \mapsto \frac{1}{\|t \mapsto 1\|}$ the real constant function of unit norm), re-scaling $\text{Scl} = \{y \overset{\text{Scl}_\lambda}{\longmapsto} \lambda \cdot (y - 0_y) + 0_y : \lambda \in \mathbb{R}^+\}$ around a reference point $0_y \in \mathbb{C}$, and rotation $\text{Rot} = \{y \overset{\text{Rot}_u}{\longmapsto} u \cdot (y - 0_y) + 0_y : u \in \mathbb{S}^1\}$ around 0_y with $\mathbb{S}^1 = \{u \in \mathbb{C} : |u| = 1\} = \{\exp(\omega\sqrt{\text{-}1}) : \omega \in \mathbb{R}\}$ the circle group reflecting counterclockwise rotations by ω radian measure. In the literature, the reference point is usually omitted setting $0_y = 0$, which can be done without loss of generality under translation invariance (i.e. in particular for shapes/forms). However, keeping other possible combinations of invariances in mind, we explicitly refer to an individual reference point and suggest the centroid $0_y = \langle \mathbb{1}, y \rangle \mathbb{1}$ or, more generally, $0_y = a(y)\mathbb{1}$ for some linear functional $a : \mathcal{Y} \to \mathbb{R}$. Assuming $\text{Trl}_\gamma(0_y) = 0_{\text{Trl}_\gamma(y)}$, as for the centroid, the definition of re-scaling and rotation around 0_y ensures that Trl_γ, Scl_λ and Rot_u commute – and that Trl, Rot and Scl present normal subgroups of the combined group actions $\{y \mapsto \lambda u y + \gamma : \gamma \in \mathbb{C}, \lambda \in \mathbb{R}^+, u \in \mathbb{S}^1\}$ of shape invariances. Thus, the combined group actions can be written as the direct product (or direct sum) $\text{Trl} \times \text{Scl} \times \text{Rot} = \{\text{Trl}_\gamma \circ \text{Scl}_\lambda \circ \text{Rot}_u : \gamma \in \mathbb{C}, \lambda \in \mathbb{R}^+, u \in \mathbb{S}^1\} \cong \mathbb{C} \times \mathbb{R}^+ \times \mathbb{S}^1$ and invariances with respect to Trl, Scl, Rot can be modularly accounted for in arbitrary order. $\text{Trl} \times \text{Rot}$, for instance, describe rigid motions. The ultimate response object is then given by the *orbit* $[y]_G = \{g(y) : g \in G\}$ (or short $[y]$), i.e. the equivalence class with respect to the direct sum G generated by the chosen combination of Trl, Scl and Rot. $[y]_{\text{Trl} \times \text{Scl} \times \text{Rot}}$ is referred to as the *shape* of y and $[y]_{\text{Trl} \times \text{Rot}}$ as its *form* or *size-and-shape* (compare Dryden and Mardia, 2016); studying $[y]_{\text{Scl}}$ is closely related to *directional* data analysis (Mardia and Jupp, 2009) where the direction of y is analyzed independent of its size $\|y\|$.

S.1.2 Parallel transport of form tangent vectors

To confirm that the parallel transport in the form space $\mathcal{Y}^*_{/\text{Trl} \times \text{Rot}}$ can be carried out via representatives in $\mathcal{Y}$ as described in the main manuscript, we closely follow Huckemann et al. (2010) in their derivation of shape parallel transport. Necessary differential geometric notions and statements are briefly introduced in the following before stating the main result in Lemma 1. For a more profound introduction, we recommend Lee (2018, in particular, p. 21, 43, 93, 124, 337, 402), as well as Tu (2011) for an illustrative introduction into some of the concepts, and Klingenberg (1995, in particular, p.103- 107) for an introduction in the light of potentially infinite dimensional manifolds.

The entire argument crucially relies on properties known for Riemannian submersions between differentiable manifolds $\widetilde{\mathcal{M}}$ and $\mathcal{M}$, which allow to relate the structure of $\mathcal{M}$ back to $\widetilde{\mathcal{M}}$. A *submersion* is a smooth surjective function $\Phi : \widetilde{\mathcal{M}} \to \mathcal{M}$, for which also the differential $d\Phi : T_{\widetilde{q}}\widetilde{\mathcal{M}} \to T_q\mathcal{M}$, $\widetilde{q} \in \widetilde{\mathcal{M}}$, $q = \Phi(\widetilde{q})$, is surjective at each $\widetilde{q} \in \widetilde{\mathcal{M}}$. For $q \in \mathcal{M}$, the $\Phi^{-1}(\{q\})$ are submanifolds of $\mathcal{M}$, and $T_{\widetilde{q}}\widetilde{\mathcal{M}} = T_{\widetilde{q}}\Phi^{-1}(\{q\}) \oplus H_{\widetilde{q}}\widetilde{\mathcal{M}}$ can be decomposed into the *vertical space* $T_{\widetilde{q}}\Phi^{-1}(\{q\}) = \ker(d\Phi)$ and its orthogonal complement $H_{\widetilde{q}}\widetilde{\mathcal{M}}$, the *horizontal space*. When restricted to the horizontal space, $d\Phi\big|_{H_{\widetilde{q}}\widetilde{\mathcal{M}}} : H_{\widetilde{q}}\widetilde{\mathcal{M}} \to T_q\mathcal{M}$ presents a linear isomorphism. A submersion Φ is called *Riemannian submersion* if $d\Phi\big|_{H_{\widetilde{q}}\widetilde{\mathcal{M}}}$ is also isometric. It gives rise to an identification $T_q\mathcal{M} \cong H_{\widetilde{q}}\widetilde{\mathcal{M}}$ of tangent spaces of $\mathcal{M}$ with horizontal spaces on $\widetilde{\mathcal{M}}$. Such an identification underlies the presentation of the response geometry in Section 2 of the main manuscript.

By construction, the quotient map $\Phi : \mathcal{Y}^* \to \mathcal{Y}^*_{/\,\mathrm{Trl}\times\mathrm{Rot}}, y \mapsto [y]$ presents a Riemannian submersion: Since $[p] = \{up + \gamma\imath : u \in \mathbb{S}^1, \gamma \in \mathbb{C}\}$ embeds $\mathbb{S}^1 \times \mathbb{R}^2$ in $\mathcal{Y}$, and, since the tangent spaces of $\mathbb{S}^1$ and $\mathbb{R}^2$ are well-known, the vertical space is given by $T_p[p] \cong \{\lambda p\sqrt{\text{-}1} + \gamma\imath : \lambda \in \mathbb{R}, \gamma \in \mathbb{C}\} \subset \mathcal{Y}$, with orthogonal complement $H_p\mathcal{Y}^* \cong \{y \in \mathcal{Y} : \langle y, \imath\rangle = 0,\ \mathrm{Im}(\langle y, p\rangle) = 0\}$ (see also Figure 1 in the main manuscript for an illustration). While Φ is obviously surjective, surjectivity and isometry of $d\Phi|_{H_p\mathcal{Y}^*}$ can be seen by expressing Φ in terms of the charts for $\mathcal{Y}^*_{/\,\mathrm{Trl}\times\mathrm{Rot}}$: for a given $p \in [p] \in \mathcal{Y}^*_{/\,\mathrm{Trl}\times\mathrm{Rot}}$, the map $\widetilde{(\cdot)} : [y] \mapsto \widetilde{y}^{\mathrm{Trl}\times\mathrm{Rot}}$ provides a chart $\mathcal{U}_{[p]} \to \mathcal{V}_p$, i.e. an isomorphism from $\mathcal{U}_{[p]} = \{[y] \in \mathcal{Y}^*_{/\,\mathrm{Trl}\times\mathrm{Rot}} : \langle \widetilde{p}^{\mathrm{Trl}}, \widetilde{y}^{\mathrm{Trl}}\rangle \neq 0\}$ to $\mathcal{V}_p = \{y \in \mathcal{Y} : \mathrm{Im}(\langle p, y\rangle) = 0, \mathrm{Re}(\langle p, y\rangle) > 0, \langle \imath, y\rangle = 0\}$ used to establish the differential structure on $\mathcal{Y}^*_{/\,\mathrm{Trl}\times\mathrm{Rot}}$. Expressed in this chart, $\widetilde{\Phi}(y) = \widetilde{(\cdot)} \circ \Phi(y) = \widetilde{y}^{\mathrm{Trl}\times\mathrm{Rot}}$ is the identity for all $y \in \mathcal{V}_p \subset \Phi^{-1}\left(\mathcal{U}_{[p]}\right)$. Thus, since $T_p\mathcal{V}_p = H_p\mathcal{Y}^*$, also $d\widetilde{\Phi}\big|_{H_p\mathcal{Y}^*}$ is the identity, which is obviously an isometric isomorphism. The latter carries over to $d\Phi\big|_{H_p\mathcal{Y}^*}$ independent of the given chart.

The isometric isomorphism $d\Phi\big|_{H_p\mathcal{Y}^*} : H_p\mathcal{Y}^* \to T_{[p]}\mathcal{Y}^*_{/\,\mathrm{Trl}\times\mathrm{Rot}}$ yields the identification $T_{[p]}\mathcal{Y}^*_{/\,\mathrm{Trl}\times\mathrm{Rot}} \cong \{y \in \mathcal{Y} : \langle y, \imath\rangle = 0,\ \mathrm{Im}(\langle y, p\rangle) = 0\}$, which we rely on in the main manuscript. Unlike there, we denote $d\Phi\big|^{-1}_{H_y\mathcal{Y}^*} : \xi \mapsto \widetilde{\xi}$ also for tangent vectors in the following, to make the identification of $\xi = d\Phi(\widetilde{\xi}) \in T_{[y]}\mathcal{Y}^*_{/\,\mathrm{Trl}\times\mathrm{Rot}}$ with the corresponding $\widetilde{\xi} \in H_y\mathcal{Y}^*$, usually referred to as *horizontal lift*, explicit in the notation.

The *covariant derivative* (Levi-Civita connection) $\nabla^{\mathcal{M}}_V W \in T\mathcal{M}$ of a vector-field $W \in T\mathcal{M}$ along a vector-field $V \in T\mathcal{M}$ provides a derivative of vector-fields in the tangent bundle $T\mathcal{M} = \{T_q\mathcal{M} : q \in \mathcal{M}\}$ of a Riemannian manifold $\mathcal{M}$. As a derivation in W and a linear function in V, $\nabla^{\mathcal{M}}$ fulfills a set of properties identifying it as unique generalization of ordinary directional derivatives of the components of $W : q \mapsto W_q \in T_q\mathcal{M}$ into the direction $V_q \in T_q\mathcal{M}$. For a submanifold $\mathcal{M}$ of a linear space $\mathcal{Y}$, $\nabla^{\mathcal{M}}_V W$ corresponds to the ordinary directional derivative orthogonally projected into $T_q\mathcal{M}$. For the linear case (with $\mathcal{M} = \mathcal{Y}$), the covariant derivative of a vector field $W(\tau) := W_{c(\tau)}$ along a differentiable curve $c(\tau)$ is directly given as

$$\nabla^{\mathcal{Y}}_{\dot{c}(t)} W(\tau) = \dot{W}(\tau) = \frac{d}{d\tau} W(\tau). \tag{1}$$

In analogy to straight lines, geodesic curves $c(\tau)$ are characterized by

$$\nabla^{\mathcal{M}}_{\dot{c}(\tau)}\dot{c}(\tau) = 0,$$

i.e. curves with zero 'second derivative'. More generally, a vector-field W is called parallel along a curve $c(\tau)$ if

$$\nabla^{\mathcal{M}}_{\dot{c}(\tau)}W(\tau) = 0. \tag{2}$$

According to that the parallel transport $\mathrm{Transp}^{c}_{q,q'} : T_q\mathcal{M} \to T_{q'}\mathcal{M}$ along a curve $c : [\tau_0, \tau_1] \to \mathcal{M}$ between $c(\tau_0) = q,\ c(\tau_1) = q' \in \mathcal{M}$ is defined to map tangent vectors $\varepsilon = W(\tau_0) \mapsto \varepsilon' = W(\tau_1)$ for some vector field W parallel along c (fulfilling Equation 2). If the curve c is clear from context, we omit it in the notation. This is especially the case in the following, where c can be chosen as the unique geodesic between two forms $[p]$ and $[p']$ with $\langle p, p'\rangle \neq 0$, yielding a canonical connection (in this case, c corresponds to the line between p and the aligned $\tilde{p}'$; for $\langle p, p'\rangle = 0$, by contrast, it is easy to see that for each $u \in \mathbb{S}^1$ the line between p and up' corresponds to a different geodesic; the second case can, however, be neglected).

The possibility to effectively carry out the parallel transport between forms $[p], [p']$ on suitable representatives $p, p' \in \mathcal{Y}^*$ stems from the following theorem and subsequent Corollary (compare, e.g, Klingenberg, 1995, p. 103-105).

Theorem 1. *Let $\Phi : \widetilde{\mathcal{M}} \to \mathcal{M}$ be a Riemannian submersion between manifolds $\widetilde{\mathcal{M}}$ and $\mathcal{M}$, and $V, W \in T\mathcal{M}$ vector-fields. Then*

$$\nabla^{\widetilde{\mathcal{M}}}_{\widetilde{V}}\widetilde{W} = \widetilde{(\nabla^{\mathcal{M}}_{V}W)} + \frac{1}{2}[\widetilde{V}, \widetilde{W}]^{\perp}$$

where $\widetilde{Z} \in H\widetilde{\mathcal{M}}$ denotes the horizontal lift of $Z \in T\mathcal{M}$ to the horizontal bundle $H\widetilde{\mathcal{M}} = \{H_{\tilde{p}}\widetilde{\mathcal{M}} : \tilde{p} \in \widetilde{\mathcal{M}}\}$, $Z = d\Phi\left(\widetilde{Z}\right)$, and $[\widetilde{V}, \widetilde{W}]^{\perp}$ is the the Lie bracket $[\widetilde{V}, \widetilde{W}] = \widetilde{V} \circ \widetilde{W} - \widetilde{W} \circ \widetilde{V}$ orthogonally projected $(\cdot)^{\perp} : T\widetilde{\mathcal{M}} \to ker\,(d\Phi)$ to the vertical space.

Corollary 1. *Let $\Phi : \widetilde{\mathcal{M}} \to \mathcal{M}$ be a Riemannian submersion and $c : (\tau_0, \tau_1) \to \mathcal{M}$ a smooth curve on $\mathcal{M}$ with $\widetilde{c} : (\tau_0, \tau_1) \to \widetilde{\mathcal{M}}$ its horizontal lift, i.e., $\Phi \circ \widetilde{c} = c$, $d\Phi \circ \dot{\widetilde{c}} = \dot{c}$ and $\dot{\widetilde{c}}(\tau) \in H_{\widetilde{c}(\tau)}\widetilde{\mathcal{M}}$ horizontal (i.e. $\dot{\widetilde{c}} = \widetilde{\dot{c}}$). Then*

i) a vector-field $W = d\Phi \circ \widetilde{W} \in T\mathcal{M}$ along c is parallel if and only if

$$\nabla^{\widetilde{\mathcal{M}}}_{\dot{\widetilde{c}}}\widetilde{W} = \frac{1}{2}[\dot{\widetilde{c}}, \widetilde{W}]^{\perp}$$

for the horizontal vector-field $\widetilde{W} \in T\widetilde{\mathcal{M}}$ along $\widetilde{c}$.

ii) c is a geodesic if and only if $\widetilde{c}$ is a geodesic.

While i) yields the basis for confirming the parallel transport computation, ii) is the underlying fact behind the identification of geodesics in form and shape spaces with geodesics of suitably aligned representatives. Note that, while Huckemann et al. (2010) generally

restrict their discussion to finite dimensional manifolds, the theorem does in fact not have this restriction. Based on these preparations, we can now verify the presented parallel transport along the lines of Huckemann et al. (2010), but for forms rather than shapes and explicitly based on a separable Hilbert space $\mathcal{Y}$ rather than on $\mathbb{C}^k$. Note that, while identifying $H_p\mathcal{Y}^* \cong T_{[p]}\mathcal{Y}^*_{/\,\mathrm{Trl}\times\mathrm{Rot}}$ and $\varepsilon \cong d\Phi(\varepsilon)$ in the main manuscript, they are distinguished here for clarity.

Lemma 1. *Let $p, p' \in \mathcal{Y}^*$ with $\langle p, p'\rangle \neq 0$ centered and mutually rotation aligned representatives of forms $[p], [p'] \in \mathcal{Y}^*_{/\,\mathrm{Trl}\times\mathrm{Rot}}$ (i.e. $p = \widetilde{p}^{\,\mathrm{Trl}\times\mathrm{Rot}}$ for notational simplicity and p' accordingly), let $\varepsilon \in T_{[p]}\mathcal{Y}^*_{/\,\mathrm{Trl}\times\mathrm{Rot}}$ with horizontal lift $\tilde{\varepsilon} \in H_p\mathcal{Y}^*$, and let $\Phi : y \mapsto [y]$ denote the quotient map. Then*

$$\mathrm{Transp}_{[p],[p']}(\varepsilon) = d\Phi\left(\tilde{\varepsilon} - \mathrm{Im}(\langle p'/\|p'\|, \tilde{\varepsilon}\rangle)\,\frac{p/\|p\| + p'/\|p'\|}{1 + \langle p/\|p\|, p'/\|p'\|\rangle}\sqrt{\text{-}1}\right) \tag{3}$$

implements the form parallel transport via its horizontal lift.

Proof. For $\langle p, p'\rangle \neq 0$ aligned and centered, the unique unit-speed geodesic (uniqueness can be seen using Corollary 1 ii)) between $[p]$ and $[p']$ is described by $\tau \to [p + \tau\frac{p'-p}{\|p'-p\|}]$. Yet, to simplify the argument, we choose a unit-angular speed parameterization instead. It takes the form $c(\tau) := [\widetilde{c}(\tau)] := [\rho(\tau)\gamma(\tau)]$ with $\gamma(\tau) = \cos(\tau)\beta + \sin(\tau)\beta'$ where $\beta = \frac{p}{\|p\|}$ and $\beta' = \frac{p'-\langle\beta,p'\rangle\beta}{\|p'-\langle\beta,p'\rangle\beta\|} = \frac{\frac{p'}{\|p'\|} - \langle\frac{p}{\|p\|}, \frac{p'}{\|p'\|}\rangle\frac{p}{\|p\|}}{\|\frac{p'}{\|p'\|} - \langle\frac{p}{\|p\|}, \frac{p'}{\|p'\|}\rangle\frac{p}{\|p\|}\|}$ form an orthonormal basis of the real plain containing the horizontal geodesic. With $\widetilde{c}(0) = p$ and $\widetilde{c}(\arccos\langle\frac{p}{\|p\|}, \frac{p'}{\|p'\|}\rangle) = p'$, $\widetilde{c}(\tau)$ describes the line connecting p and p' in polar coordinates. $[\gamma(\tau)]_{\mathrm{Trl}\times\mathrm{Rot}\times\mathrm{Scl}}$ corresponds to the shape geodesic between $[p]_{\mathrm{Trl}\times\mathrm{Rot}\times\mathrm{Scl}}$ and $[p']_{\mathrm{Trl}\times\mathrm{Rot}\times\mathrm{Scl}}$, and $\rho(\tau) = \|\widetilde{c}(\tau)\|$ reflects the size of the geodesic $c(\tau)$. An explicit definition of $\rho(\tau)$ is not needed.

Due to the alignment of p and p', $\dot{\gamma}(\tau)$ and also

$$\widetilde{W}(\tau) := \tilde{\varepsilon} + \mathrm{Im}(\langle\beta', \tilde{\varepsilon}\rangle)\,(\dot{\gamma}(\tau) - \beta')\,\sqrt{\text{-}1} \tag{4}$$

are horizontal along $\widetilde{c}(\tau)$, i.e. $\widetilde{W}(\tau) \in H_{\widetilde{c}(\tau)}\mathcal{Y}^*$ for each τ, if $\tilde{\varepsilon}$ is horizontal, i.e. if $\mathrm{Im}(\langle p, \tilde{\varepsilon}\rangle) = \langle \imath, \tilde{\varepsilon}\rangle = 0$. More concretely, this holds as

$$\begin{aligned}\mathrm{Im}\Big(\langle\widetilde{c}(\tau), \widetilde{W}(\tau)\rangle\Big) &= \rho(\tau)\ \left(\mathrm{Im}(\langle\gamma(\tau), \tilde{\varepsilon}\rangle) + \mathrm{Im}(\langle\beta', \tilde{\varepsilon}\rangle)\,\mathrm{Re}(\langle\gamma(\tau), \dot{\gamma}(\tau) - \beta'\rangle)\right)\\ &\overset{\tilde{\varepsilon}\text{ horizontal}}{=} \rho(\tau)\ \left(\sin(\tau)\,\mathrm{Im}(\langle\beta', \tilde{\varepsilon}\rangle) + \mathrm{Im}(\langle\beta', \tilde{\varepsilon}\rangle)\,(0 - \sin(\tau)\,\underbrace{\|\beta'\|^2}_{=1})\right) = 0\end{aligned}$$

and, obviously, also $\langle \imath, \widetilde{W}(\tau)\rangle = 0$ as this is the case for all involved vectors. Moreover, $\widetilde{W}$ is smooth and $\tilde{\varepsilon} \mapsto \widetilde{W}\left(\arccos\langle\frac{p}{\|p\|}, \frac{p'}{\|p'\|}\rangle\right)$ yields the transport formulated in Equation (3),

which follows from basic trigonometric relations. In detail, it follows from plugging

$$\begin{aligned}
\dot\gamma\left(\arccos\langle\tfrac{p}{\|p\|},\tfrac{p'}{\|p'\|}\rangle\right)-\beta' &= \langle\tfrac{p}{\|p\|},\tfrac{p'}{\|p'\|}\rangle\ \beta' - \sqrt{1-\langle\tfrac{p}{\|p\|},\tfrac{p'}{\|p'\|}\rangle^2}\ \beta - \beta' \\
&= \Big(\langle\tfrac{p}{\|p\|},\tfrac{p'}{\|p'\|}\rangle - 1\Big)\overbrace{\frac{\frac{p'}{\|p'\|}-\langle\frac{p}{\|p\|},\frac{p'}{\|p'\|}\rangle\frac{p}{\|p\|}}{\sqrt{1-\langle\frac{p}{\|p\|},\frac{p'}{\|p'\|}\rangle^2}}}^{\beta'} - \sqrt{1-\langle\tfrac{p}{\|p\|},\tfrac{p'}{\|p'\|}\rangle^2}\ \frac{p}{\|p\|} \\
&= \frac{\Big(\langle\frac{p}{\|p\|},\frac{p'}{\|p'\|}\rangle-1\Big)\frac{p'}{\|p'\|}}{\sqrt{1-\langle\frac{p}{\|p\|},\frac{p'}{\|p'\|}\rangle^2}} \\
&\quad + \frac{-\langle\frac{p}{\|p\|},\frac{p'}{\|p'\|}\rangle^2\frac{p}{\|p\|}+\langle\frac{p}{\|p\|},\frac{p'}{\|p'\|}\rangle\frac{p}{\|p\|}-\Big(1-\langle\frac{p}{\|p\|},\frac{p'}{\|p'\|}\rangle^2\Big)\frac{p}{\|p\|}}{\sqrt{1-\langle\frac{p}{\|p\|},\frac{p'}{\|p'\|}\rangle^2}} \\
&= \frac{-\Big(1-\langle\frac{p}{\|p\|},\frac{p'}{\|p'\|}\rangle\Big)\Big(\frac{p'}{\|p'\|}+\frac{p}{\|p\|}\Big)}{\sqrt{1-\langle\frac{p}{\|p\|},\frac{p'}{\|p'\|}\rangle^2}}
\end{aligned}$$

and

$$\mathrm{Im}(\langle\beta',\tilde\varepsilon\rangle)\overset{\tilde\varepsilon\text{ horizontal}}{=}\frac{\mathrm{Im}(\langle p',\tilde\varepsilon\rangle)}{\sqrt{1-\langle\frac{p}{\|p\|},\frac{p'}{\|p'\|}\rangle^2}}$$

into the definition of $W(\tau)=d\Phi(\widetilde W(\tau))$ using (4).

Hence, due to Corollary 1 i), we mainly need to show

$$\nabla^{\mathcal{Y}^*}_{\dot{\tilde c}}\widetilde W=\frac{1}{2}[\dot{\tilde c},\widetilde W]^\perp \tag{5}$$

where the left-hand side may directly be computed as

$$\nabla^{\mathcal{Y}^*}_{\dot{\tilde c}(\tau)}\widetilde W(\tau)\overset{(1)}{=}\dot{\widetilde W}(\tau)\overset{(4)}{=}-\mathrm{Im}(\langle\beta',\tilde\varepsilon\rangle)\,\gamma(\tau)\sqrt{\text{-}1}$$

since $\ddot\gamma(\tau)=-\gamma(\tau)$.

On the right-hand side, the orthogonal projection of a vector-field $V(\tau):=V_{\widetilde c(\tau)}\in T_{\widetilde c(\tau)}\mathcal{Y}^*$ along $\widetilde c(\tau)$ into the vertical spaces (of which $\{\gamma(\tau),\mathbb{1},\sqrt{\text{-}1}\mathbb{1}\}$ constitute an orthonormal basis) is given by

$$\begin{aligned}
V^\perp(\tau) &= \frac{\mathrm{Re}\big(\langle\sqrt{\text{-}1}\,\widetilde c(\tau),V(\tau)\rangle\big)}{\|\widetilde c(\tau)\|^2}\widetilde c(\tau)\sqrt{\text{-}1}+\langle\mathbb{1},V(\tau)\rangle\,\mathbb{1} \\
&= \frac{\omega^{\mathrm{Rot}}(V(\tau))}{\rho(\tau)}\gamma(\tau)\sqrt{\text{-}1}+\omega^{\mathrm{Trl}}(V(\tau))\,\mathbb{1}
\end{aligned}$$

with the 1-forms ω^{Rot} and ω^{Trl} defined as

$$\begin{aligned}\omega^{\mathrm{Rot}}(V_p) &:= \mathrm{Re}\Big(\langle \sqrt{\text{-}1}\, p, V_p\rangle\Big) = \mathrm{Re}\Big(-\sqrt{\text{-}1}\,\langle p, V_p\rangle\Big)\\ &= \mathrm{Re}\Big(-\sqrt{\text{-}1}\,\Big(\mathrm{Re}(\langle p, V_p\rangle) + \mathrm{Im}(\langle p, V_p\rangle)\,\sqrt{\text{-}1}\Big)\Big)\\ &= \mathrm{Re}\Big(-\sqrt{\text{-}1}\,\mathrm{Re}(\langle p, V_p\rangle) + \mathrm{Im}(\langle p, V_p\rangle)\Big)\\ &= \mathrm{Im}(\langle p, V_p\rangle)\,.\end{aligned}$$

and $\omega^{\mathrm{Trl}}(V_p) = \langle \ell, V_p\rangle$ for $p \in \mathcal{Y}^*$.

Thus, to confirm (5) and complete the proof, it remains to show $\omega^{\mathrm{Rot}}([\dot{\tilde{c}}, \widetilde{W}]) = -2\,\mathrm{Im}(\langle \beta', \tilde{\varepsilon}\rangle)\,\rho$ and $\omega^{\mathrm{Trl}}([\dot{\tilde{c}}, \widetilde{W}]) = 0$. For this, we use some statements on the exterior derivative $d\omega$ of a 1-form ω subsumed in the following auxiliary lemma (proven later):

Lemma 2. *Let V, W be smooth vector-fields.*

i) For any smooth 1-form ω it holds that $\omega\,([V,W]) = V\,(\omega\,(W)) - W\,(\omega\,(V)) - d\omega\,(V,W)$.

ii) For ω^{Rot} defined above, $d\omega^{\mathrm{Rot}}(V,W) = 2\ \mathrm{Im}(\langle V, W\rangle)$.

iii) For ω^{Trl} defined above, $d\omega^{\mathrm{Trl}}(V,W) = 0$.

Using further that

$$\omega^{\mathrm{Rot}}\left(\dot{\tilde{c}}(\tau)\right) = \mathrm{Im}(\langle \gamma(\tau), \rho(\tau)\,\dot{\gamma}(\tau)\rangle) + \mathrm{Im}(\langle \gamma(\tau), \dot{\rho}(\tau)\,\gamma(\tau)\rangle) = 0$$

and

$$\begin{aligned}\omega^{\mathrm{Rot}}\left(\widetilde{W}(\tau)\right) &= \rho(\tau)\,\mathrm{Im}(\langle \gamma(\tau), \tilde{\varepsilon}\rangle) + \rho(\tau)\,\mathrm{Im}(\langle \beta', \tilde{\varepsilon}\rangle)\,\underbrace{\langle \gamma(\tau), \dot{\gamma}(\tau) - \beta'\rangle}_{\in \mathbb{R}}\\ &= \rho(\tau)\,(\sin(\tau)\,\mathrm{Im}(\langle \beta', \tilde{\varepsilon}\rangle) - \mathrm{Im}(\langle \beta', \tilde{\varepsilon}\rangle)\sin(\tau)) = 0\end{aligned}$$

we then have

$$\begin{aligned}\omega^{\mathrm{Rot}}([\dot{\tilde{c}}(\tau), \widetilde{W}(\tau)]) &= \underbrace{\dot{\tilde{c}}(\tau)\left(\omega^{\mathrm{Rot}}(\widetilde{W}(\tau))\right)}_{=0} - \underbrace{\widetilde{W}(\tau)\left(\omega^{\mathrm{Rot}}(\dot{\tilde{c}}(\tau))\right)}_{=0} - d\omega^{\mathrm{Rot}}\left(\dot{\tilde{c}}(\tau), \widetilde{W}(\tau)\right)\\ &= -2\,\Big(\mathrm{Im}\Big(\langle \rho(\tau)\dot{\gamma}(\tau), \widetilde{W}(\tau)\rangle\Big) + \underbrace{\mathrm{Im}\Big(\langle \dot{\rho}(\tau)\gamma(\tau), \widetilde{W}(\tau)\rangle\Big)}_{=0\ (\widetilde{W}\text{ horizontal, }\rho\text{ and }\dot{\rho}\text{ real})}\Big)\\ &= -2\rho(\tau)\ \mathrm{Im}\Big(\langle \cos(\tau)\beta' - \sin(\tau)\beta, \tilde{\varepsilon}\rangle + \langle \dot{\gamma}(\tau), \mathrm{Im}(\langle \beta', \tilde{\varepsilon}\rangle)\,(\dot{\gamma}(\tau) - \beta')\,\sqrt{\text{-}1}\rangle\Big)\\ &= -2\rho(\tau)\,\Big(\cos(\tau)\,\mathrm{Im}(\langle \beta', \tilde{\varepsilon}\rangle) + \mathrm{Im}(\langle \beta', \tilde{\varepsilon}\rangle)\,\underbrace{\langle \dot{\gamma}(\tau), \dot{\gamma}(\tau) - \beta'\rangle}_{\in \mathbb{R}}\Big)\\ &= -2\rho(\tau)\,\Big(\cos(\tau)\,\mathrm{Im}(\langle \beta', \tilde{\varepsilon}\rangle) + \mathrm{Im}(\langle \beta', \tilde{\varepsilon}\rangle) - \mathrm{Im}(\langle \beta', \tilde{\varepsilon}\rangle)\cos(\tau)\Big)\\ &= -2\rho(\tau)\ \mathrm{Im}(\langle \beta', \tilde{\varepsilon}\rangle)\end{aligned}$$

and

$$\omega^{\mathrm{Trl}}([\dot{\widetilde{c}}(\tau),\widetilde{W}(\tau)]) = \dot{\widetilde{c}}(\tau)\left(\underbrace{\omega^{\mathrm{Trl}}\left(\widetilde{W}(\tau)\right)}_{=\langle \ell,\widetilde{W}\rangle=0}\right) - \widetilde{W}(\tau)\left(\underbrace{\omega^{\mathrm{Trl}}\left(\dot{\widetilde{c}}(\tau)\right)}_{=\langle \ell,\dot{\widetilde{c}}\rangle=0}\right) - \underbrace{d\omega^{\mathrm{Trl}}\left(\dot{\widetilde{c}}(\tau),\widetilde{W}(\tau)\right)}_{=0} = 0,$$

where tangent vectors $\dot{\widetilde{c}}(\tau)$ and $\widetilde{W}(\tau)$ are interpreted as directional derivatives. These are the two equations that remained to show. □

Proof of Lemma 2. i) See, e.g., Lee (2018), Proposition B.12 on page 402. This is a standard result. Note that based on an alternative (yet also common) definition of the wedge product and, hence, the exterior derivative, Huckemann et al. (2010) and others write $\omega([V,W]) = V(\omega(W)) - W(\omega(V)) - 2\,d\omega(V,W)$ instead. In this case, we also have $d\omega^{\mathrm{Rot}}(V,W) = \mathrm{Im}(\langle V,W\rangle)$ in ii) compensating for the different factor in the proof of Lemma 1.

ii) Let $\{e_r\}_r$ be an orthonormal $\mathbb{C}$-basis of $\mathcal{Y}$ (a complete orthonormal system existing since $\mathcal{Y}$ is separable) and $\{\vartheta^{(r)}(y)\}_r = \langle e_r, y\rangle$ the corresponding dual basis. The tangent vectors $\partial_{\mathrm{Re},r}\big|_p \cong e_r$ and $\partial_{\mathrm{Im},r}\big|_p \cong \sqrt{\text{-}1}\,e_r$, $p \in \mathcal{Y}$ together form an $\mathbb{R}$-basis of $T_p\mathcal{Y}^* \cong \mathcal{Y}$. The dual 1-forms are given by $d^{\mathrm{Re},r}(V_p) := V_p\left(\mathrm{Re}\circ\vartheta^{(r)}\right) \cong \mathrm{Re}\circ\vartheta^{(r)}(V_p)$ and $d^{\mathrm{Im},r}(V_p) := V_p\left(\mathrm{Im}\circ\vartheta^{(r)}\right) \cong \mathrm{Im}\circ\vartheta^{(r)}(V_p)$ where we identify tangent vectors either with directional derivatives $V_p(f) = \frac{d}{d\tau}\left(f\circ\mathrm{Exp}_p(\tau V_p)\right)\big|_{\tau=0}$ of functions $f:\mathcal{M}\to\mathbb{R}$ or with elements of $\mathcal{Y}$, and the equality follows from $\mathcal{M}=\mathcal{Y}^*$, and $\mathrm{Re}\circ\vartheta^{(r)}$, $\mathrm{Im}\circ\vartheta^{(r)}$ linear. With this given, we have

$$\begin{aligned}
\omega^{\mathrm{Rot}}(V_p) &= \mathrm{Im}\left(\langle\sum_r \langle e_r,p\rangle e_r, V_p\rangle\right) \qquad (6)\\
&= \sum_r \mathrm{Im}(\langle p,e_r\rangle\langle e_r,V_p\rangle)\\
&= \sum_r \mathrm{Re}(\langle e_r,p\rangle)\,\mathrm{Im}(\langle e_r,V_p\rangle) - \mathrm{Im}(\langle e_r,p\rangle)\,\mathrm{Re}(\langle e_r,V_p\rangle)\\
&= \sum_r \mathrm{Re}\circ\vartheta^{(r)}(p)\,d^{\mathrm{Im},j}(V_p) - \mathrm{Im}\circ\vartheta^{(r)}(p)\,d^{\mathrm{Re},j}(V_p)
\end{aligned}$$

and thus, expressing the exterior derivative in terms of wedge products

$$\begin{aligned}
d\omega^{\mathrm{Rot}} &= \sum_r \sum_l \partial_{\mathrm{Re},r} \left(\mathrm{Re} \circ \vartheta^{(r)}\right)\ d^{\mathrm{Re},l} \wedge d^{\mathrm{Im},r} + \partial_{\mathrm{Im},r} \left(\mathrm{Re} \circ \vartheta^{(r)}\right)\ d^{\mathrm{Im},l} \wedge d^{\mathrm{Im},r} \\
&\quad - \partial_{\mathrm{Re},r} \left(\mathrm{Im} \circ \vartheta^{(r)}\right)\ d^{\mathrm{Re},l} \wedge d^{\mathrm{Re},r} - \partial_{\mathrm{Im},r} \left(\mathrm{Im} \circ \vartheta^{(r)}\right)\ d^{\mathrm{Im},l} \wedge d^{\mathrm{Re},r} \\
&= \sum_r \sum_l d^{\mathrm{Re},l} \left(\partial_{\mathrm{Re},r}\right)\ d^{\mathrm{Re},l} \wedge d^{\mathrm{Im},r} + d^{\mathrm{Re},l} \left(\partial_{\mathrm{Im},r}\right) d^{\mathrm{Im},l} \wedge d^{\mathrm{Im},r} \\
&\quad - d^{\mathrm{Im},l} \left(\partial_{\mathrm{Re},r}\right) d^{\mathrm{Re},l} \wedge d^{\mathrm{Re},r} - d^{\mathrm{Im},l} \left(\partial_{\mathrm{Im},r}\right) d^{\mathrm{Im},l} \wedge d^{\mathrm{Re},r} \\
&= \sum_r d^{\mathrm{Re},r} \wedge d^{\mathrm{Im},r} - d^{\mathrm{Im},r} \wedge d^{\mathrm{Re},r} \\
&= 2 \sum_r d^{\mathrm{Re},r} \wedge d^{\mathrm{Im},r}
\end{aligned}$$

which evaluates to

$$\begin{aligned}
d\omega^{\mathrm{Rot}}(V, W) &= 2 \sum_r \left(d^{\mathrm{Re},r}(V)\, d^{\mathrm{Im},r}(W) - d^{\mathrm{Im},r}(V)\, d^{\mathrm{Re},r}(W) \right) \\
&= 2\ \mathrm{Im}(\langle V, W \rangle)
\end{aligned}$$

where the last equation follows from a computation analogous to (6).

iii) By choosing w.l.o.g. $e_1 = \imath$, we obtain

$$\omega^{\mathrm{Trl}}(V) = d^{\mathrm{Re},1} + \sqrt{\text{-}1}\, d^{\mathrm{Im},1}$$

which immediately yields $d\langle \imath, \cdot \rangle = 0$, since $d\, d^{\mathrm{Re},1} = d\, d^{\mathrm{Im},1} = 0$.

□

S.2 Tensor-product factorization

The optimality of the proposed tensor-product factorization follows from the Eckart-Young-Mirsky theorem (EYM) which can be found, e.g., in (Gentle, 2007, page 139) for matrices and, in more general terms, in (Hsing and Eubank, 2015, page 111) for Hilbert-Schmidt operators. In the following, we present a tensor-product version of EYM designed for our needs. The optimality of the tensor-product factorization is then illustrated in two corollaries – first in a theoretical model setting and second for the empirical decomposition on evaluations which can be practically conducted on given data. Consider two real vector spaces $\mathcal{B}_j$, $j \in \{0, 1\}$, with positive semi-definite bilinear forms $\langle \cdot, \cdot \rangle_j : \mathcal{B}_j \times \mathcal{B}_j \to \mathbb{R}$ inducing semi-norms $\| \cdot \|_j$. Assuming $\mathcal{B}_1$ to be, in fact, a function space of functions $f : \mathcal{X} \to \mathbb{R}$ on some set $\mathcal{X}$, the (vector space) tensor product $\mathcal{B}_1 \otimes \mathcal{B}_0$ of $\mathcal{B}_0$ and $\mathcal{B}_1$ is the vector space spanned by all $f \otimes y : \mathcal{X} \to \mathcal{B}_0, \mathbf{x} \mapsto f(\mathbf{x})\, y$ with $f \in \mathcal{B}_1$ and $y \in \mathcal{B}_0$. By linear extension, a symmetric positive semi-definite bilinear form on $\mathcal{B}_1 \otimes \mathcal{B}_0$ is defined by $\langle f \otimes y\,,\ f' \otimes y' \rangle_{\mathcal{B}_1 \otimes \mathcal{B}_0} = \langle f, f' \rangle_1 \langle y, y' \rangle_0$ for all $f, f' \in \mathcal{B}_1, y, y' \in \mathcal{B}_0$. It induces a semi-norm $\| \cdot \|_{\mathcal{B}_1 \otimes \mathcal{B}_0}$ on the tensor product space.

Theorem 2 (Eckart-Young-Mirsky for finite-dimensional tensor-products)**.** *Let $\mathcal{B}_0, \mathcal{B}_1$ be semi-normed vector spaces as defined above and $h = \sum_{r=1}^{m_0}\sum_{l=1}^{m_1}\theta^{(r,l)}\, b_1^{(l)} \otimes b_0^{(r)} \in \mathcal{B}_1 \otimes \mathcal{B}_0$ expressed as a finite linear-combination with $b_1^{(1)}, \dots, b_1^{(m_1)} \in \mathcal{B}_1$, $b_0^{(1)}, \dots, b_0^{(m_0)} \in \mathcal{B}_0$, and coefficient matrix $\{\theta^{(r,l)}\}_{r,l} = \boldsymbol{\Theta} \in \mathbb{R}^{m_0 \times m_1}$. Then we can optimally decompose $h = \sum_{r=1}^{m} d^{(r)}\, \xi_1^{(r)} \otimes \xi_0^{(r)}$ with $m = \min\{m_0, m_1\}$, $d^{(1)} \geq \cdots \geq d^{(m)} \geq 0$ and $\langle \xi_j^{(r)}, \xi_j^{(l)} \rangle_j = \mathbb{1}(r = l)$ for $\xi_j^{(r)} \in \mathcal{B}_j$, in the sense that for any $L \leq m$*

$$\| h - \sum_{r=1}^{L} d^{(r)}\, \xi_1^{(r)} \otimes \xi_0^{(r)} \|_{\mathcal{B}_1 \otimes \mathcal{B}_0} \;\leq\; \| h - \sum_{r=1}^{L} d_\star^{(r)}\, \xi_{1\star}^{(r)} \otimes \xi_{0\star}^{(r)} \|_{\mathcal{B}_1 \otimes \mathcal{B}_0} \tag{7}$$

for all $d_\star^{(r)} \in \mathbb{R}$ and $\xi_{j\star}^{(r)} \in \mathcal{B}_j$, $j \in \{0,1\}$, $r = 1, \dots, L$. Arranging $\mathbf{D} = \operatorname{diag}(d^{(1)}, \dots, d^{(m)})$ and expressing $\xi_j^{(r)} = \sum_l u_j^{(l,r)} b_j^{(l)}, r = 1, \dots, m$, with the coefficient matrices $\{u_j^{(l,r)}\}_{l,r} = \mathbf{U}_j \in \mathbb{R}^{m_j \times m}$, $j \in \{0,1\}$, an optimal decomposition is obtained as follows:

i) If for $j \in \{0,1\}$ the Gram matrices $\mathbf{G}_j = \{\langle b_j^{(r)}, b_j^{(l)} \rangle_j\}_{r,l}$ are the identity $\mathbf{G}_j = \mathbf{I}_{m_j}$, the matrices $\mathbf{D}$ and $\mathbf{U}_j$, $j \in \{0,1\}$, are directly determined via SVD of the coefficient matrix $\boldsymbol{\Theta} = \mathbf{U}_0 \mathbf{D} \mathbf{U}_1^\top$.

ii) In general, there are suitable matrices $\mathbf{M}_j \in \mathbb{R}^{\operatorname{rank} \mathbf{G}_j \times m_j}$, $j \in \{0,1\}$, such that $\boldsymbol{\Xi} = \mathbf{V}_0 \mathbf{D} \mathbf{V}_1^\top$ is the SVD of the matrix $\boldsymbol{\Xi} = \mathbf{M}_0 \boldsymbol{\Theta} \mathbf{M}_1^\top$ and $\mathbf{U}_j = \mathbf{M}_j^- \mathbf{V}_j$ with generalized inverse $\mathbf{M}_j^- = \mathbf{M}_j^\top (\mathbf{M}_j \mathbf{M}_j^\top)^{-1}$.

(a) In general, a suitable matrix is given by $\mathbf{M}_j = \sqrt{\mathbf{G}_j}^\top$ with $\mathbf{G}_j = \sqrt{\mathbf{G}_j}\sqrt{\mathbf{G}_j}^\top$ a Cholesky decomposition.

(b) If the $b_j^{(r)}$ can be identified with vectors $\mathbf{b}_j^{(r)} \in \mathbb{R}^{m_j'}$ of some length m_j', arranged as column vectors of a "design matrix" $\mathbf{B}_j \in \mathbb{R}^{m_j' \times m_j}$, such that $\langle b_j^{(r)}, b_j^{(l)} \rangle_j = (\mathbf{b}_j^{(r)})^\top \mathbf{W}_j \mathbf{b}_j^{(l)}$, with $r, l = 1, \dots, m_j$, for a symmetric positive definite weight matrix $\mathbf{W}_j$, we may equivalently set $\mathbf{M}_j = \mathbf{R}_j$ based on the QR-decomposition $\sqrt{\mathbf{W}_j}^\top \mathbf{B}_j = \mathbf{Q}_j \mathbf{R}_j$. In this case, design matrices $\mathbf{E}_j$ of vector representatives $\boldsymbol{\xi}_j^{(r)}$ for the $\xi_j^{(r)}$ can, alternatively, be obtained as $\mathbf{E}_j = \sqrt{\mathbf{W}_j}^{-\top} \mathbf{Q}_j \mathbf{V}_j$ (where $\mathbf{W}_j$ is typically diagonal and, hence, $\sqrt{\mathbf{W}_j}^{-\top}$ fast to compute).

Proof. *i)* For $j \in \{0,1\}$, denote the column vectors of $\mathbf{U}_j$ by $\mathbf{u}_j^{(r)}$, $r = 1, \dots, m$, and consider the space of $m_0 \times m_1$ matrices equipped with the inner product $\langle \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2 \rangle_F = \operatorname{tr}\left(\boldsymbol{\Theta}_1^\top \boldsymbol{\Theta}_2\right)$, for $\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2 \in \mathbb{R}^{m_0 \times m_1}$, inducing the Frobenius norm $\| \cdot \|_F$.

The EYM for matrices (e.g. Gentle, 2007, page 139) states that the matrix $\boldsymbol{\Theta}_L = \sum_{r=1}^{L} d^{(r)} \mathbf{u}_0^{(r)} (\mathbf{u}_1^{(r)})^\top$ is the best rank L approximation of $\boldsymbol{\Theta}$, in the sense that

$$\| \boldsymbol{\Theta} - \boldsymbol{\Theta}_L \|_F \leq \| \boldsymbol{\Theta} - \sum_{r=1}^{L} d_\star^{(r)} \mathbf{u}_{0\star}^{(r)} (\mathbf{u}_{1\star}^{(r)})^\top \|_F \text{ for any } d_\star^{(r)} \in \mathbb{R}, \mathbf{u}_{j\star}^{(r)} \in \mathbb{R}^{m_j}, r = 1, \dots, m.$$

To apply the theorem, we point out that, provided the Gram matrices $\mathbf{G}_j = \mathbf{I}_{m_j}$, the $\{b_j^{(r)}\}_{r=1,\dots,m_j}$ and, thus, also $\{\, b_1^{(r)} \otimes b_0^{(l)} \,\}_{r,l}$ are orthonormal bases of finite-dimensional subspaces $\mathcal{A}_j \subset \mathcal{B}_j$ and $\mathcal{A}_1 \otimes \mathcal{A}_0 \subset \mathcal{B}_1 \otimes \mathcal{B}_0$, respectively, forming Hilbert spaces. Hence, the basis representation map sending $\xi_j^{(r)} \mapsto \mathbf{u}_j^{(r)}$ to its coefficient vector w.r.t. $\{b_j^{(r)}\}_r$ presents an isometric isomorphism from $\mathcal{A}_j$ to $\mathbb{R}^{m_j}$. Accordingly, the basis representation $\mathcal{A}_1 \otimes \mathcal{A}_0 \to \mathbb{R}^{m_0 \times m_1}$, $h \mapsto \mathbf{\Theta}$ presents an isometric isomorphism identifying $\xi_1^{(r)} \otimes \xi_0^{(l)}$ with $\mathbf{u}_0^{(l)} (\mathbf{u}_1^{(r)})^\top$. The isometry follows from $\langle h_1, h_2 \rangle = \sum_{r,l,r',l'} \theta_1^{(r,l)} \theta_2^{(r',l')} \langle\, b_1^{(r)} \otimes b_0^{(l)} \,,\, b_1^{(r')} \otimes b_0^{(l')} \,\rangle = \sum_{r,l} \theta_1^{(r,l)} \theta_2^{(r,l)} = \operatorname{tr}\left(\mathbf{\Theta}_1^\top \mathbf{\Theta}_2\right)$ for basis representations $h_1 \mapsto \mathbf{\Theta}_1$ and $h_2 \mapsto \mathbf{\Theta}_2$. This lets us carry over the EYM for matrices to $\mathcal{A}_1 \otimes \mathcal{A}_0$ yielding the desired inequality (7) restricted to $\xi_{j\star}^{(r)} \in \mathcal{A}_j \subset \mathcal{B}_j$. The property $d_1 \geq \cdots \geq d_m \geq 0$ and orthonormality of the $\xi_j^{(r)}$ are also inherited from the SVD.

Moreover, we can project any $\xi_{j\star}^{(r)} \in \mathcal{B}_j$ as $\xi_{j\|}^{(r)} = \sum_l \langle b_j^{(l)}, \xi_{j\star}^{(r)} \rangle_j \, b_j^{(l)}$ into $\mathcal{A}_j$ and define $\xi_{j\perp}^{(r)} = \xi_{j\star}^{(r)} - \xi_{j\|}^{(r)}$, which yields an analogous decomposition $h_\star = \sum_{r=1}^L d_\star^{(r)} \, \xi_{1\star}^{(r)} \otimes \xi_{0\star}^{(r)} = h_\| + h_\perp$ with $h_\| \in \mathcal{A}_1 \otimes \mathcal{A}_0$ and $\langle h, h_\perp \rangle_{\mathcal{B}_1 \otimes \mathcal{B}_0} = \langle h_\|, h_\perp \rangle_{\mathcal{B}_1 \otimes \mathcal{B}_0} = 0$. Thus, we have $\|h - h_\star\|^2_{\mathcal{B}_1 \otimes \mathcal{B}_0} = \|h - h_\|\|^2_{\mathcal{B}_1 \otimes \mathcal{B}_0} + \|h_\perp\|^2_{\mathcal{B}_1 \otimes \mathcal{B}_0} \geq \|h - h_\|\|^2_{\mathcal{B}_1 \otimes \mathcal{B}_0} \underset{\text{on } \mathcal{A}_1 \otimes \mathcal{A}_0}{\overset{\text{EYM}}{\geq}} \|h - \sum_{r=1}^L d^{(r)} \, \xi_1^{(r)} \otimes \xi_0^{(r)} \|_{\mathcal{B}_1 \otimes \mathcal{B}_0}$, which completes the proof.

ii) We represent $b_j^{(r)} = \sum_{l=1} M_j^{(l,r)} a_j^{(l)}$ in an orthonormal basis $\{a_j^{(l)}\}_l$ of the Hilbert space $\mathcal{A}_j \subset \mathcal{B}_j$ spanned by $\{b_j^{(r)}\}_r$ as in *i)* with the coefficients forming the matrix $\mathbf{M}_j = \{M_j^{(l,r)}\}_{l,r}$, for $j \in \{0,1\}$, such that $\mathbf{\Xi} = \mathbf{M}_0 \mathbf{\Theta} \mathbf{M}_1^\top$ is the coefficient matrix of h w.r.t. $\{a_j^{(l)}\}_l$. Hence, due to *i)*, the matrices $\mathbf{\Xi} = \mathbf{V}_0 \mathbf{D} \mathbf{V}_1^\top$ obtained by SVD fulfill the desired properties where the $\mathbf{V}_j$ are the coefficient matrices of the $\{\xi_j^{(r)}\}_r$ w.r.t. $\{a_j^{(r)}\}_r$. We may set $\mathbf{U}_j = \mathbf{M}_j^- \mathbf{V}_j$ to represent $\{\xi_j^{(r)}\}_r$ in the original basis $\{b_j^{(r)}\}_r$ instead, since, due to $\mathbf{M}_j \mathbf{M}_j^- = \mathbf{I}_{\operatorname{rank} \mathbf{G}_j}$, we have $a_j^{(r)} = \sum_{l=1} M_j^{-(l,r)} b_j^{(l)}$ for $\mathbf{M}^- = \{M_j^{-(l,r)}\}_{l,r}$.

a) Constructing the orthonormal basis $\{a_j^{(r)}\}_r$ via $a_j^{(r)} = \sum_{l=1} M_j^{-(l,r)} b_j^{(l)}$ with $\mathbf{M}_j^- = \{M_j^{-(l,r)}\}_{l,r} = \sqrt{\mathbf{G}_j}^{\top -}$ is straight forward yielding

$$\begin{aligned}
\{\langle a_j^{(r)}, a_j^{(l)} \rangle\}_{r,l} &= \mathbf{M}_j^{-\top} \mathbf{G}_j \mathbf{M}_j^- \\
&= \left(\sqrt{\mathbf{G}_j}^\top \sqrt{\mathbf{G}_j}\right)^{-1} \sqrt{\mathbf{G}_j}^\top \sqrt{\mathbf{G}_j} \sqrt{\mathbf{G}_j}^\top \sqrt{\mathbf{G}_j} \left(\sqrt{\mathbf{G}_j}^\top \sqrt{\mathbf{G}_j}\right)^{-1} \\
&= \mathbf{I}_{\operatorname{rank} \mathbf{G}_j}.
\end{aligned}$$

b) As in this case, $\mathbf{G}_j = \mathbf{B}_j^\top \sqrt{\mathbf{W}_j} \sqrt{\mathbf{W}_j}^\top \mathbf{B}_j = \mathbf{R}_j^\top \mathbf{Q}_j^\top \mathbf{Q}_j \mathbf{R}_j \overset{\mathbf{Q}_j \text{ orthogonal}}{=} \mathbf{R}_j^\top \mathbf{R}_j$ the choice $\mathbf{M}_j = \mathbf{R}_j$ is equivalent to *a)*. Accordingly, $\mathbf{U}_j = \mathbf{R}_j^- \mathbf{V}_j$ and thus $\mathbf{E}_j = \mathbf{B}_j \mathbf{U}_j = \mathbf{B}_j \mathbf{R}_j^- \mathbf{V}_j = \sqrt{\mathbf{W}}^{-\top} \mathbf{Q}_j \mathbf{V}_j$.

□

Corollary 2 (Tensor-product factorization)**.** *Let $\{b_0^{(r)}\}_{r=1,\dots,m_0}$ elements of a Hilbert space $\mathcal{Y}$ with norm $\|\cdot\|$ and $b_1^{(l)} \in \mathcal{L}^2(\mathcal{X}) = \{f : \mathcal{X} \to \mathbb{R} : f \circ \mathbf{X}$ measurable, $\mathbb{E}(\|f(\mathbf{X})\|^2) < \infty\}$, $l = 1, \dots, m_1$, square-integrable functions of a random covariate vector $\mathbf{X}$ taking values in $\mathcal{X}$. Let further $h(\mathbf{x}) = \sum_{r=1}^{m_0} \sum_{l=1}^{m_1} \theta^{(r,l)}\, b_1^{(l)}(\mathbf{x})\, b_0^{(r)}$ for $\mathbf{x} \in \mathcal{X}$. Then we can optimally decompose $h(\mathbf{x}) = \sum_{r=1}^{m} h^{(r)}(\mathbf{x})\, \xi^{(r)}$ with $m = \min\{m_0, m_1\}$, $\xi^{(1)}, \dots, \xi^{(m)}$ orthonormal and $h^{(r)} \in \mathcal{L}^2(\mathcal{X})$ with $\mathbb{E}\left(h^{(1)}(\mathbf{X})^2\right) \geq \cdots \geq \mathbb{E}\left(h^{(m)}(\mathbf{X})^2\right)$, in the sense that for any $L \leq m$*

$$\mathbb{E}\left(\|h(X) - \sum_{r=1}^{L} h^{(r)}(X)\, \xi^{(r)}\|^2\right) \leq \mathbb{E}\left(\|h(X) - \sum_{r=1}^{L} h_\star^{(r)}(X)\, \xi_\star^{(r)}\|^2\right),$$

for any other $\xi_\star^{(r)} \in \mathcal{Y}$ and $h_\star^{(r)} \in \mathcal{L}^2(\mathcal{X})$, $r = 1, \dots, L$. An optimal decomposition is obtained by specifying $\xi^{(r)} = \xi_0^{(r)}$ and $h^{(r)} = d^{(r)}\, \xi_1^{(r)}$ as in Theorem 2 with $\langle \cdot, \cdot \rangle_0 = \langle \cdot, \cdot \rangle$ the inner product of $\mathcal{Y}$ and $\langle f, f' \rangle_1 = \mathbb{E}(f(X)\, f'(X))$ for $f, f' \in \mathcal{L}^2(\mathcal{X})$.

Proof. After applying Theorem 2, it remains to check that $\|h\|^2_{\mathcal{L}^2(\mathcal{X}) \otimes \mathcal{Y}} = \mathbb{E}(\|h(X)\|^2)$. Indeed, this holds for all simple $h = f \otimes y$, since

$$\langle\, y \otimes f\, ,\, y' \otimes f' \,\rangle_{\mathcal{L}^2(\mathcal{X}) \otimes \mathcal{Y}} = \langle y, y' \rangle\, \mathbb{E}(f(X)\, f'(X)) = \mathbb{E}(\langle f(X)\, y, f'(X)\, y' \rangle)$$

for any $y, y' \in \mathcal{Y}$ and $f, f' \in \mathcal{L}^2(\mathcal{X})$, and, therefore, carries over to all $h \in \mathcal{L}^2(\mathcal{X}) \otimes \mathcal{Y}$ in the vector space. □

Corollary 3 (Tensor-product factorization, empirical version)**.** *Let $\mathcal{F}(\mathcal{X}, \mathbb{R})$ and $\mathcal{F}(\mathcal{T}, \mathbb{C})$ denote the sets of functions $\mathcal{X} \to \mathbb{R}$ and $\mathcal{T} \to \mathbb{C}$, respectively, which are both considered real vector spaces. Let $b_0^{(r)} \in \mathcal{F}(\mathcal{T}, \mathbb{C})$, $r = 1, \dots, m_0$, and $b_1^{(l)} \in \mathcal{F}(\mathcal{X}, \mathbb{R})$, $l = 1, \dots, m_1$. Consider $h(\mathbf{x})(t) = \sum_{r=1}^{m_0} \sum_{l=1}^{m_1} \theta^{(r,l)}\, b_1^{(l)}(\mathbf{x})\, b_0^{(r)}(t)$ for $\mathbf{x} \in \mathcal{X}$, $t \in \mathcal{T}$ evaluated, for $i = 1, \dots, n$, at $\mathbf{x}_i \in \mathcal{X}$ and $t_{i,\iota} \in \mathcal{T}, \iota = 1, \dots, k_i$. Then we can decompose $h(\mathbf{x}) = \sum_{r=1}^{m} h^{(r)}(\mathbf{x})\, \xi^{(r)}$ with $m = \min\{m_0, m_1\}$ optimally, in the sense that for any $L \leq m$ and any other functions $\xi_\star^{(r)} : \mathcal{T} \to \mathbb{C}$ and $h_\star^{(r)} : \mathcal{X} \to \mathbb{R}$, $r = 1, \dots, L$,*

$$\begin{aligned} &\sum_{i=1}^{n} w_{1i} \frac{1}{n} \sum_{\ddot{\imath}=1}^{n} \sum_{\iota=1}^{k_i} w_{0\ddot{\imath}\iota} |h(\mathbf{x}_i)(t_{\ddot{\imath}\iota}) - \sum_{r=1}^{L} h^{(r)}(\mathbf{x}_i)\, \xi^{(r)}(t_{\ddot{\imath}\iota})|^2 \\ &\qquad\qquad \leq \\ &\sum_{i=1}^{n} w_{1i} \frac{1}{n} \sum_{\ddot{\imath}=1}^{n} \sum_{\iota=1}^{k_i} w_{0\ddot{\imath}\iota} |h(\mathbf{x}_i)(t_{\ddot{\imath}\iota}) - \sum_{r=1}^{L} h_\star^{(r)}(\mathbf{x}_i)\, \xi_\star^{(r)}(t_{\ddot{\imath}\iota})|^2, \end{aligned} \tag{8}$$

with integration/sample weights $w_{0i\iota} \geq 0$ and $w_{1i} \geq 0$. An optimal decomposition is obtained by specifying $\xi^{(r)} = \xi_0^{(r)}$ and $h^{(r)} = d^{(r)}\, \xi_1^{(r)}$, $r = 1, \dots, m$, specified as in Theorem 2 with $\langle y, y' \rangle_0 = \frac{1}{n} \sum_{i=1}^{n} \sum_{\iota=1}^{k_i} w_{0i\iota}\, \mathrm{Re}\big(y^\dagger(t_{i\iota}) y'(t_{i\iota})\big)$ for $y, y' \in \mathcal{F}(\mathcal{T}, \mathbb{C})$ and $\langle f, f' \rangle_1 = \sum_{i=1}^{n} w_{1i} f(x_i) f'(x_i)$ for $f, f' \in \mathcal{F}(\mathcal{X}, \mathbb{R})$.

Proof. Again, we confirm $\|h\|^2_{\mathcal{F}(\mathcal{X},\mathbb{R})\otimes\mathcal{F}(\mathcal{T},\mathbb{C})} = \sum_{i=1}^{n} w_{1i} \sum_{\iota=1}^{k_i} w_{0\ddot{\imath}\iota} \left(h(\mathbf{x}_i)(t_{\ddot{\imath}\iota})\right)^2$ by showing

$$\begin{aligned}\langle\, y \otimes f \,,\, y' \otimes f' \,\rangle_{\mathcal{F}(\mathcal{X},\mathbb{R})\otimes\mathcal{F}(\mathcal{T},\mathbb{C})} &= \langle y, y'\rangle_0 \sum_{i=1}^{n} w_{1i} f(x_i) f'(x_i) = \sum_{i=1}^{n} w_{1i} \langle f(x_i)y, f'(x_i)y'\rangle_1 \\ &= \sum_{i=1}^{n} w_{1i} \frac{1}{n} \sum_{\ddot{\imath}=1}^{n} \sum_{\iota=1}^{k_i} w_{0\ddot{\imath}\iota} \operatorname{Re}\Big(\left(f(x_i)y(t_{\ddot{\imath}\iota})\right)^\dagger f'(x_i) y'(t_{\ddot{\imath}\iota}) \Big)\end{aligned}$$

for any $y, y' \in \mathcal{F}(\mathcal{T}, \mathbb{C})$ and $f, f' \in \mathcal{F}(\mathcal{X}, \mathbb{R})$. □

Remark 1. *For the regular case with $k_1 = \cdots = k_n =: k$ and for all $\iota = 1, \ldots, k$ also $t_{i\iota} = t_{1\iota} =: t_\iota$ and $w_{0i\iota} = w_{01\iota} =: w_{0\iota}$ equal for all observations $i = 1, \ldots, n$, Inequality* (8) *simplifies to*

$$\begin{aligned}&\sum_{i=1}^{n} w_{1i} \sum_{\iota=1}^{k_i} w_{0\iota} |h(\mathbf{x}_i)(t_\iota) - \sum_{r=1}^{L} h^{(r)}(\mathbf{x}_i)\, \xi^{(r)}(t_\iota)|^2 \\ &\qquad\qquad \leq \\ &\sum_{i=1}^{n} w_{1i} \sum_{\iota=1}^{k_i} w_{0\iota} |h(\mathbf{x}_i)(t_\iota) - \sum_{r=1}^{L} h_\star^{(r)}(\mathbf{x}_i)\ \xi_\star^{(r)}(t_\iota)|^2.\end{aligned}$$

S.3 Shape differences in astragali of wild and domesticated sheep

Table S1: Distribution of covariate levels over the sheep populations in the data set.

	Sex			Age_group			
	female	male	na	juvenile	subadult	adult	na
Karakul	21	19	1	1	5	35	0
Marsch	18	5	0	5	5	13	0
Soay	21	25	12	7	8	13	30
Wild_sheep	21	20	0	5	18	14	4

	Mobility			Status		
	confined	pastured	free	domestic	feral	wild
Karakul	31	10	0	41	0	0
Marsch	23	0	0	23	0	0
Soay	0	0	58	0	58	0
Wild_sheep	0	0	41	0	0	41

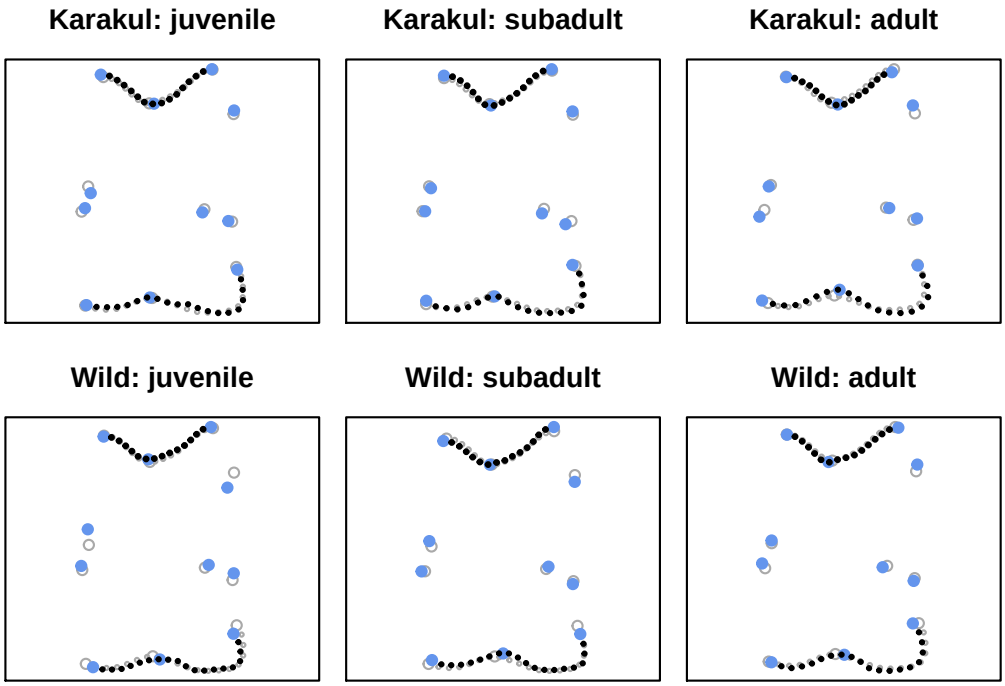

Figure S1: Six example sheep astragalus shape configurations consisting of landmarks *(blue dots)* and semi-landmarks describing two outline curves *(black dots)* recorded in male Karakul and wild sheep of different age. Points are weighted such that the total weight of each curve corresponds to three landmarks *(weights reflected in point-size)*. Shapes are depicted aligned to their overall mean shape *(grey circles)*.

S.4 Cellular Potts model parameter effects on cell form

In the graphics below, the CPM parameters are abbreviated as

- b: bulk stiffness $x_{i1} \in [0.003, 0.015]$
- m: membrane stiffness $x_{i2} \in [0.001, 0.015]$
- a: substrate adhesion $x_{i3} \in [30, 70]$
- r: signaling radius $x_{i4} \in [5, 40]$

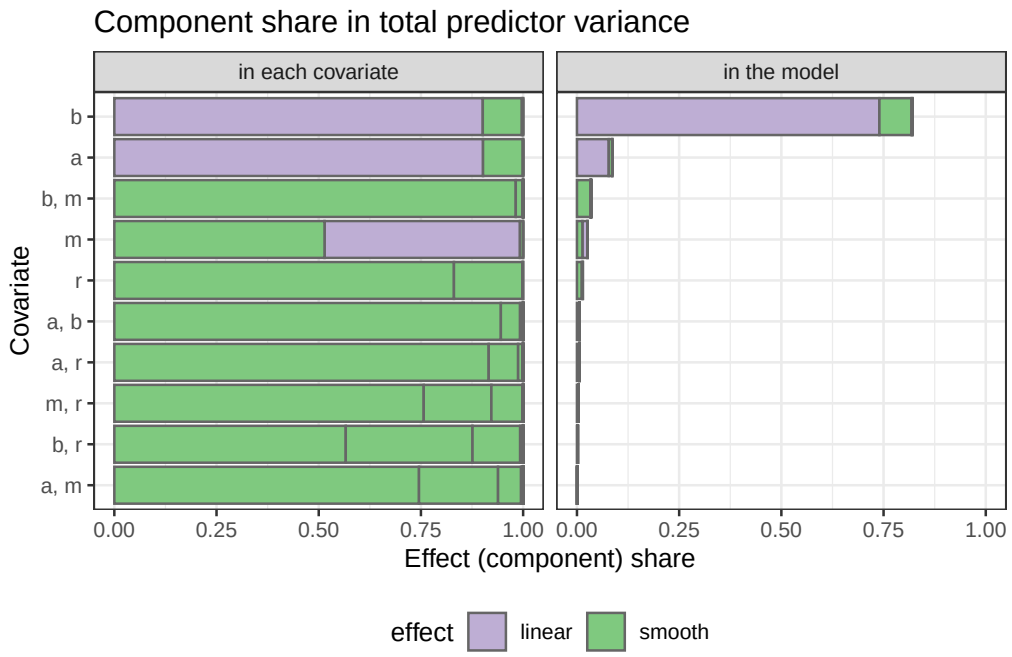

Figure S2: Tensor-product effect factorization: Predictor variance share explained by each effect direction *(separated by vertical lines)* relative to the total predictor variance of the effects of each covariate *(left)* and of the overall model *(right)*. Linear effect components are presented together with the respective nonlinear effects of a covariate – they point, however, in individual directions. Interaction effects are listed separately. We observe that for many covariates the nonlinear effect is already almost entirely captured by its first component.

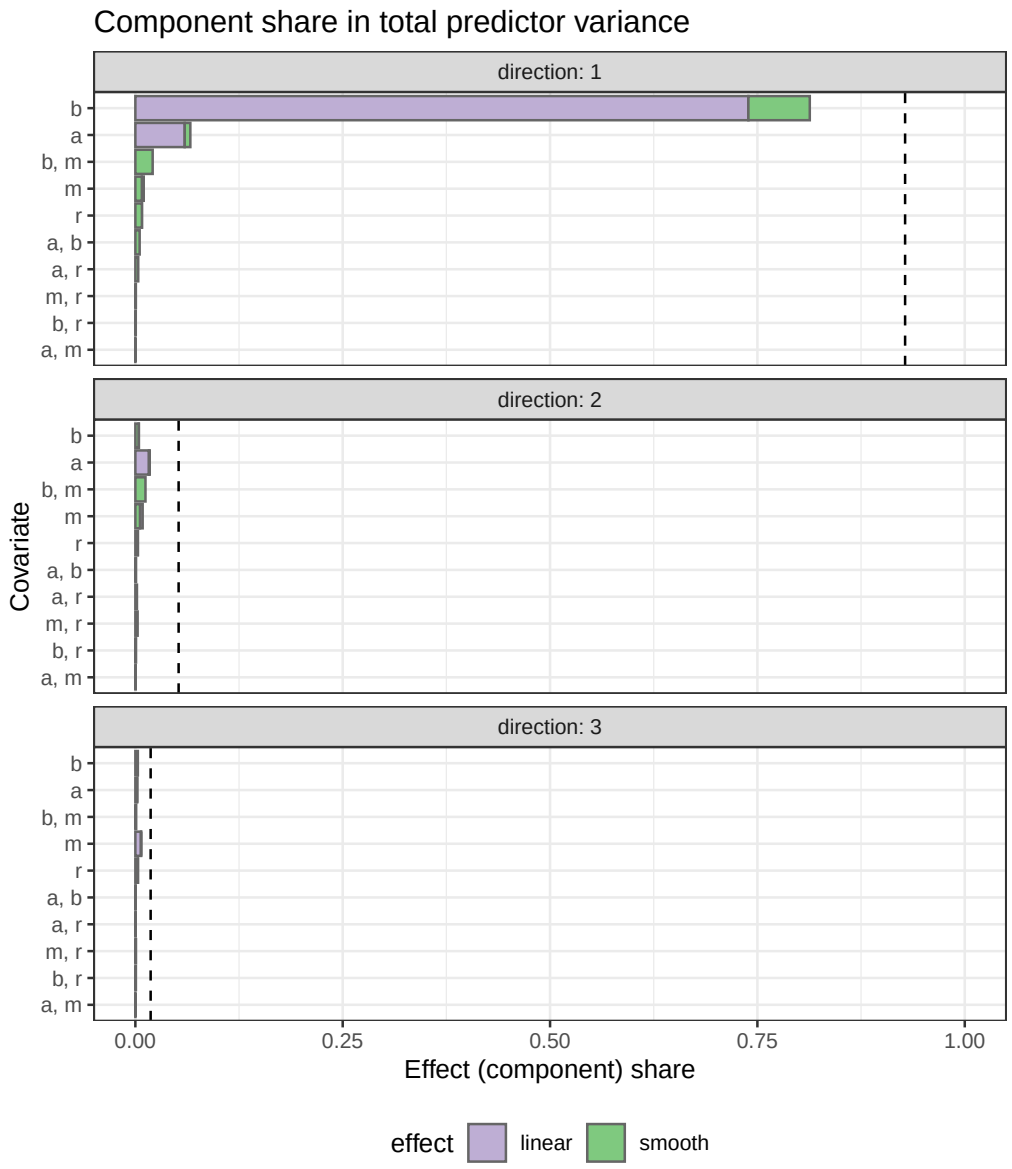

Figure S3: Tensor-product model factorization: Predictor variance shares into the first three directions *(dashed vertical lines)* resulting from joint model factorization (unlike individual factorization of effects in Figure S2). Horizontal bars reflect the variance of the single covariate effects within each model predictor component. They roughly – but due to potential correlation not precisely – add up to the predictor component variance shares.

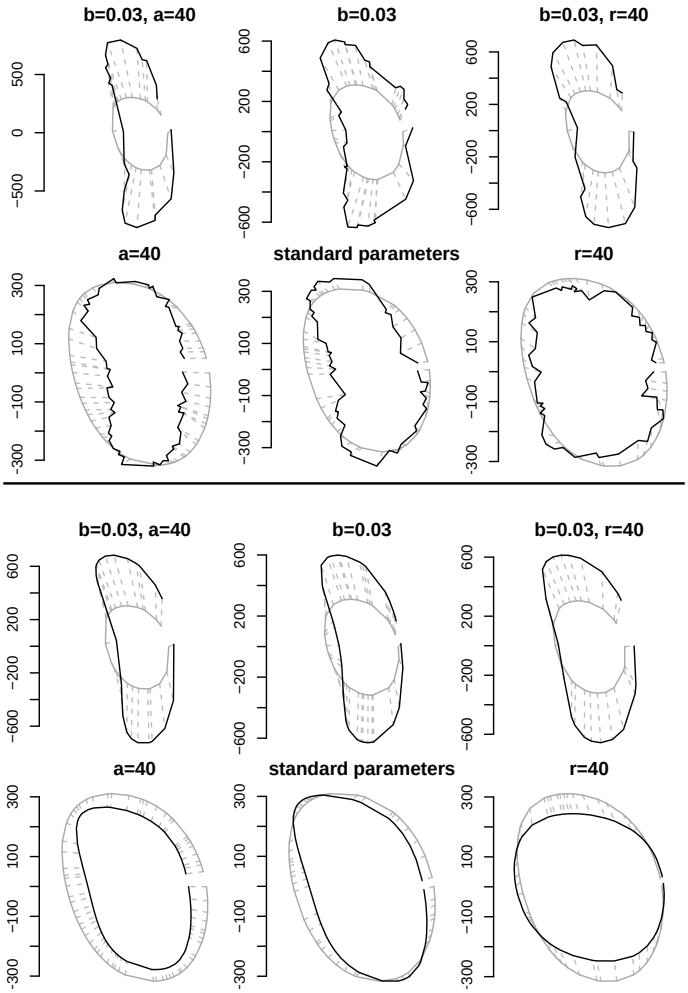

Figure S4: *Top:* Example cell outline *(black)*, one randomly selected out of 33 for each of six different CPM parameter (covariate) configurations chosen for visualization, aligned to the overall mean form *(grey)*. Note that while panel scales are individually adjusted for better visibility, contrasting plotted forms with the overall mean, which is equal in all plots, also allows to compare their sizes across panels. Headers show parameter deviations from a standard configuration with $b = 0.009$, $m = 0.003$, $a = 50$ and $r = 20$. Dashed lines indicate point correspondences. Cell outlines are oriented as cells migrating rightwards and not connected between $y(0)$ and the point left of it (while outlines are modeled as closed forms in the model). *Bottom:* Predictions for the corresponding mean form of our cell form model described in Section 5.2.

S.5 Realistic shape and form simulation studies

S.5.1 Sampling of response observations

Response curves are generated separately for the shape and form scenario as follows: we obtain the true underlying models by fitting original beer and whisky bottles and 3D rotated versions of them, four successively rotated towards the viewer and four away from the viewer, and compute transported residuals ϵ_i of a total of $N = 360$ bottle outlines $y_1, \ldots, y_N$ (20 whisky and 20 beer brands, each from 9 different angles z_1). For each simulated dataset, a sample of the desired size n is randomly drawn (with replacement) from the model residuals $\epsilon_1, \ldots, \epsilon_N$. To obtain irregular data with an average grid length $k = \frac{1}{n}\sum_{i=1}^{n} k_i$, we subsample the original evaluations $\epsilon_i(t_{i1}), \ldots, \epsilon_i(t_{iK_i})$, with original grid sizes $K_i \geq 123$, in two steps: first we randomly pick three evaluations as minimal sample size; then we draw evaluations independently with $\frac{k-3}{K_i-3}$ probability to enter the dataset. To preserve the original covariate distribution of the data, covariates are not randomly picked but we select batches of 9 beer and 9 whisky bottles with $z_1 \in [-60, 60]$ as in the original dataset. Sample sizes n are, therefore, multiples of 18. With the conditional means $[\mu_i]$ determined by the covariates, the evaluated residuals ϵ_i (on k_i points) are parallel transported to $\varepsilon_{[\mu_i],i} \in T_{[\mu_i]}\mathcal{Y}^*_{i/G}$, into the tangent space of the true conditional mean, to generate the simulated shape/form dataset $[y_i] = \mathrm{Exp}_{[\mu_i]}(\varepsilon_{[\mu_i],i})$, $i = 1, \ldots, n$.

S.5.2 Simulation results

In order to systematically and efficiently assess model behavior, we vary key aspects of the model setup and compare fitting performance in selected settings. Here, we list the different aspects and how they are referred to in subsequent graphical visualizations:

- Scenario: Shape or form responses.
- Sample size n of curves and mean grid size k that curves are evaluated on.
- Setting: simulations adjusted in an additional aspect compared to a default setup

  equal weight: Constant inner product weights $w_{i\iota} = \frac{1}{k_i}$, $\iota = 1, \ldots, k_i$, are utilized for curve evaluations $y_i(t_{i1}), \ldots, y_i(t_{ik_i})$ instead of trapezoidal rule weights (default).

  no nuisance: No constant and smooth nuisance effects h_0 and $f_2(z_2)$ are included into the model, which are included by default.

  `pre-aligned`: This setting concerns the pre-alignment of the curves $y_1, \ldots, y_n$ representing the forms/shapes in the simulated data. Note, however, that due to alignment to the pole p in the very beginning of the Riemannian L^2-Boosting algorithm, all of this only effects the preliminary pole p_0 used for estimation of p. In the models fit in the paper, we estimated p_0 by using a functional L^2-Boosting algorithm (without any alignment), which makes sense for typical data where the curves occur roughly aligned.

Consequently, this aspect translates to a "good or worse starting point p_0", which is then replaced by p in the actual model fit. In pre-aligned settings, simulated response curves $\tilde{y}_i = \mathrm{Exp}_{\mu_i}(\varepsilon_{\mu_i,i})$ are directly used for fitting. In the default, by contrast, the model is fit on random representatives of $[y_i]$ to mimic realistic scenarios, where $y_i = \lambda u \tilde{y}_i + \gamma \in [\tilde{y}_i]$ with $u = \exp(\sqrt{\text{-}1}\omega)$, $\omega \sim N(0, \frac{\pi}{20})$, with $\gamma = \sigma_1\gamma_1 + \sigma_2^2\gamma_2\sqrt{\text{-}1}$, $\gamma_1, \gamma_2 \sim N(0,1)$, $(\sigma_1^2, \sigma_2^2) = \frac{1}{nk}\sum_{i=1}^{n}\sum_{\iota=1}^{k_i}(\mathrm{Re}(\tilde{y}_i(t_\iota)), \mathrm{Im}(\tilde{y}_i(t_\iota)))$, and with $\lambda = 1$ for forms and $\lambda \sim \mathrm{Gamma}(10^2, 10^{-2})$ for shapes.

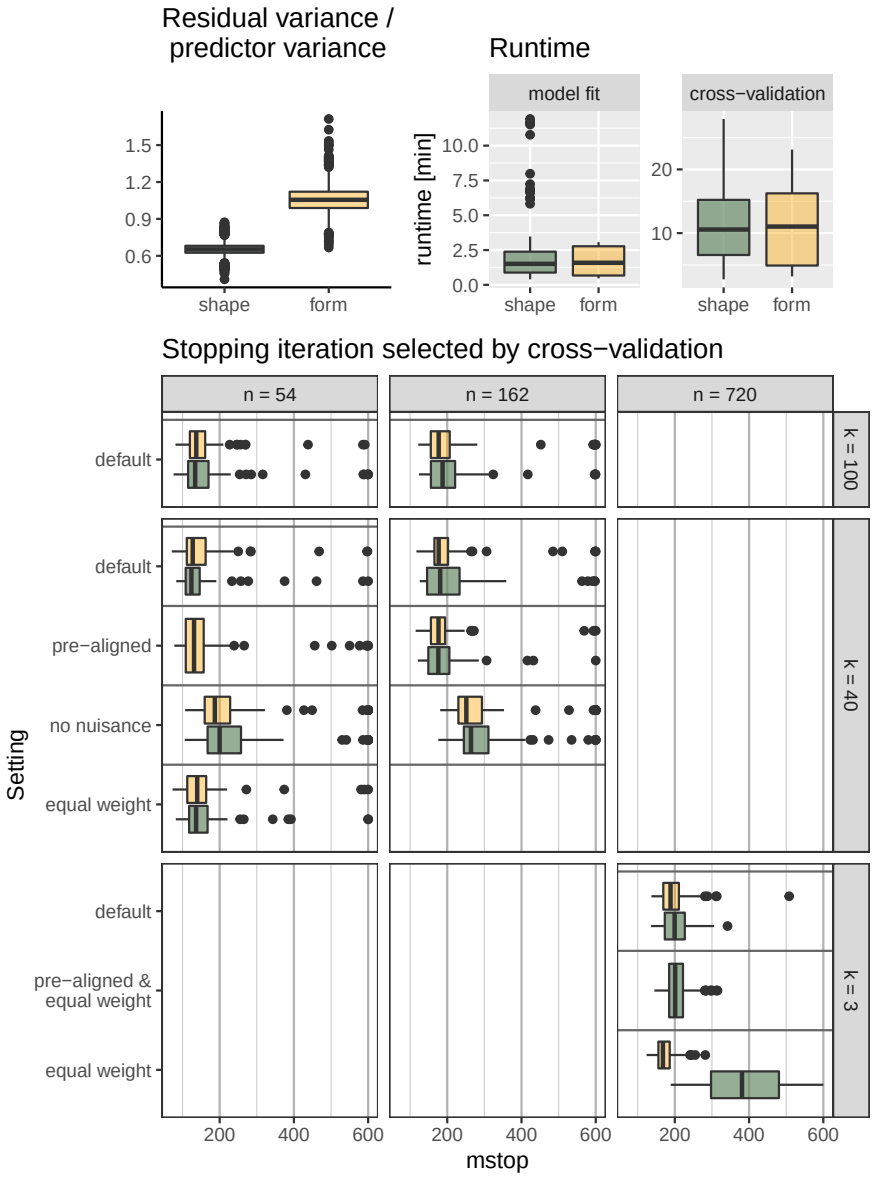

Figure S5: *Top, left:* Noise-to-signal ratio: distribution of empirical residual variance / predictor variance ratio in all simulations. *Top, right:* Runtime distribution of model fits and subsequent cross-validations (always running 600 boosting iterations). *Bottom:* Distribution of stopping iteration m_{stop} selected by 10-fold curve-wise cross-validation for different simulation settings. All plots displayed separately for the shape and form scenario (top and bottom row within sub-panel, respectively).

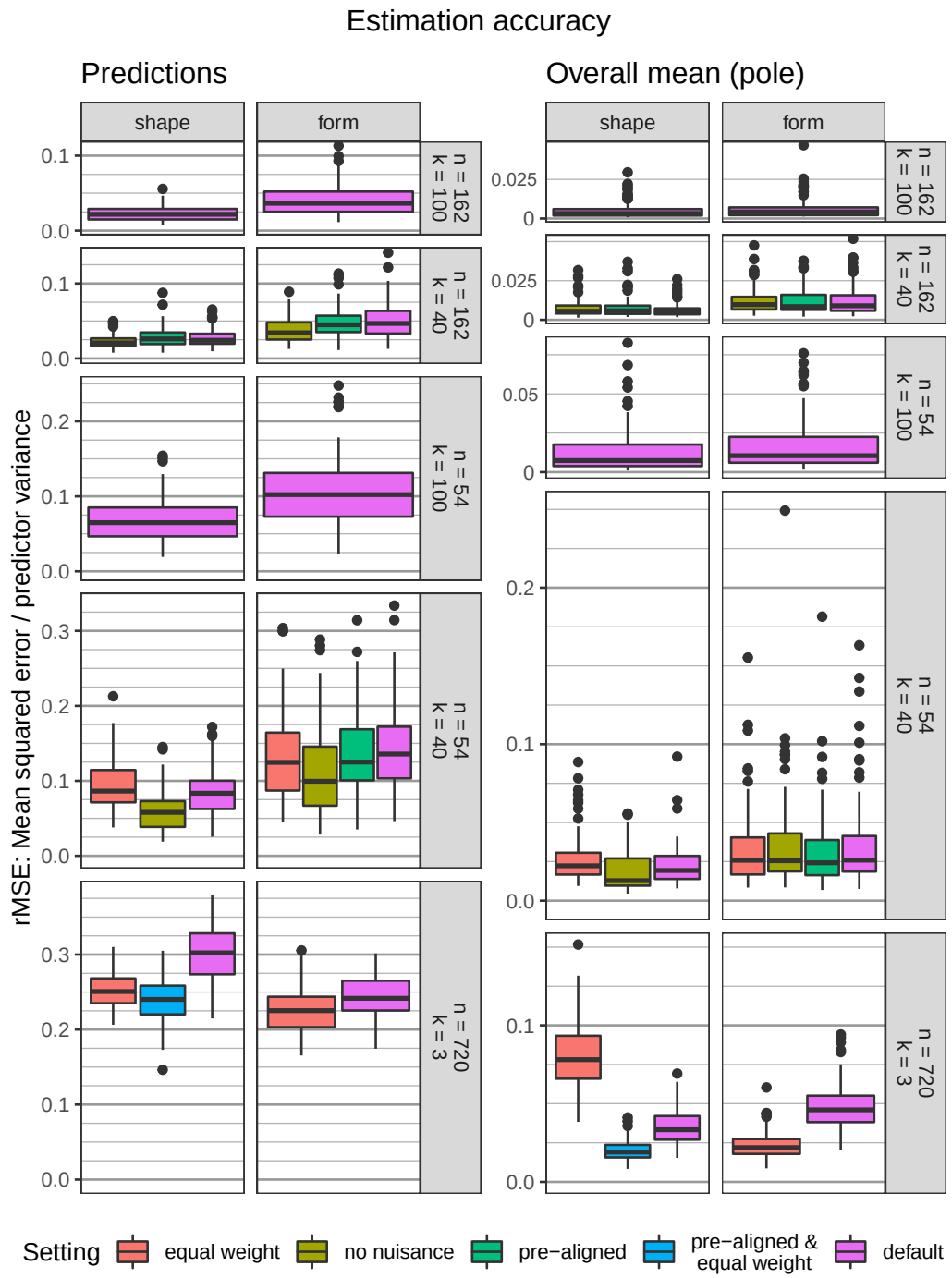

Figure S6: Accuracy in estimating the unconditional mean (pole) and conditional means (predictions), where the MSE is averaged over the covariate values in the dataset.

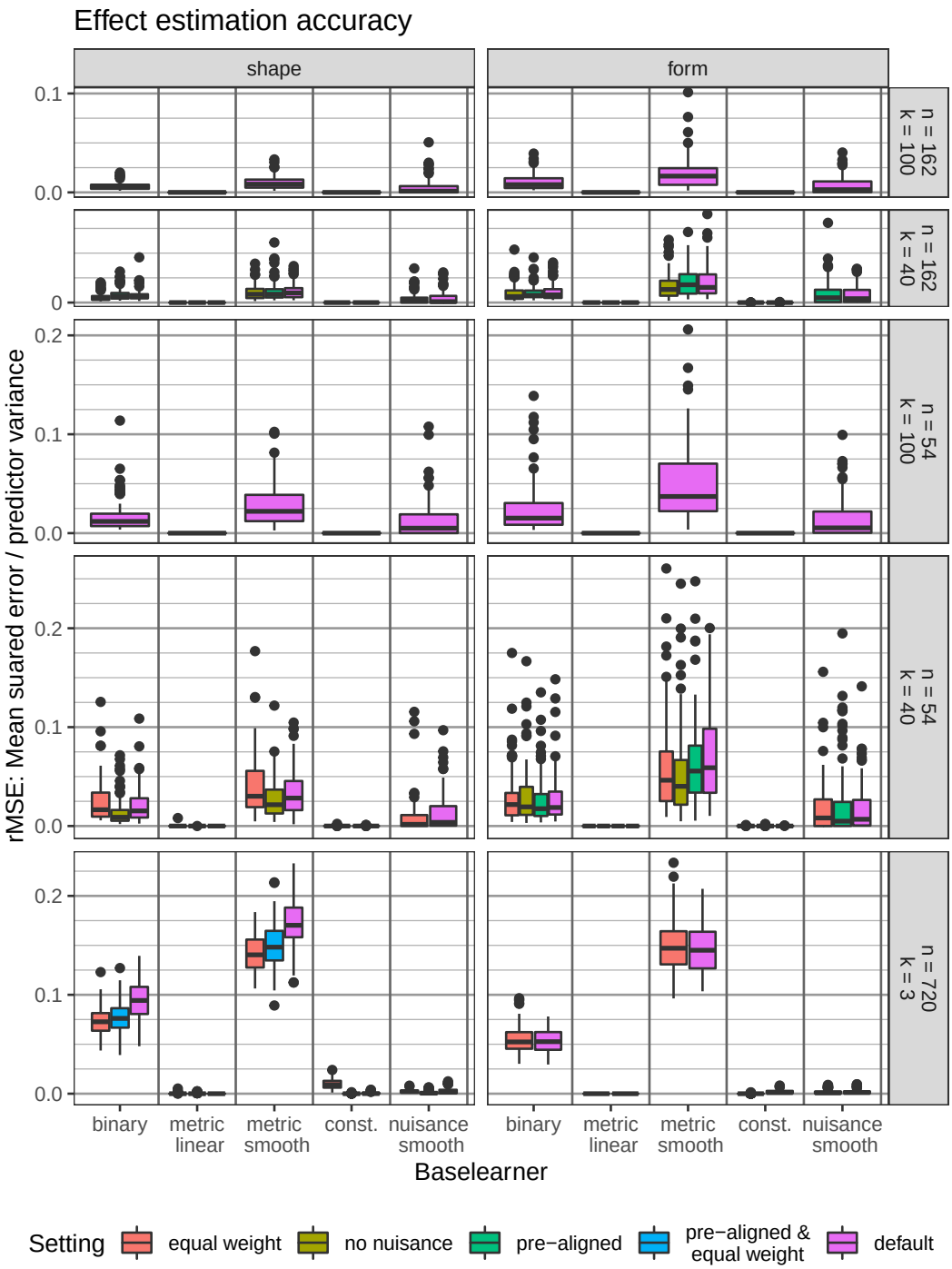

Figure S7: Accuracy of estimated effects on tangent space level.

S.6 Coefficient level modeling

In the main manuscript, we consider the space of complex valued functions $\mathcal{Y}$ mostly a vector space over $\mathbb{R}$ and utilize real coefficients to formulate the tensor-product effect structure in Section 3.1 in corresponding bases. In particular for form tangent spaces, identified with real linear subspaces that do not correspond to complex subspaces, this is useful to implement respective constraints via basis transforms. By contrast, here we represent $h_j(\mathbf{x}) = \sum_{r,l} \vartheta_j^{(r,l)} b_j^{(l)}(\mathbf{x}) b_0^{(r)}$ with complex coefficients $\vartheta_j^{(r,l)} \in \mathbb{C}$, $r = 1, \ldots, m_0$, $l = 1, \ldots, m_j$, with (possibly all real-valued) basis functions $b_0^{(1)}, \ldots, b_0^{(m_0)} \in \mathcal{Y}$ corresponding to the basis used for construction of the tangent space basis $\{\partial_r\}_r$ in Section 3.1. This representation lets us illustrate the link between evaluation level and coefficient level modeling of shapes and forms:
Consider the case where $y_i \in \mathcal{Y}$, $i = 1, \ldots, n$, can be expanded as $y_i = \sum_{r=1}^{m_0} \check{y}_i^{(r)} b_0^{(r)}$ in the same basis with complex coefficient vectors $\check{\mathbf{y}}_i = (\check{y}_i^{(1)}, \ldots, \check{y}_i^{(m_0)})^\top \in \mathbb{C}^{m_0}$, and let also the pole $[p] = [\sum_{r=1}^{m_0} \check{p}^{(r)} b_0^{(r)}]$, $\check{\mathbf{p}}_i = (\check{p}^{(1)}, \ldots, \check{p}^{(m_0)})^\top$, be expanded accordingly. With $\check{\imath} = (\check{\imath}^{(1)}, \ldots, \check{\imath}^{(m_0)})$ the coefficient vector of $\imath = \sum_{r=1}^{m_0} \check{\imath}^{(r)} b_0^{(r)}$ (for B-splines simply $\check{\imath} = \frac{1}{|\mathcal{T}|}(1, \ldots, 1)^\top$), we have $u y_i + \gamma \imath = \sum_{r=1}^{m_0} (u\, \check{y}_i^{(r)} + \gamma\, \check{\imath}) b_0^{(r)}$ for $u, \gamma \in \mathbb{C}$, such that basis representation yields an isomorphism between shapes/forms $[y]$ of curves and the shapes/forms $[\check{\mathbf{y}}]$ of their coefficients as alternative "landmarks". Moreover, when choosing inner products on $\mathcal{Y}$ and $\mathbb{C}^{m_0}$ such that $y \to \check{\mathbf{y}}$ is isometric, it follows that $[y] \to [\check{\mathbf{y}}]$ presents an isometric isomorphism.

Under these assumptions, modeling the mean shape/form $[\check{\boldsymbol{\mu}}] = \mathrm{Exp}_{[\check{\mathbf{p}}]}\left(\check{\mathbf{h}}(\mathbf{x})\right)$ of the coefficients $\check{\mathbf{y}}_i$, with predictor $\check{\mathbf{h}}(\mathbf{x}) = \sum_{j=1}^{J} \check{\mathbf{h}}_j(\mathbf{x}) \in \mathbb{C}^{m_0}$ and $\check{\boldsymbol{\mu}}_i = (\check{\mu}_i^{(1)}, \ldots, \check{\mu}_i^{(m_0)})^\top \in \mathbb{C}^{m_0}$, is equivalent to our presented model on the original level of curves, if coefficient level effects $\check{\mathbf{h}}_j(\mathbf{x}) = \sum_{r,l} \vartheta_j^{(r,l)} b_j^{(l)}(\mathbf{x}) \mathbf{e}_r$ are specified with the canonical basis $\mathbf{e}_r = (\mathbb{1}(r = 1), \ldots, \mathbb{1}(r = m_0))^\top$, since $[\mu] = [\sum_{r=1}^{m_0} \check{\mu}^{(r)} b_0^{(r)}] \overset{(*)}{=} \mathrm{Exp}_{[\sum_{r=1}^{m_0} \check{p}^{(r)} b_0^{(r)}]}\left(\sum_{r=1}^{m_0} \check{h}^{(r)}(\mathbf{x}) b_0^{(r)}\right) = \mathrm{Exp}_{[p]}(h(\mathbf{x}))$ with $\check{\mathbf{h}}(\mathbf{x}) = (\check{h}^{(1)}(\mathbf{x}), \ldots, \check{h}^{(m_0)}(\mathbf{x}))^\top$. For shapes, equality $(*)$ follows from

$$
\begin{aligned}
\mathrm{Exp}_{\sum_{r=1}^{m_0} \check{p}^{(r)} b_0^{(r)}}\left(\sum_{r=1}^{m_0} \check{h}^{(r)}(\mathbf{x}) b_0^{(r)}\right) &= \cos(\|\check{\mathbf{h}}(\mathbf{x})\|) \sum_{r=1}^{m_0} \check{p}^{(r)} b_0^{(r)} + \sin(\|\check{\mathbf{h}}(\mathbf{x})\|) \frac{\sum_{r=1}^{m_0} \check{h}^{(r)}(\mathbf{x}) b_0^{(r)}}{\|\check{\mathbf{h}}(\mathbf{x})\|} \\
&= \sum_{r=1}^{m_0} \left(\cos(\|\check{\mathbf{h}}(\mathbf{x})\|) \check{p}^{(r)} + \sin(\|\check{\mathbf{h}}(\mathbf{x})\|) \frac{\check{h}^{(r)}(\mathbf{x})}{\|\check{\mathbf{h}}(\mathbf{x})\|}\right) b_0^{(r)} \\
&= \sum_{r=1}^{m_0} \mathbf{e}_r^\top \mathrm{Exp}_{\check{\mathbf{p}}}\left(\check{\mathbf{h}}(\mathbf{x})\right) b_0^{(r)} = \sum_{r=1}^{m_0} \check{\mu}^{(r)} b_0^{(r)}
\end{aligned}
$$

where Exp is the exponential map on the sphere (first on the function space and then on the coefficient level), we use that due to the isometry $\|\check{\mathbf{h}}(\mathbf{x})\| = \|\sum_{r=1}^{m_0} \check{h}^{(r)}(\mathbf{x}) b_0^{(r)}\|$ and, we assume w.l.o.g. $\|p\| = \|\check{\mathbf{p}}\| = 1$ and $\langle p, \imath \rangle = \sum_r \check{p}^{(r)} = 0$. Accordingly for forms.

However, the expansion $y_i \approx \sum_{r=1}^{m_0} \check{y}_i^{(r)} b_0^{(r)}$ is typically only approximate. In terms of the inner product, $\langle y_i, y_i' \rangle_i^0 = \check{\mathbf{y}}_i^\dagger \check{\mathbf{W}} \check{\mathbf{y}}_i'$, with $\check{\mathbf{W}}$ the Gramian matrix of $\{b_0^{(r)}\}_r$, presents an alternative empirical substitute for the inner product $\langle y_i, y_i' \rangle$ of curves $y_i, y_i' \in \mathcal{Y}$, which is

computed on the coefficients instead of $\langle y_i, y_i' \rangle_i = \mathbf{y}_i^\dagger \mathbf{W}_i \mathbf{y}_i'$ computed on evaluation vectors $\mathbf{y}_i = (y_i(t_{i1}), \ldots, y_i(t_{ik_i}))^\top, \mathbf{y}_i' = (y_i'(t_{i1}), \ldots, y_i'(t_{ik_i}))^\top$ as suggested in Section 2. When, for dense grids, it can be assumed that both $\langle y_i, y_i' \rangle_i^0 \approx \langle y_i, y_i' \rangle_i \approx \langle y_i, y_i' \rangle$ approximate the inner product on the level of curves well, the approach based on the coefficients $\check{\mathbf{y}}_i$ may be computationally preferable, guaranteeing regular and typically more sparse representations that necessitate operations on smaller design matrices (in particular when utilizing the linear array framework (Brockhaus et al., 2015)). By contrast, in comparably sparse irregular scenarios, expanding single observed y_i in a basis in a first step might involve unwanted pre-smoothing. To give a consistent presentation on the original level of curves, we rely on an evaluation based approach in all applications presented in the main manuscript.

S.7 Functional Principal Component Representation

Various approaches in the literature (e.g., Müller and Yao, 2008; Scheipl et al., 2015; Cederbaum et al., 2016; Volkmann et al., 2021) have employed functional principal component (FPC) basis representations for modeling functional responses in regression models. In combination with covariance smoothing (e.g., Yao et al., 2005; Cederbaum et al., 2018) this can be particularly useful in sparse/irregular scenarios, allowing to estimate the functional covariance structure from single curve evaluations. In fact, two variants of corresponding approaches directly fit into our proposed framework, either a) representing curves using predicted FPC scores or b) estimating inner products based on the covariance structure. In the following, we outline both approaches and briefly discuss related perspectives beyond the scope of this paper.

Prediction of FPC scores and inner products are carried out along the lines of Yao et al. (2005) and, in the complex case, Stöcker et al. (2022). Assume we have given (an estimate of) the complex covariance surface $C(s,t) = \mathbb{E}\left(Y^\dagger(s)Y(t)\right)$ of the process Y generating the curve samples $y_1, \ldots, y_n$ in the data, with point-wise mean $\mathbb{E}(Y(t)) = 0$ for all $t \in \mathcal{T}$ without loss of generality in the following. Under standard assumptions, this yields a (truncated) FPC basis $b_0^{(r)} : \mathcal{T} \to \mathbb{C}$, $r = 1, \ldots, m_0$ with respective eigenvalues $\lambda_1 \geq \cdots \geq \lambda_{m_0} \geq 0$. Observing only evaluation vectors $\mathbf{y}_i = (y_{i1}, \ldots, y_{ik_i})^\top = (y_i(t_{i1}) + \epsilon_{i1}, \ldots, y_i(t_{ik_i}) + \epsilon_{ik_i})^\top$ at time-points $t_{i1}, \ldots, t_{ik_i} \in \mathcal{T}$ subject to some iid. white noise measurement errors $\epsilon_{i1}, \ldots, \epsilon_{ik_i} \sim N(0, \sigma^2)$, predicted FPC score vectors $\check{\mathbf{y}}_i$, comprising predicted basis coefficients of y_i expanded in the basis $\{b_0^{(r)}\}_r$, can be obtained via the conditional expectations

$$\check{\mathbf{y}}_i = \mathbb{E}\left((\langle b_0^{(1)}, Y \rangle, \ldots, \langle b_0^{(m_0)}, Y \rangle)^\top \mid \mathbf{Y}_i + \epsilon_i = \mathbf{y}_i \right) = \mathbf{\Lambda} \mathbf{B}_i^\dagger \mathbf{\Sigma}_i^{-1} \mathbf{y}_i \tag{9}$$

under a working normality assumption, with matrices $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_{m_0})$, $\mathbf{B}_i$ with columns $(b_0^{(r)}(t_{i1}), \ldots, b_0^{(r)}(t_{ik_i}))^\top$, $r = 1, \ldots, m_0$, and $\mathbf{\Sigma}_i$ the covariance matrix of $\mathbf{Y}_i + \epsilon_i = (Y(t_{i1}) + \epsilon_{i1}, \ldots, Y(t_{ik_i}) + \epsilon_{i1})^\top$ obtained from corresponding evaluations of C plus σ^2 on the diagonal.

Approach a) directly analyses shapes of the predicted score vectors $\check{\mathbf{y}}_1, \ldots, \check{\mathbf{y}}_n$ as described in Section S.6 of the supplementary material.

Approach b) uses (9) to motivate integration weights $\mathbf{W}_i = \mathbf{\Sigma}_i^{-1} \mathbf{B}_i \mathbf{\Lambda} \mathbf{B}_i^\dagger \mathbf{\Sigma}_i^{-1}$ for the empirical inner products $\langle y_i, y_i' \rangle_i = \mathbf{y}_i^\dagger \mathbf{W}_i \mathbf{y}_i'$, $i = 1, \ldots, n$, introduced in Section 2 of the

main manuscript for $\mathbf{y}_i, \mathbf{y}_i' \in \mathbb{C}^{k_i}$, such that with this choice

$$\langle y_i, y_i' \rangle_i = \mathbf{y}_i^\dagger \mathbf{W}_i \mathbf{y}_i' = \mathbb{E}\left(\langle Y, Y' \rangle \mid \mathbf{Y}_i + \epsilon_i = \mathbf{y}_i, \mathbf{Y}_i' + \epsilon_i' = \mathbf{y}_i'\right)$$

for an independent copy Y' of Y, with $\mathbf{Y}_i'$ and ϵ_i' defined as $\mathbf{Y}_i$ and ϵ_i. Approach b) might be refined by approximating $\|y_i\|^2$ with $\mathbb{E}\left(\langle Y, Y \rangle \mid \mathbf{Y}_i + \epsilon_i = \mathbf{y}_i\right)$ and $\langle b^*, y_i \rangle$ with $\mathbb{E}\left(\langle b^*, Y \rangle \mid \mathbf{Y}_i + \epsilon_i = \mathbf{y}_i\right)$ for a known function $b^* : \mathcal{T} \to \mathbf{C}$ as described by Stöcker et al. (2022), which is slightly different from $\langle y_i, y_i \rangle_i$ and $\langle b^*, y_i \rangle_i$, respectively. However, basing all computations on $\langle y_i, y_i' \rangle_i$ as described in the main manuscript, holds the advantage of a unified definition of the shape geometry on evaluation vectors and curves.

Both a) and b) rely, however, on the covariance $C(s,t)$ of the process Y underlying the realizations y_i, while we ultimately analyze shapes/forms $[y_i]$, $i = 1, \ldots, n$, presenting equivalence classes. In practice, this might in many cases not be a problem, when the y_i are in fact roughly aligned and not as arbitrarily recorded as they might be in theory. However in general, it renders FPC based approaches for such settings more complicated and beyond the scope of this work (compare Stöcker et al., 2022, for related work in a different non-regression setting). Carrying out the FPC alternatively on tangent space level (i.e. in a linear space) would require computation of $\text{Log}_{[p]}([y_i])$ at some shape/form $[p]$ involving already computation/prediction of inner products.

We leave such considerations to future research, and focus instead on simpler weight matrices $\mathbf{W}_i$ which are known to also work reasonably well in regression scenarios with sparsely/irregularly sampled functional response (Scheipl et al., 2015, 2016; Brockhaus et al., 2015, 2017; Rügamer et al., 2018; Stöcker et al., 2021).

S.8 Tensor-product structure in non-parametric regression

We illustrate the broad applicability of the proposed TP factorization (Section 3.2) for the example of *Additive Regression with Hilbertian Responses* proposed by Jeon and Park (2020) showing that also approaches avoiding (finite-dimensional) basis representations may lead to the desired form of effect estimates $\hat{h}_j(\mathbf{x})$. Hence, although they do not consider manifold valued responses, TP factorization can be directly applied to visualize and investigate their effect estimates. We adapt relevant equations to fit our notation and refer for details to their work.

Jeon and Park (2020) consider regression with an additive predictor $h(\mathbf{x}) = \sum_{j=1}^J h_j(x_j)$ with $h_j(x_j)$ depending on the jth scalar covariate in $\mathbf{x} = (x_1, \ldots, x_J)^\top$. In Section 2.5 p. 2679, they point out that the estimator $\hat{h}_j(x_j)$ of $h_j(x_j)$ is a linear smoother if the initial estimate of their back-fitting algorithm is (as, e.g., in all their numerical studies). Assuming this in the following, the expression becomes

$$\hat{h}_j(x_j) = \frac{1}{n} \sum_{i=1}^n w_{ij}^{[g]}(x_j)\, y_i$$

with weight functions $w_{ij}^{[g]}(x_j)$, $i = 1, \ldots, n$, $j = 1, \ldots, J$ after g fitting iterations. In fact, this immediately has the desired TP form given in Section 3.1, setting $m = m_j = n$,

$\theta_j^{(r,l)} = \frac{1}{n}\mathbb{1}(r = l)$, $b_j^{(l)} = w_{lj}^{[g]}$ and $\partial_r = y_r$ for all $l, r = 1, \ldots, n$ and j. Here, tangent vectors are naturally identified with elements of the Hilbert space, as we are in the linear case.

It might seem odd to have the effect basis functions $b_j^{(i)}$ only implicitly defined depending on the fitting iteration. Yet in fact, the $w_{ij}^{[g]}$ are all in the span of

$$b_j^{(i)}(x_j) = \frac{\mathcal{K}_j(x_j, x_{ij})}{\sum_{i=1}^n \mathcal{K}_j(x_j, x_{ij})}, \quad i = 1, \ldots, n$$

with some kernels $\mathcal{K}_j$ evaluated around covariate realizations $\mathbf{x}_i = (x_{i1}, \ldots, x_{iJ})^\top$, $i = 1, \ldots, n$. This can be seen by re-writing the definition of $w_{ij}^{[g]}$ (Jeon and Park, 2020, Sec. 2.5, p. 2679):

$$\begin{aligned}
w_{ij}^{[g]}(x_j) &= \frac{\mathcal{K}_j(x_j, x_{ij})}{\hat{P}_j(x_j)} - 1 - \sum_{\jmath \neq j} \int_0^1 w_{ij}^{[g-\mathbb{1}(\jmath \geq j)]}(x_\jmath) \frac{\hat{P}_{j\jmath}(x_j, x_\jmath)}{\hat{P}_j(x_j)}\, dx_\jmath \\
&= \frac{\mathcal{K}_j(x_j, x_{ij})}{\hat{P}_j(x_j)} - 1 - \sum_{l=1}^n \sum_{\jmath \neq j} \int_0^1 w_{ij}^{[g-\mathbb{1}(\jmath \geq j)]}(x_\jmath) \frac{\mathcal{K}_j(x_j, x_{lj}) \mathcal{K}_\jmath(x_\jmath, x_{l\jmath})}{\hat{P}_j(x_j)}\, dx_\jmath \\
&= \underbrace{\frac{\mathcal{K}_j(x_j, x_{ij})}{\hat{P}_j(x_j)}}_{= n\, b_j^{(i)}(x_j)} - 1 - \sum_{l=1}^n \frac{\mathcal{K}_j(x_j, x_{lj})}{\hat{P}_j(x_j)} \underbrace{\sum_{\jmath \neq j} \int_0^1 w_{ij}^{[g-\mathbb{1}(\jmath \geq j)]}(x_\jmath) \mathcal{K}_\jmath(x_\jmath, x_{l\jmath})\, dx_\jmath}_{=: a_{lj}^{[g]} \in \mathbb{R} \text{ or } \mathbb{C}, \text{ respectively}}, \\
&= n \sum_{l=1}^n (\mathbb{1}(l = i) - \frac{1}{n} - a_{lj}^{[g]})\, b_j^{(l)}(x_j),
\end{aligned}$$

where by definition

$$\hat{P}_j(x_j) = \frac{1}{n} \sum_{i=1}^n \mathcal{K}_j(x_j, x_{ij}), \quad \hat{P}_{j\jmath}(x_j, x_\jmath) = \frac{1}{n} \sum_{i=1}^n \mathcal{K}_j(x_j, x_{ij}) \mathcal{K}_\jmath(x_\jmath, x_{i\jmath}).$$

and by construction $1 \equiv \sum_{i=1}^n b_j^{(i)}(x_j)$. (Starting values for the back-fitting algorithm presented in the paper are given simply by $w_{ij}^{[0]} = 0$ or the Nadaraya-Watson-type estimator $w_{ij}^{[0]} = \frac{1}{n} \sum_{i=1}^n (\frac{\mathcal{K}_j(x_j, x_{ij})}{\hat{P}_j(x_j)} - 1)\, y_i$.)
Consequently, also this non-parametric approach leads to the TP effect structure

$$\hat{h}_j(x_j) = \sum_{r=1}^n \sum_{l=1}^n \hat{\theta}_j^{(r,l)} \underbrace{\frac{\mathcal{K}_j(x_j, x_{lj})}{\sum_{i=1}^n \mathcal{K}(x_j, x_{ij})}}_{= b_j^{(l)}(x_j)} \underbrace{y_r}_{\partial_r}$$

with $\hat{\theta}_j^{(r,l)} = \mathbb{1}(l = r) - \frac{1}{n} - a_{lj}^{[g]}$.

References Supplement

Brockhaus, S., M. Melcher, F. Leisch, and S. Greven (2017). Boosting flexible functional regression models with a high number of functional historical effects. *Statistics and Computing 27*(4), 913–926.

Brockhaus, S., F. Scheipl, and S. Greven (2015). The Functional Linear Array Model. *Statistical Modelling 15*(3), 279–300.

Cederbaum, J., M. Pouplier, P. Hoole, and S. Greven (2016). Functional linear mixed models for irregularly or sparsely sampled data. *Statistical Modelling 16*(1), 67–88.

Cederbaum, J., F. Scheipl, and S. Greven (2018). Fast symmetric additive covariance smoothing. *Computational Statistics & Data Analysis 120*, 25–41.

Dryden, I. L. and K. V. Mardia (2016). *Statistical Shape Analysis: With Applications in R*. John Wiley & Sons.

Gentle, J. E. (2007). *Matrix algebra: Theory, Computations and Applications in Statistics*. Springer International Publishing.

Hsing, T. and R. Eubank (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*. John Wiley & Sons.

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.

Klingenberg, W. (1995). *Riemannian geometry*. de Gruyter.

Lee, J. M. (2018). *Introduction to Riemannian manifolds*. Springer.

Mardia, K. V. and P. E. Jupp (2009). *Directional statistics*, Volume 494. John Wiley & Sons.

Müller, H.-G. and F. Yao (2008). Functional additive models. *Journal of the American Statistical Association 103*(484), 1534–1544.

Rügamer, D., S. Brockhaus, K. Gentsch, K. Scherer, and S. Greven (2018). Boosting factor-specific functional historical models for the detection of synchronization in bioelectrical signals. *Journal of the Royal Statistical Society: Series C 67*(3), 621–642.

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.

Stöcker, A., S. Brockhaus, S. A. Schaffer, B. v. Bronk, M. Opitz, and S. Greven (2021). Boosting functional response models for location, scale and shape with an application to bacterial competition. *Statistical Modelling 21*(5), 385–404.

Stöcker, A., M. Pfeuffer, L. Steyer, and S. Greven (2022). Elastic full Procrustes analysis of plane curves via Hermitian covariance smoothing.

Tu, L. W. (2011). *An Introduction to Manifolds*. Springer.

Volkmann, A., A. Stöcker, F. Scheipl, and S. Greven (2021). Multivariate functional additive mixed models. *Statistical Modelling*.

Yao, F., H. Müller, and J. Wang (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association 100*(470), 577–590.

C. Supplementary material for Chapter 7 “Elastic Full Procrustes Analysis of Plane Curves via Hermitian Covariance Smoothing”

Online supplement for the contribution:

Stöcker, A., Pfeuffer, M., Steyer, L., and Greven, S. (2022). Elastic Full Procrustes Analysis of Planar Curves via Hermitian Covariance Smoothing. *arXiv pre-print*.
DOI: 10.48550/ARXIV.2203.10522.

Supplementary material for "Elastic Full Procrustes Analysis of Plane Curves via Hermitian Covariance Smoothing"

BY A. STÖCKER, M. PFEUFFER, L. STEYER AND S. GREVEN

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

almond.stoecker@hu-berlin.de manuel.pfeuffer@esmt.org lisa.steyer@hu-berlin.de sonja.greven@hu-berlin.de

SUMMARY

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 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 framework. Identifying the real plane with the complex numbers, we establish a connection to covariance estimation in irregular/sparse functional data analysis and propose Hermitian covariance smoothing for (in)elastic full Procrustes mean estimation. 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.

1. HERMITIAN COVARIANCE SMOOTHING

1.1. Complex processes and rotation invariance

In the following, we detail prerequisites on linear operators and proof Theorem 1 and 2. Subsequently, Proposition S2 substantiates the relation of complex and real covariance surfaces indicated in the main manuscript.

We widely follow Hsing & 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 S1. *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 & Youngson (2007), Chapter 7.3. □

DEFINITION S1. *Let Y be a random element in $\mathbb{H}$ with $\mathbb{E}\left(\|Y\|^2\right) < \infty$. Then*

i) the mean element $\mu \in \mathbb{H}$ of Y is defined by $\langle f, \mu \rangle = \mathbb{E}\left(\langle f, Y \rangle\right)$ for all $f \in \mathbb{H}$.

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

PROPOSITION S1. *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 $\mathbb{E}\left(\|Y\|^2\right) < \infty$, Jensen's inequality yields $\mathbb{E}\left(\|Y\|\right) < \infty$, and therefore $\mathbb{E}\left(\langle f, Y\rangle\right) < \infty$ and also $\mathbb{E}\left(\langle Y-\mu, f\rangle\langle e, Y-\mu\rangle\right) < \infty$ 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 $\langle \Sigma(e), f\rangle = \mathbb{E}\left(\langle Y, f\rangle\langle e, Y\rangle\right) = \langle e, \Sigma(f)\rangle$ and nonnegative-definiteness $\langle \Sigma(e), e\rangle = \mathbb{E}\left(\langle Y, e\rangle\langle e, Y\rangle\right) = \mathbb{E}\left(|\langle e, Y\rangle|^2\right)$ immediately follow from the definition. Σ is trace-class, since for an orthonormal basis $\{e_k\}_k$ of $\mathbb{H}$ it holds that

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

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

COROLLARY S1. *The covariance operator Σ of Y with $\mathbb{E}(\|Y\|^2) < \infty$ has an eigenbasis as descibed in Theorem S1.*

Proof. Immediately follows from Theorem S1 and the self-adjointness and compactness of Σ shown in Proposition S1. □

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 $\mathbf{Y}$:

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

Proof 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. □

PROPOSITION S2. *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*

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

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

Proof.

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

which shows the desired form. □

2. Elastic full Procrustes analysis

2.1. 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.

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

$$\begin{aligned}
d^2_{\mathcal{G}}((\beta_1),(\beta_2)) &= \inf_{u\in\mathbb{C}} \|q_1 - u\,q_2\|^2 = \inf_{u\in\mathbb{C}} \big[1 - \overbrace{u}^{=r_1\exp(\mathbf{i}\,\omega_1)} \underbrace{\langle q_1, q_2\rangle}_{=r_2\exp(\mathbf{i}\,\omega_2)} - u^\dagger\langle q_2, q_1\rangle + |u|^2\big] \\
&= \inf_{r_1>0,\,\omega_1\in\mathbb{R}} \big[1 - r_1 r_2\exp(\mathbf{i}\,(\omega_1+\omega_2)) - r_1 r_2\exp(-\mathbf{i}\,(\omega_1+\omega_2)) + r_1^2\big] \\
&= \inf_{r_1>0,\,\omega_1\in\mathbb{R}} \big[1 - 2r_1 r_2\cos(\omega_1+\omega_2) + r_1^2\big] \overset{\omega_1=-\omega_2}{=} \inf_{r_1>0} \big[1 - 2r_1 r_2 + r_1^2\big] \qquad \text{(S1)} \\
&= \inf_{r_1>0} \big[1 - r_2^2 + (r_1 - r_2)^2\big] \overset{r_1=r_2}{=} 1 - |\langle q_1, q_2\rangle|^2 = \|q_1 - \langle q_2, q_1\rangle q_2\|^2 \qquad \text{(S2)}
\end{aligned}$$

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 $d_{\mathcal{G}}((\beta_1),(\beta_3)) = \inf_{u\in\mathbb{C}} \|q_1 - u\,q_3\| \overset{\mathbb{L}^2}{\underset{\text{tr. ineq.}}{\leq}}$ $\underbrace{\|q_1 - v^*\,q_2\|}_{\overset{(S2)}{=}\inf_{v\in\mathbb{C}}\|q_1 - v\,q_2\|} + \underbrace{\inf_{u\in\mathbb{C}} \|v^*\,q_2 - u\,q_3\|}_{=|v^*|\inf_{u\in\mathbb{C}}\|q_2-u\,q_3\|} \overset{|v^*|\leq 1}{\leq} d_{\mathcal{G}}((\beta_1),(\beta_2)) + d_{\mathcal{G}}((\beta_2),(\beta_3))$. This shows i). ii) directly follows from (S1), since $\exp(-\mathbf{i}\omega_2) = \langle q_1, q_2\rangle/|\langle q_1, q_2\rangle|$. □

Proof Proposition 3 iii). $\min_{(\beta)\in\tilde{\mathfrak{B}}} \mathbb{E}\left(d^2_{\mathcal{G}}((\beta),(B))\right) = \min_{y:\|y\|=1} \mathbb{E}\left(1 - |\langle y, Q\rangle|^2\right) =$ $1 - \max_{y:\|y\|=1} \mathbb{E}\left(|\langle y, Q\rangle|^2\right)$. Hence, $\psi_{\mathcal{G}} \in \operatorname{argmax}_{y:\|y\|=1} \mathbb{E}\left(|\langle y, Q\rangle|^2\right)$, and $\mathbb{E}\left(|\langle y, Q\rangle|^2\right) = \langle y, \Sigma(y)\rangle = \langle y, \sum_k \lambda_k\langle e_k, y\rangle e_k\rangle = \sum_k \lambda_k|\langle e_k, y\rangle|^2 \leq \lambda_1 \sum_k |\langle e_k, y\rangle|^2 = \lambda_1\|y\|^2 = \lambda_1$, 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 $(\mu_{\mathcal{G}})$ and $\sigma^2_{\mathcal{G}}$. □

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

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

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 $\tilde{\mathfrak{B}}$ by isometries, $\inf_{u\in\mathbb{C},\gamma\in\Gamma}\|q_1 - u\,q_2\circ\gamma\,\dot{\gamma}^{1/2}\| = \inf_{\gamma\in\Gamma} d_{\mathcal{G}}((\beta_1),(\beta_2))$ 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 $0 = d_{\mathcal{E}}([\beta_1],[\beta_2]) = \inf_{\gamma\in\Gamma}\lim_{l\to\infty}\|q_1 - u_l q_2\circ\gamma\dot{\gamma}^{1/2}\|$. 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 $\lim_{h\to\infty} u_{l_h} = u$, and $0 = \inf_{\gamma\in\Gamma}\lim_{h\to\infty}\|q_1 - u_{l_h} q_2\circ\gamma\dot{\gamma}^{1/2}\| \overset{\text{continuity}}{=} \inf_{\gamma\in\Gamma}\|q_1 - u\,q_2\circ\gamma\dot{\gamma}^{1/2}\|$ 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. □

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

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

THEOREM S2. *Let $\beta : [0,1]\to\mathbb{C}$ 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 $\phi : [0,\infty)\to\mathbb{C}$, $r\mapsto\beta(t_1) + r\,\mathbf{i}$ 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, $\zeta(t_1) > \Im(a_k)$ for $k = 2,\ldots,4$ 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. □

COROLLARY S2 (FEASIBLE SAMPLING). *If* $\beta^* : [0,1] \to \mathbb{C}$ *is continuous and* $\beta^* : (t^*_{j-1}, t^*_j) \to \mathbb{C}$ *continuously differentiable for* $j = 1, \dots, n_0$, $t^*_0 < \cdots < t^*_{n_0}$ *with non-vanishing derivative, then for any time points* $0 < t_1 < \cdots < t_{n_0} < 1$ *and speeds* $w_1, \dots, w_{n_0} > 0$*, there exists a* $\gamma \in \Gamma$ *such that for the SRV-transform* q *of* $\beta = \beta^* \circ \gamma$*,* $q(t_j) = w_j^{1/2} \left(\beta^*(t^*_j) - \beta^*(t^*_{j-1})\right) = w_j^{1/2}\, \Delta_j$ *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 S2, there exists $c \in (0,1)$ with $\dot\beta(c)^* = a\, \Delta_1$ for some $a > 0$. Choose $\gamma \in \Gamma$ such that $\gamma(t_1) = c$ and $\dot\gamma(t_1) = w_1 a^{-2}$. Then, $q(t_j) = \beta^* \circ \gamma(t_j) \dot\gamma(t_j)^{1/2} = a\, \Delta_1 w_1^{1/2} a^{-1} = w_1^{1/2} \Delta_1$ for all $j = 1, \dots, n_0$. □

2.3. 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 S3 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 $\hat C(s,t) = \mathbf{f}^\top(s)\, \widehat{\boldsymbol\Xi}\, \mathbf{f}(t)$ of the same form by setting $\widehat{\boldsymbol\Xi} = \frac{1}{n} \sum_{i=1}^n \hat{\boldsymbol\vartheta}_i \hat{\boldsymbol\vartheta}_i^\dagger$ 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 $\mathcal{N}_1 \cup \cdots \cup \mathcal{N}_N = \{1, \dots, n\}$ 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

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

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 "`gaulss`" 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 S3 iii) Equation (S3)), its optimal rotation to the mean (S4), its length (S5), and its distance from the mean (S6) 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 S3 (CONDITIONAL GAUSSIAN PROCESS). *Consider a random element Y in a complex Hilbert space $\mathbb{H}$ of functions $\mathcal{T} \to \mathbb{C}$ 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 $\tau^2 : \mathcal{T} \to \mathbb{R}$. 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 $\mathbf{T}_+ = \operatorname{Diag}(\tau^2(t_1), \dots, \tau^2(t_{n_+}))$ for the diagonal $n_+ \times n_+$ noise covariance matrix. Let $r_0 = \operatorname{rank}(\mathbf{E}_0)$ 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 $\mathbf{E}_0^\top = \mathbf{QR}$. 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* $\operatorname{span}(\mathbf{N})$ *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* $\mathbf{y}_0 \notin \operatorname{span}(\mathbf{E}_0)$*, 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*

$$\hat{\mathbf{z}} = \mathbb{E}\left(\mathbf{Z} \mid Y(t_j) + \varepsilon(t_j) = y_j, j = 1, \dots, n_*\right) = \mathbf{N}\hat{\mathbf{z}}_+ + \mathbf{z}_0 \tag{S3}$$

$$\mathbb{E}\left(\langle Y, x \rangle \mid Y(t_j) + \varepsilon(t_j) = y_j, j = 1, \dots, n_*\right) = \hat{\mathbf{z}}^\dagger \mathbf{g}_x \tag{S4}$$

$$\mathbb{E}\left(\|Y\|^2 \mid Y(t_j) + \varepsilon(t_j) = y_j, j = 1, \dots, n_*\right) = \operatorname{tr}(\mathbf{S}\,\mathbf{G}) + \hat{\mathbf{z}}^\dagger \mathbf{G} \hat{\mathbf{z}}. \tag{S5}$$

$$\mathbb{E}\left(|\langle Y, x \rangle|^2 \mid Y(t_j) + \varepsilon(t_j) = y_j, j = 1, \dots, n_*\right) = \mathbf{g}_x^\dagger \mathbf{S}\,\mathbf{g}_x + \mathbf{g}_x \hat{\mathbf{z}}^\dagger \hat{\mathbf{z}}\,\mathbf{g}_x^\dagger. \tag{S6}$$

Proof. The computation is analogous to the real case. Defining $\boldsymbol{Y} = (Y(t_1), \dots, Y(t_{n_*}))^\top$, i.e. $\boldsymbol{Y} = \mathbf{E}\mathbf{Z}$, 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 $\boldsymbol{Y} + \boldsymbol{\epsilon} = \mathbf{y}$ has a density proportional to

$$
\begin{aligned}
p_{\tilde{\mathbf{Z}}}(\tilde{\mathbf{z}} \mid \boldsymbol{Y} + \boldsymbol{\epsilon} = \mathbf{y}) &\propto p_{\tilde{\mathbf{Z}}, \boldsymbol{Y}+\boldsymbol{\epsilon}}(\tilde{\mathbf{z}}, \boldsymbol{Y} + \boldsymbol{\epsilon}) \propto p_{\mathbf{Z},\boldsymbol{\epsilon}}(\overbrace{\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\boldsymbol{\Lambda}^{-1}\mathbf{Q}\tilde{\mathbf{z}}\right) \cdot \\
&\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}}) \\
&\overset{(*)}{\propto} \exp\left(-\frac{1}{2}\left(\tilde{\mathbf{z}}_+^\dagger\mathbf{N}^\dagger\boldsymbol{\Lambda}^{-1}\mathbf{N}\,\tilde{\mathbf{z}}_+\right) - \Re\left(\tilde{\mathbf{z}}_+^\dagger\mathbf{N}^\dagger\boldsymbol{\Lambda}^{-1}\mathbf{z}_0\right)\right) \cdot \\
&\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\Bigg(-\frac{1}{2}\tilde{\mathbf{z}}_+^\dagger \underbrace{\mathbf{N}^\dagger\left(\boldsymbol{\Lambda}^{-1} + \mathbf{E}_+^\dagger\mathbf{T}_+^{-1}\mathbf{E}_+\right)\mathbf{N}}_{=\mathbf{S}^{-1}}\tilde{\mathbf{z}}_+ + \\
&\quad + \Re\Bigg(\tilde{\mathbf{z}}_+^\dagger \underbrace{\mathbf{N}^\dagger\left(\mathbf{E}_+^\dagger\mathbf{T}_+^{-1}(\mathbf{y}_+ - \mathbf{E}_+\mathbf{z}_0) - \boldsymbol{\Lambda}^{-1}\mathbf{z}_0\right)}_{=\mathbf{S}^{-1}\hat{\mathbf{z}}_+}\Bigg)\Bigg) 1_{\{\mathbf{M}^\dagger\mathbf{z}_0\}}(\tilde{\mathbf{z}}_0) \\
&\propto \exp\left(-\frac{1}{2}\left(\tilde{\mathbf{z}}_+ - \hat{\mathbf{z}}_+\right)^\dagger\mathbf{S}^{-1}\left(\tilde{\mathbf{z}}_+ - \hat{\mathbf{z}}_+\right)\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), (S3) and (S4) follow directly by linearity and (S5) from variance decomposition (omitting conditions for brevity):

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

The computation for (S6) is analogous. □

Warping alignment: Generally, we consider it advisable to base warping alignment of the ith curve directly on its original SRV-evaluations $q_{i1}^{[h]}, \ldots, q_{in_i}^{[h]}$ but, when considerable measurement error presents an issue, it might also be useful to employ a smoothed reconstruction $\tilde{q}_i : [0,1] \to \mathbb{C}$ 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 $\tilde{q}_i^{[h]}(t) = (\hat{\mathbf{z}}_i^{[h]}/\|\hat{\mathbf{z}}_i\|)^\top\hat{\mathbf{e}}^{[h]}(t)$ with $\hat{\mathbf{z}}_i^{[h]} = (\hat{z}_{i1}^{[h]}, \ldots, \hat{z}_{im}^{[h]})^\top$ the predicted score vector for the eigenbasis $\hat{\mathbf{e}}^{[h]} = (\hat{e}_1^{[h]}, \ldots, \hat{e}_m^{[h]})^\top$.

Following Steyer et al. (2021), 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 hyper-parameter $\rho \in [0,1]$, we control the balance between original $q_{ij}^{[h]}$ (for $\rho = 0$) and smoothed reconstruction $\tilde{q}_i$ (for $\rho = 1$) and set $\hat{q}_{ij}^{[h]} =$

$\hat{u}_i^{[h]} \left(\varrho\, \tilde{q}_i^{[h]}(t_{ij}^{[h]}) + (1-\varrho)\, q_{ij}^{[h]} \right)$ at nodes $s_{i0}^{[h]} = 0$, $s_{ij}^{[h]} = 2t_{ij}^{[h]} - s_{ij-1}^{[h]}$, $j = 1, \ldots, n_i$. 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 $i = 1, \ldots, n$ by finding an optimal $\hat{q}_i^* \in \mathbb{L}^2([0,1], \mathbb{C})$ with

$$\|\hat{q}_i^* - \hat{\psi}^{[h]}\| \leq \|\hat{q}_i^{[h]} \circ \gamma\, \dot{\gamma}^{1/2} - \hat{\psi}^{[h]}\| \quad \text{for all } \gamma \in \Gamma \tag{S7}$$

where the polygon approximation yields a practically feasible optimization problem and has proven suitable for sparse/irregular curves (Steyer et al., 2021). As shown by Steyer et al. (2021), the optimizers of (S7) 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 $w_i : [0,1] \to \mathbb{R}$ are given by $w_i^2(t) = (s_{ij}^{[h]} - s_{ij-1}^{[h]}) \Re \left(\psi^{[h]}(t)^\dagger \hat{q}_{ij}^{[h]} \right)_+^2 / \int_{s_{ij-1}^{[h+1]}}^{s_{ij}^{[h+1]}} \Re \left(\psi^{[h]}(t)^\dagger \hat{q}_{ij}^{[h]} \right)_+^2 dt$ for $t \in [s_{ij-1}^{[h+1]}, s_{ij}^{[h+1]})$, and fully determined by the warped time points

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

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 $w_{ij}^* = (s_{ij}^{[h]} - s_{ij-1}^{[h]})^{1/2} (s_{ij}^{[h+1]} - s_{ij-1}^{[h+1]})^{-1/2}$ for $s_{ij}^{[h+1]} > s_{ij-1}^{[h+1]}$ and omit double time points for $j = 1, \ldots, n_i$. The chosen time-points hereby approximate $t_{ij}^{[h+1]} \approx t_{ij}^* \in (s_{ij}^{[h+1]}, s_{ij-1}^{[h+1]})$ with $w_i(t_{ij}^*) = w_{ij}^*$ existing by the Mean Value Theorem.

3. 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 S1 exemplifies elastic full Procrustes mean estimation on the entire "*fda*" handwritings contained in the dataset `handwrit.dat` in the R package `fda` (Ramsay & Silverman, 2005). To visualize different degrees of sparsity, means are fitted after subsampling recorded points to $n_i = n_{points}$, $i = 1, \ldots, n$, $n = 20$, 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 $n_{points} = 71$.

REFERENCES

BRUVERIS, M. (2016). Optimal reparametrizations in the square root velocity framework. *SIAM Journal on Mathematical Analysis* **48**, 4335–4354.

HSING, T. & EUBANK, R. (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*. John Wiley & Sons.

RAMSAY, J. O. & SILVERMAN, B. W. (2005). *Functional Data Analysis*. Springer New York.

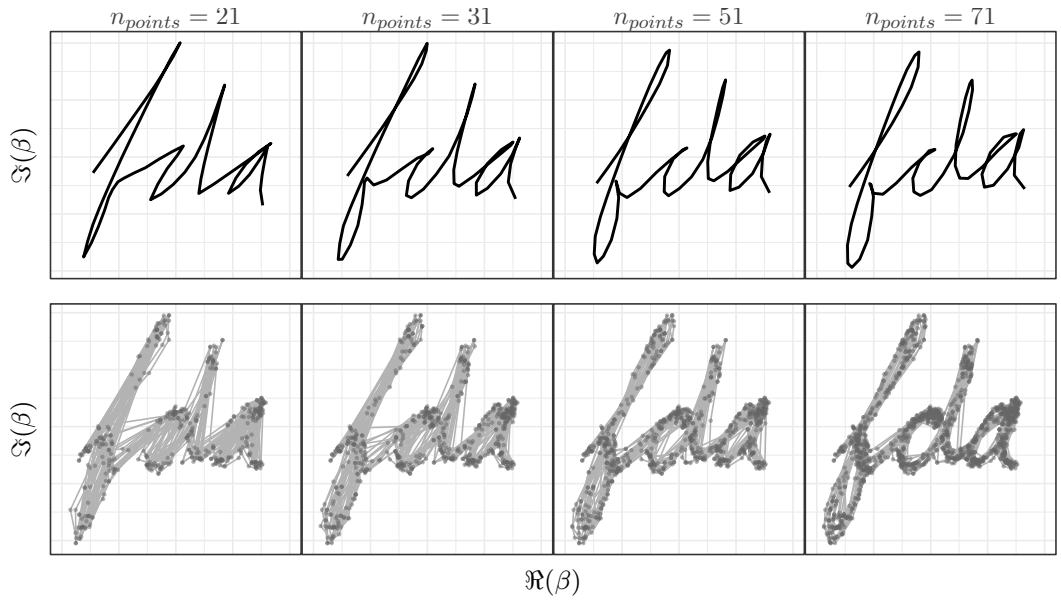

Fig. S1: *Top:* Elastic full Procrustes means estimated over 20 handwritten "*fda*"s sampled with different degrees of sparsity. *Bottom:* Underlying datasets with 20 curves from the `handwrit.dat` dataset subsampled with higher acceptance probability on points important for curve reconstruction. Points sampled for each curve are connected by light-grey lines.

RYNNE, B. & YOUNGSON, M. A. (2007). *Linear functional analysis*. Springer Science & Business Media.
STEYER, L. (2021). *elasdics: Elastic Analysis of Sparse, Dense and Irregular Curves*. R package version 0.1.3.
STEYER, L., STÖCKER, A. & GREVEN, S. (2021). Elastic analysis of irregularly or sparsely sampled curves. *arXiv preprint arXiv:2104.11039* .
WOOD, S. (2017). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2nd ed.
YAO, F., MÜLLER, H. & WANG, J. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* **100**, 577–590.

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 angegeben Quellen verwendet habe und die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Gräfelfing, den 27.07.2022

Ort, Datum Unterschrift Jan Almond Stöcker