Alexander Bauer

Flexible Approaches in Functional Data and Age-Period-Cohort Analysis with Application on Complex Geoscience Data

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

Eingereicht am 04.03.2022

Alexander Bauer

Flexible Approaches in Functional Data and Age-Period-Cohort Analysis with Application on Complex Geoscience Data

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

Eingereicht am 04.03.2022

Erster Berichterstatter: Prof. Dr. Helmut Küchenhoff
Zweiter Berichterstatter: PD Dr. Fabian Scheipl
Dritter Berichterstatter: Assoc. Prof. Jeff Goldsmith

Tag der Disputation: 24.06.2022

> *Wo man massvolle Verkündung der Resultate erwartet, die sich als Vorläufer einer genaueren Erkenntniss ausgeben, hört man rechthaberisches Pochen auf das wenige gefundene, Vertheidigung sogar der richtigen Ansichten durch halbwahre Gründe. Keine planvolle Ermittlung der thatsächlichen Verhältnisse, kein selbstbewusstes Hinstreben nach wohlerkanntem Ziele; vielmehr zufällig entstandene Tabellen aufs willkürlichste ausgebeutet. Wie der Reiter, wenn er die Zügel verloren hat, seinem Ross, so folgt häufig der Schriftsteller willenlos seinem Rechenstifte, und ein Gebiet, worauf der menschliche Geist seine Herrschaft, wie überall, befestigen sollte – wie oft sieht es den menschlichen Geist der rohsten Empirie unterliegen.*

Knapp (1868) on the struggle to let pure empirical evidence shape scientific discussions.

Acknowledgments

Many wonderful people have accompanied and greatly supported me over the past five years to make this work possible. My deepest gratitude goes to …

… *first and foremost my supervisors Helmut and Fabian, for your steady support and assistance over all the years. Without you, this work would have definitely not been possible. I greatly appreciate the opportunity to have done my Ph.D. under your supervision. Thank you, Helmut, for all the amazing, inspiring years at the StaBLab!*

… *Jeff Goldsmith for acting as the third reviewer for this dissertation, and Volker Schmid and Christian Heumann for your availability to be part of the examination panel at the disputation.*

… *my office mates André, Max, Vroni and Simon for the chill and supporting atmosphere, including many fruitful and – sometimes a bit too philosophical – discussions.*

… *all other members of the StaBLab family, especially Andreas, Christian, Felix, Pia, Yegi, Noemi (I miss your Christmas cookies...), Linda and Yvonne.*

… *all the other awesome people at our department for the enjoyable and positive vibe, especially Giacomo (porca puttana!), Ben, Martje, Xudong, Cornelius, Marc, Michael and Michi. Thank you for countless epic football matches, poker nights and other activities. Special props to the MedBay crew for keeping up the spirits during all the lockdowns! Also a special thanks to Brigitte and Pia and your strong nerves, for the countless hours you were forced to spend on discussing, preparing and steadily readjusting my contract, just a bit too often.*

… *Jojo for extensive expeditions on Artificial Island, always accompanied by some thorough five o'clock tea testing (or sometimes beer).*

… *all other collaborators who have crossed my way, for lots of fascinating discussions about all fields of science. Most prominently, I thank Marion Karl, Jürgen Schmude, Alice-Agnes Gabriel, Johanna Gernert, Elisabeth Bartl, Julia Wrobel and Hawre Jalal.*

Summary

Functional data analysis and age-period-cohort (APC) analysis both deal with the estimation of flexible (association) structures over domains like time and space. This dissertation focuses on the development of new approaches to robustly estimate such structures. All methods are motivated and showcased with applications in seismological research or tourism science. The outlined approaches are implemented in open-source software packages in the statistical software R and are accompanied by detailed instructions on how to properly apply them.

The main methodological contribution of this dissertation focuses on the separation of amplitude and phase variation as a central component of functional data analysis. When functional data comprise a relevant amount of phase variation, they have to be *registered* to align the phase of the individual curves by elastically deforming their domains. While registration became an active field of research over the last decades, approaches for common data structures like generalized and not completely observed data are still lacking. The first contributing article adapts a likelihood-based registration method for such generalized and incomplete data, and joins it with an approach for generalized functional principal component analysis (GFPCA) to represent the obtained solution in a low-dimensional basis. The utility of the method is showcased on simulated seismic data with a highly volatile Gamma structure, used to assess critical geophysical parameters associated with strong ground motion in the event of an earthquake. The method's performance is compared to other established registration approaches. An implementation is available in the R package `registr` which is outlined in the second article.

After the separation of amplitude and phase, functional data can be analyzed in ways conceptually similar to scalar data. One common research question is to estimate the association between observed structures in curves and a set of scalar covariates. The third article gives an introduction to generalized functional additive models (GFAMs) as a flexible semiparametric approach to estimate such function-on-scalar regression models. Practical guidelines are discussed for all relevant aspects based on the analysis of the above-mentioned seismic data. Practical researchers are guided by offering fully reproducible code as well as the R package `FoSIntro` which contains numerous utility functions.

The third part of this dissertation covers APC analysis as a technique to analyze the underlying drivers of long-term temporal processes. The critical challenge in APC analysis is the linear dependency of the three dimensions age, period, and cohort (cohort = period - age). Flexible methods for estimation and visualization are needed to properly disentangle observed temporal structures. The fourth article introduces ridgeline matrices and *partial APC plots* as novel visualization techniques, refining the concepts of established techniques like Lexis diagrams. Model-based separation of the temporal dimensions is performed utilizing the semiparametric estimation of a two-dimensional tensor product surface with a generalized additive model (GAM). The usefulness of the methods is showcased with data from tourism science, analyzing drivers for altering travel distances of German tourists over the last decades. The fifth article covers the R package `APCtools` which implements the newly introduced methods as well as additional visualization techniques.

Zusammenfassung

Funktionale Datenanalyse und Alters-Perioden-Kohorten (APC-)Analyse befassen sich mit der Schätzung flexibler (Assoziations-)Strukturen über Domains wie Zeit und Raum. Die vorliegende Dissertation behandelt die Entwicklung neuer Ansätze zur robusten Schätzung solcher Strukturen. Alle Methoden werden anhand von Anwendungen in der seismologischen Forschung oder der Tourismuswissenschaft motiviert und eingeführt. Die methodischen Ansätze sind in Open-Source-Softwarepaketen in der Statistik-Software R implementiert und werden begleitet von detaillierten Ausführungen zu ihrer korrekten Anwendung.

Der zentrale methodische Beitrag dieser Dissertation fokussiert sich auf die Trennung von Amplituden- und Phasenvariation als eine der zentralen Komponenten funktionaler Datenanalyse. Enthalten funktionale Daten eine relevante Menge an Phasenvariation, so müssen sie *registriert* werden, um die Phasen der einzelnen Kurven durch elastische Verformung ihrer Domains anzugleichen. Obwohl sich die Registrierung über die letzten Jahrzehnte zu einem aktiven Forschungsgebiet entwickelt hat, fehlen weiterhin flexible Ansätze zur Analyse verbreiteter Datenstrukturen wie etwa generalisierter oder nicht vollständig beobachteter Daten. Der erste Artikel dieser Dissertation adaptiert eine Likelihood-basierte Registrierungsmethode für solche generalisierten und unvollständigen Daten und verbindet sie mit einem Ansatz zur generalisierten funktionalen Hauptkomponentenanalyse (GFPCA), um die erhaltene Lösung in einer niedrigdimensionalen Basis zu repräsentieren. Die Nützlichkeit der Methode wird an simulierten seismischen Daten mit einer hochvolatilen Gamma-Struktur demonstriert. Diese Daten werden zur Einschätzung zentraler geophysikalischer Parameter verwendet, welche im Falle eines Erdbebens mit starken Bodenbewegungen assoziiert sind. Die Performanz der Methode wird mit anderen etablierten Registrierungsansätzen verglichen. Eine Implementierung ist im R-Paket `registr` verfügbar, das im zweiten Artikel beschrieben wird.

Nach der Separierung von Amplitude und Phase können funktionale Daten auf konzeptionell ähnliche Weisen analysiert werden wie skalare Daten. Eine häufige Forschungsfrage ist die Schätzung des Zusammenhangs zwischen beobachteten Strukturen in Kurven und einer Reihe von skalaren Kovariablen. Der dritte Artikel gibt eine Einführung in generalisierte funktionale additive Modelle (GFAMs), welche einen flexiblen semiparametrischen Ansatz zur Schätzung solcher Funktion-auf-Skalar-Regressionsmodelle darstellen. Anhand der Analyse der erwähnten seismischen Daten werden praktische Leitlinien für alle relevanten Aspekte der Methode diskutiert. Mit Blick auf Fachwissenschaftler wurde über den vollständig reproduzierbaren Code hinaus das R-Paket `FoSIntro` entwickelt, welches bei der Durchführung von Regressionsanalysen Unterstützung bietet.

Der dritte Teil dieser Dissertation befasst sich mit APC-Analyse als einer Technik zur Analyse der zugrundeliegenden Treiber langfristiger zeitlicher Prozesse. Die zentrale Herausforderung von APC-Analysen bildet die lineare Abhängigkeit der drei Dimensionen Alter, Periode und Kohorte (Kohorte = Periode - Alter). Flexible Methoden zur Schätzung und Visualisierung sind erforderlich, um die beobachteten zeitlichen Strukturen adäquat zu entflechten. Der vierte Artikel führt Ridgeline-Matrizen und *partial APC plots* als neue Visualisierungstechniken ein. Diese verfeinern etablierte Techniken wie etwa Lexis-Diagramme. Die modellbasierte Trennung der zeitlichen Dimensionen wird durch die semiparametrische Schätzung einer zweidimensionalen Tensorprodukt-Oberfläche mit einem generalisierten additiven Modell (GAM) durchgeführt. Die Nützlichkeit der Methoden wird anhand von Daten aus der Tourismuswissenschaft demonstriert, anhand welcher die Treiber analysiert werden, welche mit sich verändernden Reisedistanzen deutscher Touristen über die letzten Jahrzehnte assoziiert sind. Der fünfte Artikel behandelt das R-Paket `APCtools`, welches die neuen Methoden sowie zusätzliche Visualisierungstechniken implementiert.

Contents

Part I.

Introduction and Background

1. Introduction

1.1. Outline

This dissertation tackles the statistical analysis of data where some process of interest is observed over continuous domains like time and space, and where patterns in the development over these domains are of central interest. Contributions are made to the fields of functional data analysis and age-period-cohort (APC) analysis and focus on one-dimensional curves and multi-dimensional temporal processes, respectively. For functional data analysis, advances cover the alignment of salient structures along the functional domain in a *registration step* to properly disentangle the underlying sources of variation of incompletely observed curves with a non-Gaussian structure. Further, the focus is on methods for function-on-scalar regression, where the association structures between a functional response and purely scalar covariates are of interest. The contributions to APC analysis most prominently comprise the development of adequately complex, yet accessible visualization techniques for a – potentially model-based – analysis. The main goal is to disentangle observed developments with respect to their variation structure along the three temporal dimensions age, period and cohort. All methods are motivated by real-world applications in either seismological research or tourism science.

The remainder of this introductory chapter is organized as follows. This section gives a brief overview of the research questions covered by this dissertation and motivates their statistical relevance. Sections 2 to 4 introduce the individual statistical problems, summarize the utilized methodological approaches and potential alternative techniques, comment on the current state of statistical research in the respective field and shortly discuss potential directions for future research. The individual contributions that are the core of this dissertation are listed in Chapters II to IV. In the spirit of open science and to ensure full reproducibility, all contributions are accompanied by versatile implementations of the methodological approaches in the form of packages for the statistical open-source software `R` (R Core Team, 2021).

1.2. Motivation and Scope

Many fields in modern science are grounded in the quantitative analysis of data. Over the last centuries, not only have surveys grown in size and has the design of experiments itself become a branch of research, but scientific studies have also become more complex with respect to the variety of different parameters that they control for. Further accelerated by steady advances in information technology, researchers were enabled to collect and analyze increasingly large, densely observed and high-dimensional data. Two research areas that benefitted greatly from this development are functional data analysis and age-period-cohort (APC) analysis. APC analysis already became an established field of research in the 1800s, driven by the growing interest in the development of mortality rates (see e.g. Knapp, 1868). Building on advances in the understanding of time-dependent stochastic processes, functional data analysis became more prominent in the late 1900s, culminating in the 1997 publication of the first edition of the standard work of Ramsay and Silverman (2005). Nowadays, both APC analysis and functional data analysis are still active fields of research and are applied in diverse settings. However, substantial challenges remain that require further research.

Functional data are shaped by repeated measurements, e.g. observed over time. While sharing many aspects with longitudinal data analysis, functional data analysis typically deals with more densely observed measurements per curve and shifts the overall focus towards alterations in the curves' shape. Over time, unique strategies were developed to tackle these challenges. Two approaches that are central to the analysis of functional data are the alignment of curves in a *registration* step and the estimation of association structures in a regression framework. Registering curves is crucial since observed processes are not only shaped by their variation along some parameter of interest (*amplitude variation* along the y-axis), but also by their variation along the functional domain (*phase variation* along the x-axis). Both types of variation have to be properly represented and disentangled to ensure unbiased results in subsequent analyses. In regression-based analyses, functional data require the estimation of flexible and often multi-dimensional association structures. To ensure their practical applicability, methods for functional data analysis have to be robust and efficient even in data situations when data is observed on irregular, sparse or only incompletely observed domains. Chapter 2 focuses on the first contribution Bauer et al. (2022a) which introduces a method to register curves whose processes entail a non-Gaussian structure and are only observed incompletely, as well as on the second contribution Wrobel and Bauer (2021), outlining the respective implementation in the R package `registr`. Chapter 3 summarizes the third contribution Bauer et al. (2018) which highlights a flexible semi-parametric approach for function-on-scalar regression and thoroughly discusses related practical considerations.

In contrast to functional data analysis, APC analysis simultaneously focuses on its three eponymous temporal domains. Observed developments of a process of interest can be associated with a person's life cycle (age effect), with changes affecting the whole population over some time period, like macro-economic developments or scientific progress in medicine (period effect) or with structural differences between members of different generations like socialization or exposure processes (cohort effect). Nowadays, such research questions appear frequently not only in demographic or epidemiological contexts but also in economic, social or general medical sciences. Similar to functional data analysis, flexible and robust approaches are required both for the estimation and for the visualization of nonlinear associations with said temporal dimensions. The central challenge for statistical approaches in APC analysis is to circumvent the *identification problem*, which describes the linear dependency $cohort = period - age$ of the temporal dimensions. Many regression-based approaches solve this problem by estimating linear effects while putting hard constraints on specific parameters. Chapter 4 summarizes the fourth contribution Weigert et al. (2021) which highlights a semiparametric approach without hard constraints for effect estimation and introduces novel visualization techniques. Further, Chapter 4 also covers the fifth contribution Bauer et al. (2022b) outlining the respective implementation in the R package `APCtools`.

All methodological developments in this dissertation are driven by interdisciplinary research projects. The advances in functional data analysis in Chapters 2 and 3 are motivated by seismic ground motion data which are derived from large-scale *in silico* earthquake scenario simulations with the open-source software SeisSol (Pelties et al., 2014; Uphoff et al., 2017, github.com/SeisSol/SeisSol), based on a real seismic event that took place in Northridge (California) in 1994. The aim of the statistical analysis is to gain a better understanding of the associations between initial seismic conditions like fault stress and fault strength prior to earthquakes as well as local topography and geology with the temporal and spatial distribution of ground movement caused by an earthquake. Especially since the prediction of when and where the next seismic event will occur remains an unsolved problem, the results contribute to the necessary risk assessment for such events.

The advances in APC analysis in Chapter 4 are motivated by research questions from tourism science and are based on extensive data obtained between 1971 and 2018 in the German *Reiseanalyse* survey, an annual cross-sectional survey on pleasure travel among approximately 7 500 German residents (FUR Forschungsgemeinschaft Urlaub und Reisen e.V., 2020). Main interest is on how the travel behavior of the German population changed over the last decades, and what the drivers were for these changes. Travel behavior is the result of a (partly sub)conscious travel decision process, which itself is shaped by the personal circumstances under which it takes place. These circumstances can change over someone's life cycle (e.g. the family situation) or over calendar years (e.g. the price for long-distance trips) and can vary between generations (e.g. as a result of different socialization processes). Adequately accounting for all relevant aspects that shape the travel decision process (e.g. also financial and health constraints) is crucial to draw potential conclusions about the drivers of observed developments. Eventually, since individual travel behavior is the result of such a highly individualistic decision process, the underlying research project also partly contributes to a better understanding of societal change in the German society over the last 50 years.

As focus in Chapters 2 to 4 is on the advances of the statistical methodology, conclusions about the underlying geoscientific research questions are only briefly discussed. For full detail I refer to the individual contributions in Chapters II to IV.

2. Registration of Functional Data

Functional data have two central modes of variation, namely phase variation and amplitude variation. As already shortly outlined, amplitude and phase variation define the variation of a process along some parameter of interest (i.e., along the y-axis) and the variation along the functional domain (i.e., along the x-axis), respectively. For example, when analyzing curves consisting of seismic ground motion measurements over time and focusing on their initial peaks, amplitude variation represents the overall strength of the observed ground motion. In contrast, phase variation represents the effective time distortion at the individual peaks, caused by differing propagation speeds of the seismic waves under different physical conditions. Even if phase variation is often not of main interest, methods for functional data analysis have to properly differentiate between both types of variation by *registering* the individual curves to prevent bias in subsequent analyses.

This section focuses on the contributions Bauer et al. (2022a) and Wrobel and Bauer (2021), which introduce and implement a likelihood-based approach for registering curves with a non-Gaussian structure and whose processes were only observed incompletely. The registration approach is paired with an approach for generalized functional principal component analysis (GFPCA) to represent the aligned curves in a lower-dimensional basis.

2.1. Incomplete Curve Registration

Let $Y_i(t_i^*)$, $i = 1, \dots, N$ be discretized incomplete curves, observed over individual *chronological time domains* T_i^*, where the observed grids $\boldsymbol{t}_i^* = [t_{ij}^*]_{j=1,\dots,D_i}$ may be irregular or sparse, and chronological domains $T_i^* = [t_{\min,i}^*, t_{\max,i}^*]$ are defined by the individual observation periods. Without loss of generality, we assume that all observed curves are realizations of stochastic processes over a common underlying *internal time domain* $T = [0,1]$ and $T_i^* \subseteq T \; \forall \, i$. Registering the curves requires estimating inverse warping functions $h_i^{-1} : T_i^* \mapsto T$ that account for the data's phase variation and map the individual chronological time domains T_i^* to the internal time domain T of the underlying process. The resulting registered curves $Y_i(t) = Y_i(h_i^{-1}(t_i^*))$ only contain amplitude variation.

We differentiate between three types of incompleteness that can be present in observed curves. *Leading incompleteness* and *trailing incompleteness* are present when information is missing only at the beginning and only at the end of some process, respectively. If both are present at the same time, we use the term *full incompleteness*. While incompleteness occurs in many practical settings, most existing approaches only focus on the registration of completely observed curves. As visualized for synthetic data with trailing incompleteness in Figure 2.1 based on the established approach of Srivastava et al. (2011), the application of conventional "complete curve" registration methods to incomplete data often leads to nonsensical results. In contrast, our approach can handle all three types of incompleteness and is able to properly align the curves.

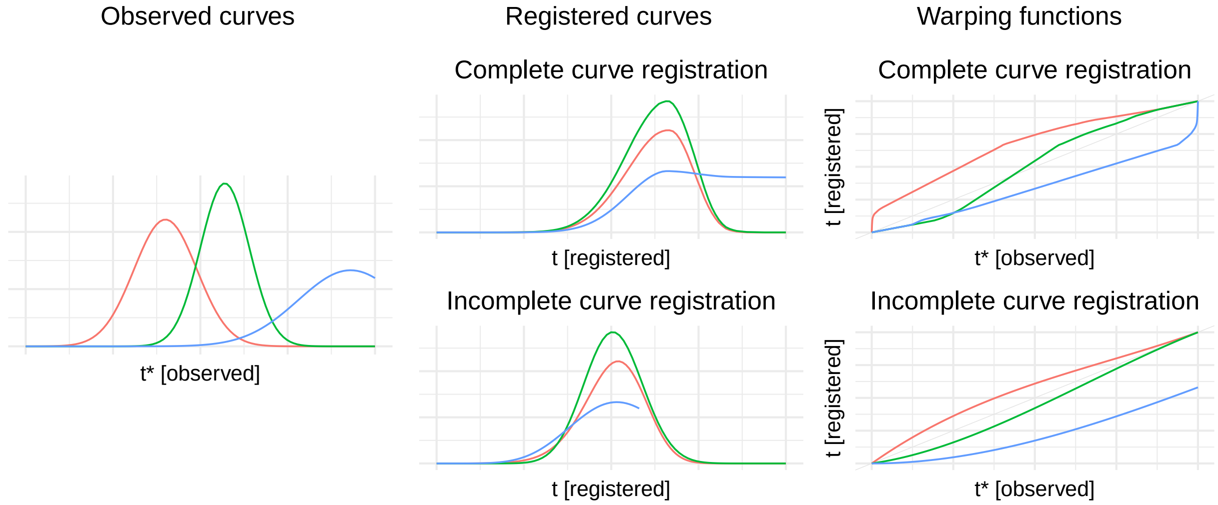

Figure 2.1.: Simulated observed curves with trailing incompleteness (left column), registered curves only comprising amplitude variation (center), and inverse warping functions visualizing phase variation (right). Registration was performed with the complete curve SRVF approach of Srivastava et al. (2011) with function `time_warping()` from the R package `fdasrvf` (top row, Tucker, 2020) and our incomplete curve approach (bottom row, Bauer et al., 2022a). Note the extreme time dilation around the blue curve's peak in the top row, which yields a highly implausible registered curve.

For our registration approach, we extend the likelihood-based framework of Wrobel et al. (2019) for registering complete curves from exponential family distributions. Inverse warping functions h_i^{-1} are estimated by mapping an unregistered curve $Y_i(t_i^*)$ toward a suitable template function $\mu_i(t)$ so that

$$\begin{aligned} \mathbb{E}\left[Y_i\left(h_i^{-1}(t_i^*)\right)|h_i^{-1}\right] &= \mu_i\left(t\right), \\ \text{with} \quad h_i^{-1}(t_i^*) &= \mathbf{\Theta}_h(t_i^*)\boldsymbol{\beta}_i. \end{aligned} \tag{2.1}$$

The inverse warping functions are represented through a B-spline basis with design matrix $\mathbf{\Theta}_h \in \mathbb{R}_{D_i \times K_h}$, with D_i the number of measurements for curve i, K_h the number of basis functions and the corresponding coefficient vector $\boldsymbol{\beta}_i$. Given some distribution from the exponential family, this yields the following log-likelihood for curve i:

$$\ell\left(h_i^{-1}|y_i,\mu_i\right) = \log\left(\prod_{j=1}^{D_i} f_{i,j}\left[y_i(t_{i,j}^*)\right]\right), \tag{2.2}$$

with $f_{i,j}(\cdot)$ the corresponding density with expected value $\mu_i\left(h_i^{-1}(t_{i,j}^*)\right)$ and observed vectors of function evaluations $y_i(t_{ij}^*)$, $j = 1,\dots,D_i$. We impose working assumptions of mutual conditional independence across functions $[Y_i \perp Y_{i'}]_{i\neq i'}|\mu_i,\mu_{i'}$ as well as within functions $[Y_i(t_{ij}) \perp Y_i(t_{ik})]_{j\neq k}|\mu_i$. The basis coefficients $\boldsymbol{\beta}_i$ are constrained in the optimization to ensure strictly increasing warping functions that do not exceed the overall domain of the underlying process.

If all curves were observed on an identical time interval, domain preservation would require to map t_{min}^* and t_{max}^* to themselves so that the warping functions begin and end on the diagonal line. Forcing observed and registered domain lengths to be identical, however, is clearly unsuitable for incomplete curve settings. Accordingly, we allow warping functions to start and/or end at any point inside the overall time domain. To avoid large domain deformations that are not strongly supported by the data, we penalize the total amount by which the registration changes the duration of an observed time domain, leading to the penalized log-likelihood

$$\begin{aligned} \ell_{\text{pen}}\left(h_i^{-1}|y_i,\mu_i\right) &= \ell\left(h_i^{-1}|y_i,\mu_i\right) - \lambda \cdot n_i \cdot \text{pen}\left(h_i^{-1}\right), \\ \text{with} \qquad \text{pen}\left(h_i^{-1}\right) &= \left(\left[h_i^{-1}(t^*_{\max,i}) - h_i^{-1}(t^*_{\min,i})\right] - \left[t^*_{\max,i} - t^*_{\min,i}\right]\right)^2. \end{aligned} \tag{2.3}$$

For settings with leading incompleteness with $h_i^{-1}(t^*_{\max,i}) = t^*_{\max,i}\ \forall\, i$, this simplifies to $\text{pen}\left(h_i^{-1}\right) = \left[h_i^{-1}(t^*_{\min,i}) - t^*_{\min,i}\right]^2$ and for trailing incompleteness with $h_i^{-1}(t^*_{\min,i}) = t^*_{\min,i}\ \forall\, i$ to $\text{pen}\left(h_i^{-1}\right) = \left[h_i^{-1}(t^*_{\max,i}) - t^*_{\max,i}\right]^2$. These penalties for one-sided incompleteness represent the squared distance of the respective endpoint of h_i^{-1} to the diagonal. The penalization parameter λ controls the overall amount of time dilation or compression and is scaled by the number of measurements n_i of curve i to ensure that the impact of the penalization relative to the likelihood is not affected by the number of measurement points per function. The choice of λ should be based on substantive knowledge so that estimated warping functions represent realistic accelerations and/or decelerations of the observed processes.

Further, we adapt the *two-step approach* of Gertheiss et al. (2017) to estimate a low-rank GFPCA representation of the registered curves $Y_i(t) = Y_i\left(h_i^{-1}(t_i^*)\right)$. Following Hall et al. (2008) and the groundwork of Yao et al. (2005), functional principal components (FPCs) are estimated using a marginal, semiparametric method based on assuming a latent Gaussian process $X_i(t)$, so that

$$\begin{aligned} \mathbb{E}[Y_i(t)] = \mu_i(t) &= g[X_i(t)], \\ X_i(t) &\approx \alpha(t) + \sum_{k=1}^{K} c_{i,k} \cdot \psi_k(t), \end{aligned} \tag{2.4}$$

where each observed $Y_i(t)$ corresponds to the transformation of a latent process $X_i(t)$ with some fixed response function $g(\cdot)$, and the latent process itself can be decomposed into a smooth global mean $\alpha(t)$, FPCs $\psi_k(t)$ with respective eigenvalues $\tau_k > 0$, and FPC scores $c_{i,k} \sim N(0,\tau_k)$. With known $\psi_k(t)$, model (2.4) is a generalized functional additive mixed model along the lines of Scheipl et al. (2016) with a smooth conditional latent mean function in a P-spline representation, functional random effects in an FPC basis representation, and functional random effect scores $c_{i,k} \sim N(0,\tau_k)$.

To derive the GFPCA solution, we first center the observed $Y_i(t)$ based on their marginal mean $\mu_Y(t) = \mathbb{E}[Y_i(t)]$, estimated through a simple smoother of all $(t_{ij}, Y_i(t_{ij}))$-pairs via a generalized additive model (GAM, Wood, 2017) with response function $g(\cdot)$ and an appropriate exponential family for the response. The covariance of the latent process can then be approximated by

$$\widehat{\mathsf{Cov}}\left[X_i(s), X_i(t)\right] \approx \frac{\hat{\sigma}_Y(s,t)}{g^{(1)}[\mu_X(s)] \cdot g^{(1)}[\mu_X(t)]}, \tag{2.5}$$

with $\sigma_Y(s,t) = \mathbb{E}[Y_{c,i}(s) \cdot Y_{c,i}(t)]$ based on the centered curves $Y_{c,i}(t)$, the marginal mean $\mu_X(t)$ estimated accordingly to $\mu_Y(t)$, and $g^{(1)}(\cdot)$ the first derivative of the response function. For given time points s_1 and s_2, $\sigma_Y(s_1,s_2)$ is estimated as the mean of all pairwise products $y_{c,i}(s_1) \cdot y_{c,i}(s_2)$.

The estimated surface $\hat{\sigma}_Y(s,t)$ is a smoothed version of $\sigma_Y(s,t)$ based on a bivariate tensor product P-spline basis (Fahrmeir et al., 2013). The FPCs $\psi_k(t)$ and their respective eigenvalues τ_k are then estimated from the spectral decomposition of $\widehat{\mathsf{Cov}}\left[X_i(t), X_i(s)\right]$.

Finally, we combine the approaches for registration and GFPCA by utilizing the iterative algorithm of Wrobel et al. (2019). Our aims to do so are twofold: (i) register all observed curves $Y_i(t_i^*)$ to suitable template functions and (ii) adequately represent the registered curves $Y_i(t) = Y_i(h_i^{-1}(t_i^*))$ through a low-rank GFPCA basis. We solve this problem by alternating the registration step (conditional on the current GFPCA representations $\mu_i(t)$) and the GFPCA step (conditional on the current estimates of the warping functions h_i^{-1}). The initial registration step is performed with respect to a fixed common template function $\mu(t)^{[0]}$ which has to be set by the user. Subsequent iterations then use the low-rank GFPCA representations $\mu_i(t)$ as curve-specific template functions, with the number of FPCs in each iteration chosen, e.g., based on the explained share of variance.

2.2. Alternative Approaches

A range of alternative approaches exists for curve registration. Many early approaches focused on the alignment of curves towards given (salient) structures, i.e., landmark registration (e.g. Kneip and Gasser, 1992). More recent proposals mostly perform alignment towards *template functions* that are either based on domain knowledge or estimated based on some measure of centrality. Warping functions are commonly estimated purely nonparametrically (e.g. Tucker, 2020; Chakraborty and Panaretos, 2021), as (piecewise) linear functions (e.g. Sangalli et al., 2010; Vitelli, 2019; McDonnell et al., 2021) or in a basis expansion like a (penalized) B-spline (Telesca and Inoue, 2008; Wrobel et al., 2019) or Fourier (Mattar et al., 2009) basis or using warplets (Claeskens et al., 2010).
One of the most popular approaches is the square-root velocity function (SRVF) framework introduced by Srivastava et al. (2011), who showed that the warping-invariant Fisher-Rao metric quantifies pure amplitude distances and is equivalent to the simple L_2-metric in the SRVF space. Recently, Guo et al. (2020) have extended this framework to jointly analyze amplitude, phase and spatial variation. Cheng et al. (2016), Kurtek (2017), Lu et al. (2017) and Ebert et al. (2021) adapted the SRVF approach to perform registration in a Bayesian setting. Further Bayesian approaches were introduced for data settings with stronger noise under informative priors (Matuk et al., 2019; Tucker et al., 2021). The method developed by Horton et al. (2021) allows to include prior information about the specific placement of landmarks on the registered domain. While Bayesian approaches can provide a full representation of the joint phase and amplitude uncertainty, they are also computationally very demanding.
More recently, Nunez et al. (2021) and Chen and Srivastava (2021) introduced the neural network-based frameworks SrvfNet and SrvfRegNet, respectively. Both approaches build on the SRVF framework and enable a highly efficient estimation and prediction of warping functions in large-scale data settings.

In contrast to our likelihood-based method, all the above approaches are limited to continuous functional data, mostly under the assumption of Gaussian errors. Some extensions to binary functional data circumenvent this restriction by utilizing a pre-smoothing step to obtain a continuous representation of the curves (e.g. Wu and Srivastava, 2014; Panaretos and Zemel, 2016). Alternatively, the *congealing* approach of Learned-Miller (2005) (adapted to functional data by Mattar et al., 2009) iteratively optimizes general measures of alignment and is applicable in diverse data situations.

For analyzing the main modes of amplitude and phase variation, it is common to estimate a low-rank representation of the registered curves and potentially also of the estimated warping functions. Tucker et al. (2013), Hadjipantelis et al. (2015), Lee and Jung (2016) and Happ et al. (2019) use (joint) functional principal component analysis for finding compact representations of both phase and amplitude modes of variation.
Tucker (2014) (utilizing the SRVF approach of Srivastava et al., 2011, for registration), Wagner and Kneip (2019) and Kneip and Ramsay (2008) (optimizing a least squares criterion) and Wrobel et al. (2019) (optimizing an exponential family likelihood) all utilize iterative algorithms to successively refine warping functions that lead to registered curves whose amplitude variation can be represented in terms of a low-rank FPC basis.

Comparatively few approaches exist for the registration of incomplete curves. Some heuristic approaches were developed in the field of dynamic time warping (DTW) for time series analysis. Subsequence DTW aims to find a subsequence of a (fully observed) template curve to which a partially observed curve can be matched (see Müller, 2015, 7.2). Tormene et al. (2009) introduce an algorithm for "open-begin and open-end" DTW (OBE-DTW) that is also able to handle full incompleteness.
More sophisticated registration approaches were only introduced recently. Sangalli et al. (2010) and Vitelli (2019) make use of linear warping functions with free starting points and endpoints. These warpings are constrained so that they dilate or compress the observed domains by maximal factors between 0.9 and 1.1 to ensure reasonable results. The Bayesian approach of Matuk et al. (2019) allows to analyze sparse and fragmented Gaussian functional data, but their approach is feasible only for small to intermediate datasets.
Bryner and Srivastava (2021) very recently introduced an approach for *elastic partial matching* to tackle trailing incompleteness. Before registering each curve to its template using the complete curve SRVF approach of Srivastava et al. (2011), they estimate the time scaling necessary to (partially) match the observed domain of a specific curve to the domain of the template curve, and perform the registration only on the intersection of the curves' domains. Both steps are combined in a joint, gradient-based algorithm.

2.3. Contribution and Prospects

Incomplete data are very common in longitudinal settings but remain under-discussed in many fields of functional data analysis. Our likelihood-based approach for joint registration and generalized FPCA allows for analyzing curves with leading, trailing or full incompleteness in the presence of substantial phase variation and is able to handle non-Gaussian data.
The simulation study results in Bauer et al. (2022a) indicate that accounting for incompleteness improves the performance in different data settings. While our approach shows some bias in the estimation of the underlying FPC structure for curves with a Gamma structure, its substantially better estimation of the warping functions leads to improved overall performance in terms of the representation of the joint phase and amplitude variation structure of the individual curves. Stronger incompleteness does not seem to structurally harm the overall performance.
In contrast to methods based on the SRVF framework of Srivastava et al. (2011), we deliberately do not utilize the warping-invariant Fisher-Rao metric. Instead, our flexible penalized

likelihood-based approach allows for representing more complex structures of variation in diverse non-Gaussian data situations and is backed by robust optimization algorithms. While SRVF approaches rely on the availability of functional derivatives evaluated on a common, regular grid and may struggle in the presence of stronger (non-Gaussian) noise, this is generally not the case for our method.

The statistical approach for jointly applying registration and GFPCA is implemented in R package `registr` (Wrobel and Bauer, 2021). The package comprises an implementation of the original approach of Wrobel et al. (2019) as well as the incomplete curve approach of Bauer et al. (2022a), and allows for the registration of Gaussian and non-Gaussian curves from several exponential family distributions with leading, trailing or full incompleteness.

The joint registration and GFPCA approach proved its worth in the application on the seismic ground velocity data. The method takes into account both the incompleteness of the observed curves and their non-Gaussian structure. Bauer et al. (2022a) provide a thorough discussion of estimated amplitude and phase variation patterns. All obtained results are geophysically plausible and in line with previous analyses of the seismic experiments (Bauer et al., 2017).

Several issues exist that should be addressed by future research on curve registration. First, the registration literature still lacks sophisticated methods to robustly handle incomplete data settings. This was the motivation for us to approach this issue and – together with the works of Matuk et al. (2019) and Bryner and Srivastava (2021) – we made an important step towards filling this research gap. However, each of these three approaches still has its specific shortcomings. Second, the practical applicability of general registration approaches is most often limited by the computational cost for larger-scale data and / or the missing ability to naturally handle non-Gaussian data structures.

With many techniques for functional data analysis being grounded on utilizing the covariance structure in the data, one central topic for future research on GFPCA is a thorough evaluation of the consistency and robustness of different covariance estimators. This comprises questions like at what point in the estimation procedure smoothing and centering (of the raw curves or the final covariance surface) should be performed to obtain the best estimator. Covariance estimators should be evaluated for common practical data settings entailing relevant non-Gaussian noise in combination with small numbers of curves and measurements per curve and different levels of their respective density over the domain.

A practical constraint for the application of the evaluated methods remains their computational efficiency in large-scale data settings. In this regard, a promising strain of research are the recently proposed neural network based frameworks by Nunez et al. (2021) and Chen and Srivastava (2021) for registration (as noted in Section 2.2) and by Sarkar and Panaretos (2021) for covariance estimation.

Regarding the seismic application, future research can build on the aligned curves and estimated warping functions and use them as (functional) response to perform regression analyses, for example similar to the seismic application performed in Bauer et al. (2018). As outlined in Section 3.4, both amplitude and phase variation are relevant for geophysicists in the setting at hand. Respective regression analyses on amplitude and phase variation could also be performed jointly in a multivariate functional regression setting (see e.g. Volkmann et al., 2021).

3. Function-on-Scalar Regression

When functional data only comprise amplitude variation – potentially after registering observed curves in a pre-processing step – subsequent analyses can be performed. One widely used approach in functional data analysis is functional regression, which deals with the problem of estimating potentially nonlinear association structures between a set of scalar or functional covariates and a scalar or functional response variable. This section focuses on function-on-scalar regression, where the response variable is functional – i.e., it consists of multiple measurements over some domain like time – and where all covariates are scalar.
The corresponding contribution Bauer et al. (2018) offers guidance on the use of generalized functional additive regression models to estimate such association structures. It centrally builds on the flexibility and robustness of generalized additive models (GAMs) for scalar data. Accordingly, the following section first gives a brief introduction to the basic concepts of the GAM framework before discussing the functional approach, potential alternative techniques and future directions for the research problem.

3.1. Generalized Additive Models

Generalized additive models (Hastie and Tibshirani, 1990) are an established approach to flexibly estimate nonlinear association structures. Estimation is based on a semiparametric framework where nonlinear effects are represented by spline bases and where overfitting is prevented by penalizing overly flexible effects. The following compact overview of the basic structure of GAMs is based on the standard work of Wood (2017). For full details on the following concepts as well as for a complete overview of existing extensions to the framework – including mixed modeling for hierarchical data structures, generalized additive models for location, scale and shape (GAMLSS) for the joint analysis of mean and variance structures, more efficient estimation algorithms, etc. – I also refer the reader to Wood (2017).

Based on a sample of size n with observation index i, a generalized additive model with some scalar response variable Y and a set $\boldsymbol{\mathcal{X}}$ of scalar covariates has the following structure:

$$\begin{aligned} &Y_i|\boldsymbol{\mathcal{X}}_i \sim F(\mu_i, \boldsymbol{\nu}) \\ &g(\mu_i) = \beta_0 + \sum_{r=1}^{R} f_r(\boldsymbol{\mathcal{X}}_{ri}), \qquad i = 1, \dots, n, \end{aligned} \tag{3.1}$$

with $\mu_i = \mathbb{E}(Y_i|\boldsymbol{\mathcal{X}}_i)$ the conditional expected value of the response given the covariates, link function $g(\cdot)$ and an intercept β_0. The conditional response follows some exponential family distribution $F(\cdot)$ with expected value μ_i and dispersion and shape parameters $\boldsymbol{\nu}$. The set $\boldsymbol{\mathcal{X}}_{ri}$ contains the observed value of one covariate for which a linear or nonlinear effect structure $f_r(\cdot)$ should be estimated. Alternatively, $\boldsymbol{\mathcal{X}}_{ri}$ can comprise multiple covariates, in which case the

corresponding effect $f_r(\cdot)$ represents some interaction between these covariates.
Nonlinear effects are represented by some spline basis consisting of J individual basis functions $B_{rj}(\cdot)$. Based on observed values x_{ri} of a single covariate the corresponding nonlinear effect is defined as follows:

$$f_r(x_{ri}) = \sum_{j=1}^{J} \gamma_{rj} \cdot B_{rj}(x_{ri}), \tag{3.2}$$

where each basis function is scaled by a linear parameter γ_{rj} which is estimated in the regression model. Whenever estimating nonlinear effect structures in a GAM-based framework in the contributions to this work, we utilize penalized B-splines (P-splines, Eilers and Marx, 1996) as basis functions. P-spline bases combine the versatility of B-splines with a penalization of the flexibility of each nonlinear effect, and are a robust tool for estimating nonlinear structures in diverse data situations (Eilers and Marx, 2021).
In addition to the above definition of univariate nonlinear effects based on some spline basis, the GAM framework in formula 3.2 can also include multidimensional nonlinear effects $f_r(\boldsymbol{\mathcal{X}}_{ri})$. Such multidimensional effects can be useful in various situations, for example for the estimation of interaction structures or of time-varying nonlinear effects. The corresponding functions $f_r(\cdot)$ are then represented by an appropriate multidimensional spline basis. One established way to define such multidimensional spline bases is the use of a tensor product representation, where a spline basis is created by taking the Kronecker product of multiple marginal, one-dimensional spline bases. A major advantage of this method is its large flexibility, since the marginal bases and penalties can be chosen freely and penalization is done separately for each dimension, allowing for different roughnesses of the various marginal dimensions.

The extent to which nonlinear effects $f_r(\cdot)$ are penalized towards linearity is controlled by penalization parameters λ_r which are estimated as part of the model. Given the vector of all such penalization parameters $\boldsymbol{\lambda}$ and the dispersion and shape parameters $\boldsymbol{\nu}$, the estimation of the joint parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$ is based on the maximization of the penalized log-likelihood

$$\ell_{\text{pen}}(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = \sum_{i=1}^{n} \log f_{\boldsymbol{\theta},\boldsymbol{\nu}}(y_i) - \sum_{r=1}^{R} \lambda_r \cdot \text{pen}(\boldsymbol{\theta}_r), \tag{3.3}$$

with $f_{\boldsymbol{\theta},\boldsymbol{\nu}}(\cdot)$ the respective exponential family density parametrized by $\boldsymbol{\theta}$ and $\boldsymbol{\nu}$, and $\text{pen}(\boldsymbol{\theta}_r)$ an appropriate penalization term for the respective coefficients $\boldsymbol{\theta}_r$ of effect r, e.g. penalizing the second-order differences between the coefficients of consecutive basis functions. To estimate the parameters $\boldsymbol{\theta}$, this penalized likelihood can be optimized utilizing the penalized iteratively reweighted least squares (PIRLS) algorithm. This PIRLS step is iteratively alternated with an optimization step for the penalization parameters $\boldsymbol{\lambda}$, which are for example estimated based on optimizing the generalized cross-validation (GCV) criterion or by optimizing the marginal REML (restricted maximum likelihood) likelihood. If necessary, the estimation of dispersion and shape parameters $\boldsymbol{\nu}$ can be performed jointly with the estimation of $\boldsymbol{\lambda}$ in the optimization of the marginal REML likelihood, or can be based on alternative measures, for example utilizing the Pearson statistic.

3.2. Generalized Functional Additive Models

The framework for generalized additive models was extended by Greven and Scheipl (2017) to handle both functional response variables and functional covariates. Similar to Section 2, we focus on some functional variable $Y(t)$ and, without loss of generality, refer to its functional domain t as *time domain.* For a function-on-scalar regression setting with a functional response and purely scalar covariates, the generalized functional additive model (GFAM) of Greven and Scheipl (2017) has the following form:

$$\begin{aligned}
&Y_i(t)|\boldsymbol{\mathcal{X}}_i \sim F(\mu_{it}, \boldsymbol{\nu}) \\
&g(\mu_i(t)) = \beta_0(t) + \sum_{r=1}^{R} f_r(\boldsymbol{\mathcal{X}}_{ri}, t), \qquad i = 1, \ldots, n,
\end{aligned} \tag{3.4}$$

with $\mu_i(t) = \mathbb{E}(Y_i(t)|\boldsymbol{\mathcal{X}}_i)$ the time-varying conditional expected value of the functional response $Y(t)$, $\beta_0(t)$ the time-dependent intercept curve and $f_r(\boldsymbol{\mathcal{X}}_{ri}, t)$ flexible effect structures which are based on one or multiple of the individual covariates and which are either linear or (multi-dimensionally) nonlinear. The functional response $Y(t)$ is assumed to come from some given distribution $F(\cdot)$ with conditional expectation $\mu_i(t)$ and dispersion and shape parameters $\boldsymbol{\nu}$. We utilize the semiparametric GFAM approach as it is rather flexible in terms of incorporating different types of covariate effects, is applicable for both regular and irregular data with possible missing values, and is accompanied by a flexible software implementation for function-on-scalar regression.

Being a direct extension of the GAM framework, the same principles as outlined in the previous section hold for effect specification and model estimation in the functional case. Nonetheless, some important differences remain between the functional and the scalar modeling approach, mainly because of the additional focus on the functional domain in all aspects of the analysis. In contrast to purely scalar regression models, model evaluation in the functional case should specifically focus on potential autocorrelation of error terms along the functional domain. If functional residuals show substantial intra-curve correlation, uncertainty estimates will be overly optimistic. In such cases, curve-specific functional random intercept terms can be added to the regression model to explicitly represent the structural variation that is present in the residuals.

Regarding the assessment of uncertainty, there exist different types of confidence intervals for nonlinear effect structures as well as confidence or prediction intervals for predictions of the functional response which differ in their coverage properties. Confidence or prediction intervals for curves can either be constructed *globally* (or simultaneously), *pointwise* or *intervalwise*, the interpretation being that the interval overlaps the true process globally, at a specific point or in a specific interval with a given probability, respectively. Depending on the targeted interpretation of the uncertainty intervals, different estimates are available for their computation. In the contribution Bauer et al. (2018) we give an overview of existing approaches to compute these different types of uncertainty intervals and additionally outline a generally applicable bootstrap-based estimation scheme. (Non)parametric bootstrapping can be used to construct all different types of uncertainty intervals. However, bootstrapping can be computationally expensive – often prohibitively so for high-dimensional data or complex models.

The versatility of the semiparametric approach is backed by sophisticated software packages for `R`. In specific, the GFAM framework is implemented in the `refund` package (Goldsmith et al., 2021), which builds on the sophisticated implementation of the estimation of generalized additive models in package `mgcv` (Wood, 2017). In the `FoSIntro` package (Bauer, 2017) that accompanies the contribution Bauer et al. (2018), we implemented some further functionalities for data visualization, model evaluation and the bootstrap-based estimation of confidence and prediction intervals, based on models estimated with function `pffr` from the `refund` package.

3.3. Alternative Approaches

Several alternative approaches exist to perform function-on-scalar regression. One such alternative class of methods, which can deal with functional data that were potentially observed over non-regular grids and with missing values, utilizes a pre-smoothing step prior to model estimation. In doing so, each functional observation is smoothed and the resulting smooth curve is treated as the novel functional observation (e.g. Ramsay and Silverman, 2005). While this can allow for more efficient estimation as of a compact representation of the smooth curve, the method comes with the major disadvantage that the measurement error removed by the smoothing step is not taken into account in subsequent inference.

An overview of nonparametric methods and their applications is provided by Ferraty and Vieu (2006). Respective regression approaches are usually based on kernel methods and are able to model highly nonlinear association structures. However, nonparametric methods are mostly limited to univariate models with a single covariate only. For a more recent survey of nonparametric functional regression approaches and potential future research directions see Ling and Vieu (2018). Very recently, Rao and Reimherr (2021) have introduced a framework that allows to estimate (nonlinear) function-on-function regression models with multiple predictor functions, based on a neural network architecture.

As another alternative, fully Bayesian functional regression can be used. The currently most comprehensive Bayesian framework is the functional mixed model (FMM) framework of Morris (2017) and collaborators, who also provide a comprehensive comparison to the approach of Greven and Scheipl (2017). Generally speaking, fully Bayesian approaches have the advantage that diverse between- and within-function correlation structures can be incorporated into the model in a very flexible way. Also, handling inference is much easier as approximate posterior distributions of all parameters are available in the form of MCMC samples.

The componentwise gradient boosting framework of Brockhaus et al. (2017) is spline-based and extremely versatile. The advantages are most noticeable when working with very high-dimensional data requiring an efficient estimation technique or when dealing with data situations with more parameters than observations, as such settings remain computationally feasible using a boosting approach. However, uncertainty quantification for boosting is currently only possible using computationally expensive resampling techniques like bootstrapping (Hastie et al., 2009).

Based on further groundwork of Brockhaus et al. (2018), the boosting approach has been extended by Stöcker et al. (2021) to the distributional regression setting to simultaneously model location, scale and shape parameters of the distribution of a functional response. Maier et al. (2021) utilize a boosting-based approach to estimate *density-on-scalar* regression which specifically accounts for the properties of a density function as a special case of a functional response. Further, several recent works introduced frameworks for performing functional quantile regression, either utilizing a Bayesian (e.g. Liu et al., 2020) or a frequentist (Beyaztas et al., 2021) approach.

When interest is on multiple functional response variables at once, multivariate functional regression approaches can be used. Different methods were developed for this case. More sophisticated versions are able to model hierarchical data structures and were introduced by Goldsmith and Kitago (2016), Zhu et al. (2017) and more recently Volkmann et al. (2021). The latter introduces a multivariate functional additive mixed model (multiFAMM) framework which is implemented in R package `multifamm` (Volkmann, 2021).

Apart from this latter package that was only published very recently, a sophisticated overview of existing software packages that implement approaches for function-on-scalar regression is given in the contribution Bauer et al. (2018).

3.4. Contribution and Prospects

The contribution Bauer et al. (2018) provides an introduction into the general concepts of function-on-scalar regression. Important practical considerations and best practices are outlined for the most important modeling tasks. The paper is aimed at researchers to use this work as a starting point for applying functional regression models to their own data. The contribution is accompanied by the R package `FoSIntro` (Bauer, 2017) which implements functionalities for data visualization, model evaluation and bootstrap-based uncertainty quantification.

We concentrated on the semiparametric approach of Greven and Scheipl (2017) as this framework is rather flexible in terms of incorporating different types of covariate effects, is applicable for both regular and irregular data with possible missing values, and is accompanied by a flexible implementation of function-on-scalar regression in the **refund** package (Goldsmith et al., 2021). However, important differences regarding practical aspects of the application of the existing function-on-scalar regression frameworks are also outlined. Furthermore, current limitations like the problem of accounting for phase variation and intra-functional correlation are made clear.

Also based on the preceding analyses published in Bauer (2016) and Bauer et al. (2017), the semiparametric function-on-scalar approach showed promising results when applied to the seismic ground velocity curves. Most prominently, in contrast to a purely scalar analysis, the use of a functional data analysis method allowed for a flexible estimation of association structures that potentially vary both over the domain of the respective covariate and the functional domain. Regarding the estimated effects, the strongest association with the observed ground velocities could be found – as expected – with the hypocentral distance of a measurement station. Additionally, among the evaluated parameters describing the physical conditions at the underground fault, our analyses show that the second most relevant parameter with a large impact on the resulting ground velocities in the evaluated setting is the dynamic coefficient of friction.

Given the discussion of alternative function-on-scalar regression approaches in the previous subsection, it is apparent that functional regression is currently a very active field of research. Taking its current state into account, several main directions for future research exist that still require substantial progress.

While the GFAM framework has important benefits compared to alternative approaches, it also has several limitations. As Morris (2017) points out, the underlying semiparametric approach is mainly suited for “relatively smooth functions sampled on coarse or moderately sampled 1D Euclidean domains”. Most importantly, the computational feasibility of the framework can be limited, especially when estimating models on huge datasets with a multitude of nonlinear effects while accounting for remaining intra-curve correlation by including curve-specific random intercept functions.

A recent branch of research is the estimation of functional regression based on deep learning architectures. Such approaches are especially promising to improve how models scale to larger-scale data settings with huge amounts of parameters. As already noted, Rao and Reimherr (2021) introduced a framework to estimate function-on-function regression based on a neural network structure. Rügamer et al. (2021) also recently introduced an approach which combines a neural network architecture with the semiparametric estimation of scalar GAMs for performing distributional regression, implemented in the R package `deepregression`. Combining these two frameworks would be a promising step towards joining the flexibility and robustness of semiparametric functional regression with an architecture that allows for a more efficient model estimation.

Most alternative approaches for function-on-scalar regression are also accompanied by relevant issues that still need future research. Regarding their practical applicability, the vision should be to develop methods which are able to robustly handle sparse, non-Gaussian data on irregular grids. These methods should allow for the estimation of flexible association structures and for accounting for hierarchical data structures. All this should be backed by robust uncertainty estimates and should be accompanied by efficient estimation algorithms and implemented in comprehensive software packages. Potential pre-processing steps like curve smoothing or curve registration should optimally be integrated into a joint estimation scheme with the regression to fully represent the observed variation in the method. While big advances were made in many of the listed fields in the last two decades alone, no approach has yet been developed which fully ticks off all the above boxes.

Regarding the seismic application, two major limitations of the current modeling approach should be addressed by future research. First, phase variation in curves was only partly accounted for in Bauer et al. (2018) by discarding leading zero measurements up to the first relevant ground movement as part of a pre-processing step. To better differentiate amplitude variation and phase variation, functional regression should be applied to aligned curves after the application of a registration step as outlined in Section 2, as well as to the resulting warping functions that comprise the curves' phase variation. This especially suggests itself since both amplitude variation (comprising information on the strength of seismic ground motion) and phase variation (comprising information on the propagation speed of seismic waves) are of central interest to geophysicists. Further, the current regression models implicitly rely on the strict assumption that seismic waves originate from the hypocenter as a fixed point source on the fault. Initial analyses hinted that this assumption does not hold for the evaluated seismic setting (Bauer, 2016), due to a generally large-scale rupture that emits relevant seismic waves from different regions on the fault.

4. APC Analysis

In contrast to the previous two sections on functional data analysis, typically focusing on a single functional domain, the main interest in age-period-cohort (APC) analysis is on its three eponymous dimensions. APC analysis aims to flexibly estimate the developments over these individual dimensions to describe temporal alterations of some process of interest and to deduce potential drivers for the respective changes. Sophisticated techniques both for descriptive and for model-based analysis can help to reach this goal.
The following section outlines the contributions Weigert et al. (2021) and Bauer et al. (2022b) which utilize novel descriptive visualization techniques as well as a state-of-the-art modeling approach to disentangle the temporal association structures. Similar to the function-on-scalar regression approach outlined in the previous section, the model-based approach is based on generalized additive models (GAMs) as a flexible and robust modeling framework. The basic concepts of GAM modeling are outlined in Section 3.1.

4.1. Semiparametric APC Analysis

When trying to differentiate between the individual temporal dimensions, statistical methods have to account for their linear dependency (see e.g. Yang and Land, 2013):

$$cohort = period - age. \tag{4.1}$$

Due to this *identification problem*, a perfect separation of the temporal effects is not possible. Still, different techniques were developed that aim to circumvent this issue. Descriptive approaches typically focus on jointly visualizing all three temporal dimensions, usually building on the concept of Lexis diagrams where age and period groups are depicted in horizontal and vertical direction, respectively, so that individual cohorts are represented along the diagonals (Carstensen, 2007). One drawback of Lexis-based visualizations, however, is that they typically only focus on developments in the mean structure. We adapt the concept of Lexis diagrams in the newly introduced *density matrices* or *ridgeline matrices* to shift the focus from a single statistical moment to the whole distribution of the main variable. As exemplarily visualized in Figure 4.1 for the analysis of travel distances of German travelers, density matrices are able to simultaneously highlight alterations of multiple salient characteristics of a distribution.

When focus is not on the overall distribution, but on a specific characteristic like the mean, the variance or a specific quantile, a heatmap can be used to visualize the developments in this parameter. Similarly to other Lexis-based visualization schemes, however, classical heatmaps depict cohorts along the diagonals in an orthogonal coordinate system and accordingly have one central problem: Compared to the age and period dimensions – depicted along the y-axis and x-axis, respectively – changes over cohorts are hard to grasp since developments along the diagonals are visually underrepresented (Jalal and Burke, 2020). To resolve this issue, Jalal and Burke

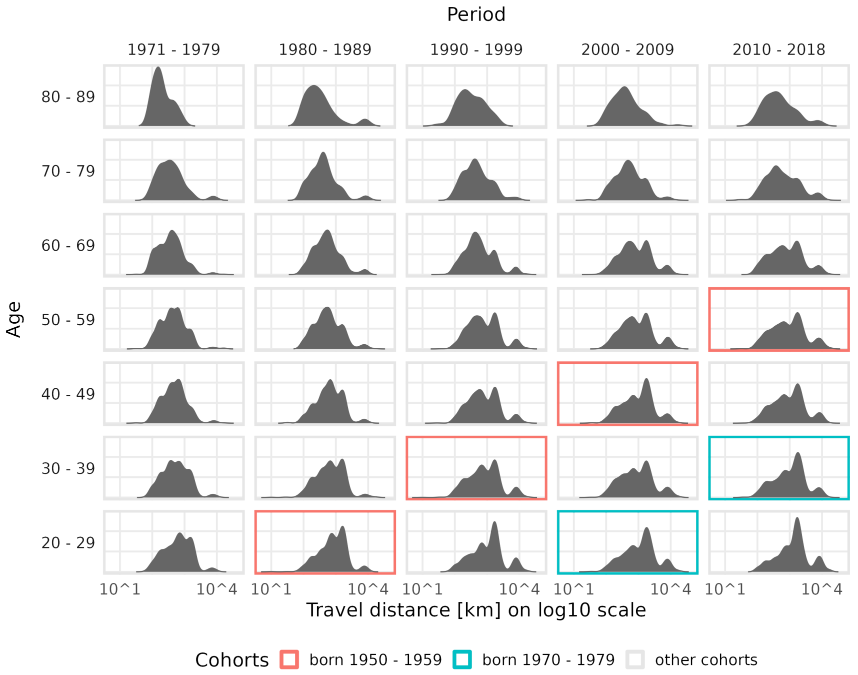

Figure 4.1.: Density matrix of the main trips' travel distance of German tourists in different age and period groups. Two cohort groups are exemplarily highlighted.

(2020) recently introduced *hexamaps* (i.e., hexagonally binned heatmaps) as a novel visualization concept where a heatmap is depicted in a coordinate system with tilted axes to ensure similar visual emphasis on all temporal dimensions. Hexagonal tiles are used instead of rectangular ones since they are a natural fit in the utilized coordinate system which mutually uses 60° angles between each two of its three axes. Similarly to the other visualization approaches outlined in this section, we integrated hexamaps as an alternative to classical heatmaps in the R package `APCtools` described in Bauer et al. (2022b).

To estimate individual association structures with the temporal dimensions, we utilize a regression model-based approach. Weigert et al. (2021) outlines a semiparametric approach based on the estimation of a two-dimensional interaction surface to represent all three temporal dimensions. Following Clements et al. (2005), we estimate a generalized additive model (see Chapter 3.1) with the following structure:

$$g(\mu_i) = \beta_0 + f_{ap}(age_i, period_i) + \eta_i, \qquad i = 1, \ldots, n, \tag{4.2}$$

with observation index i, μ_i the expected value of an exponential family response, link function $g(\cdot)$ and the intercept β_0. The interaction surface $f_{ap}(age_i, period_i)$ is represented by a two-dimensional tensor product spline basis based on two marginal P-spline bases. η_i represents an optional linear predictor that contains further covariates. In contrast to alternative approaches – which are often based on the estimation of linear parameters under specific hard constraints (see Section 4.2) – this semiparametric estimation approach has the main benefit that the linear dependency of the temporal dimensions is dealt with implicitly. Instead of using explicit constraints, a nonlinear, highly flexible approach is used where one temporal dimension (typically cohort) is naturally represented as the interaction of the other two dimensions (typically age and period). On the resulting interaction surface, the age, period and cohort effects are represented along the y-axis, x-axis and the diagonals, respectively.

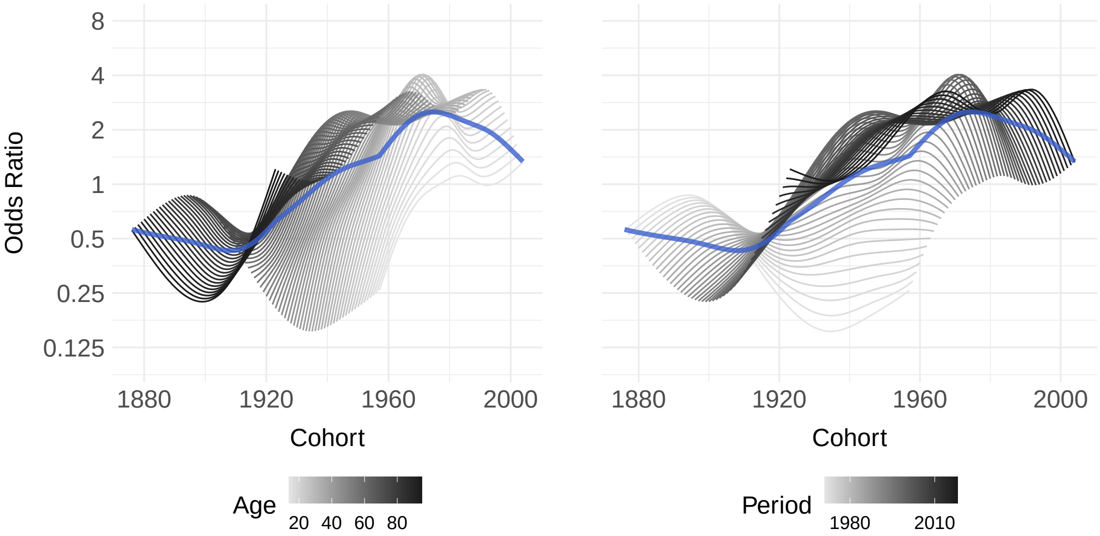

Figure 4.2.: Partial APC plot of estimated odds ratios for the cohort effect dependent on age group (left panel) and period (right), based on a logistic regression model analyzing if people's main trip was a long-distance trip (with trip length $> 6\,000$km) or not. The mean marginal effect is marked as bold blue line.

The estimated interaction surface can again be visualized using classical heatmaps or the outlined hexamaps. Marginal effects of age, period and cohort can be extracted by averaging over individual slices of the surface. Additionally, since perfect separation of the association structures is not possible, APC analyses should always comprise a thorough evaluation of how the temporal interdependencies shape the observed developments. We introduce *partial APC plots* as a novel visualization technique to tackle this issue. In addition to one marginal effect, partial APC plots display appropriate slices of the tensor product surface so that, e.g., one line represents the non-linear variation over cohorts for one specific age group only. In this way, the visualization shows potential interaction structures, for example if differences between cohorts structurally vary when focusing on different age groups. Further, this also highlights how the composition of the underlying sample might affect the estimated marginal effect in different regions of the domain. This can be used to highlight potentially problematic structures in the data, for example when the available data on people from young cohorts only consist of people in young age groups. Such interdependent data structures exacerbate the separation of the temporal effects. Since similar problems with the composition of the data are, however, often inevitable in the studies that require APC analysis, highlighting this information enables a better evaluation of which regions of the marginal effects have to be interpreted with care. An exemplary partial APC plot is visualized in Figure 4.2.

4.2. Alternative Approaches

As noted, descriptive approaches for the analysis of APC data are most commonly based on the concept of Lexis diagrams. Typical visualizations focus on the depiction of average values or rates

of some parameter, similarly to the heatmaps motivated in the previous subsection. Adaptations of this basic technique exist for different data situations, for example when observations are only available on differently coarse categorizations of the temporal dimensions.
Variation over one specific temporal dimension can be visualized in classical line plots, showing, e.g., *age-specific* mortality rates which depict the rates observed in different age groups along the x-axis, potentially drawing separate lines for separate period or cohort groups (Yang and Land, 2013). In a similar manner, it is common to visualize *age-standardized* developments where the temporally varying age distribution in the population is accounted for based on some standardization technique (see e.g. Ahmad et al., 2001). The resulting developments can then be plotted against period or cohort along the x-axis. Combined with faceting, such visualizations are also used to control for or display association structures with other covariates. In the end, however, all standardization techniques implicitly rely on substantial assumptions that, e.g., give specific age groups higher or lower weight in the resulting statistic (Ahmad et al., 2001).

More sophisticated, model-based approaches to differentiate the temporal effects often build on a three-factor regression model which includes a linear effect for every temporal dimension (see e.g. Holford, 1983). For cross-sectional data aggregated over age groups, periods and cohorts, this is a generalized linear model (GLM Nelder and Wedderburn, 1972) of the form

$$g(\mu_{apc}) = \beta_0 + \beta_a \cdot age_a + \beta_p \cdot period_p + \beta_c \cdot cohort_c, \tag{4.3}$$

where μ_{apc} denotes the expected value of an exponential family response for age group $a = 1, \dots, A$, period $p = 1, \dots P$, and cohort $c = 1, \dots, C$, $g(\cdot)$ denotes the link function, β_0 the intercept, and β_j $(j \in \{a, p, c\})$ the linear coefficients. When modeling the absolute number or the relative rate of some event (e.g. death), this model is usually estimated with some logarithmic link function, leading to a multiplicative interpretation of the effects. For such applications, the model framework also allows for the incorporation of an offset term to control for different characteristics or exposure levels in the observational units. Finally, the model structure is easily adaptable to individual data (e.g. Fannon et al., 2018). Panel data can be analyzed by introducing random effects into the model (Diggle et al., 2002; Yang and Land, 2013).
As of the identification problem, the estimation of model 4.3 is only possible when imposing additional constraints in the estimation process. While early methods often used strict linear constraints such as the equality of two of the three effects (e.g. Fienberg and Mason, 1979), modern approaches rely on less restrictive assumptions. Carstensen (2007) estimates nonlinear instead of linear effects for the temporal dimensions and bases the estimation on more subtle constraints for which he also motivates how they still allow for reasonable substantial interpretations of the individual effects. Schmid and Held (2007) use Bayesian hierarchical models in which they restrict first- and second-order differences of the effects. Fu (2000) introduced the *intrinsic estimator*, which utilizes a form of principal components regression to estimate association structures. A thorough overview of existing methodology is given by Yang and Land (2013). An overview of existing software packages that implement routines for APC analysis is given in the contribution Bauer et al. (2022b).

4.3. Contribution and Prospects

Despite substantial methodological progress over the last decades, circumventing the identification problem to separate the effects of age, period, and cohort remains the crucial challenge in APC analysis. In the contributions Weigert et al. (2021) and Bauer et al. (2022b) we utilize a semiparametric regression framework as a flexible estimation approach. Building on the framework of generalized additive models, the respective optimization algorithms are robust and efficiently applicable to a large variety of data settings, including cross-sectional and panel data as well as aggregated and individual-level data. Nonetheless, even if this approach does allow for the proper separation of the temporal dimensions, the interpretation of the intertwined association structures remains challenging, especially when communicating results to practitioners without a thorough knowledge of the statistical concepts.

To tackle this issue, both publications contribute to developing better, more accessible visualizations and summary statistics that communicate the association structures in adequate complexity. Weigert et al. (2021) introduces ridgeline matrices and partial APC plots as novel visualization concepts both for the descriptive analysis of APC structures as well as for a model-based analysis. Bauer et al. (2022b) introduces the R package `APCtools` which implements all methods outlined in Weigert et al. (2021) as well as the hexamap visualization concept introduced by Jalal and Burke (2020) and several convenience functions for performing an APC analysis.

The GAM framework is an adequate basis for estimating APC structures in widespread research settings, also because of available extensions that allow to account for diverse statistical issues. Among others, these include mixed regression models for analyzing panel data (Diggle et al., 2002), more efficient estimation schemes for larger-scale data situations (Wood et al., 2017), as well as approaches for distributional regression to jointly model location, scale and shape parameters (GAMLSS, Rigby and Stasinopoulos, 2005).

Regarding the analysis of destination choice patterns in Weigert et al. (2021), the applied statistical framework proved its worth in contributing to a deeper understanding for explaining alterations in the travel behavior of German travelers over the last decades. Our results confirm that such alterations in travel behavior occur in accordance with life cycle theory (age), macro-level developments in economy and society (period), and generational theory (cohort).

Since perfect separation of the three temporal effects is not possible, and since the outlined semiparametric framework is able to adequately represent the temporal association structures in widespread research settings, future research should focus on two main fields (in addition to further refining the underlying statistical models). First, specific focus should be on the development and refinement of tools to make the interpretation of the complex temporal patterns more accessible. Especially model-based visualization techniques like partial APC plots can provide valuable insights into specific temporal structures. Descriptive APC analyses currently mostly highlight potential alterations in the mean structure of some process. Similarly to ridgeline matrices, existing techniques should be adapted to also specifically highlight changes in the variation structure. As a side note, while we designed ridgeline matrices specifically for cross-sectional data, it remains to be evaluated if they can also be adapted for panel data settings. Second, outdated inferential approaches for APC analysis are still established in many fields of science. To this day, researchers often only evaluate linear effect structures and build on restrictive constraints, thus preventing a reasonable representation of the temporal patterns present in the data. It is also in the duty of statisticians to work towards establishing flexible state-of-the-art approaches in other fields of science.

Future research regarding our tourism science application should follow multiple directions. While our travel distance analysis in Weigert et al. (2021) mainly focuses on specific distance categories (e.g. long-distance trips), more advanced statistical methods such as functional density-on-scalar regression approaches (see Section 3.4 and Maier et al., 2021) could provide further insight into specific aspects of the observed developments. Further, similar analyses should be conducted for other dimensions of individual travel behavior and other touristic source markets to eventually obtain a holistic picture of how tourism has changed over the last decades. Given the ability of the outlined regression approach to include covariates, such analyses should also account for specific characteristics of individual travelers like financial or health constraints that shape their travel decision process. In this way, the modeling framework can be used to estimate temporal changes in travel behavior based on the comparison of individuals with similar initial conditions.

Part II.

Registration of Functional Data

5. Disentangling the Variation Structure of Seismic Ground Velocities – Registration for Incomplete Non-Gaussian Functional Data

Contributing article

Bauer, A., Scheipl, F., Küchenhoff, H., and Gabriel, A.-A. (2022a). Disentangling the Variation Structure of Seismic Ground Velocities – Registration for Incomplete Non-Gaussian Functional Data. *Manuscript submitted for publication.*

Code repository

https://github.com/bauer-alex/IncFunDatRegistration_supp

Author contributions

The idea to tackle the registration of incomplete functional data was jointly brought up by Alexander Bauer and Fabian Scheipl. Alexander Bauer performed the literature review and proposed to base the approach on the exponential family registration approach of Wrobel et al. (2019) and the generalized functional principal component analysis (GFPCA) approach of Gertheiss et al. (2017). Alexander Bauer wrote the initial draft of the paper and the complete codebase for the analyses. Fabian Scheipl was closely involved in all parts of the work, extensively proofread the publication and contributed to make the covariance estimation step of the GFPCA implementation more efficient. Helmut Küchenhoff was involved in initial methodological discussions of the approach and extensively proofread the paper. Alice-Agnes Gabriel created the seismic data, was closely involved in their analysis and substantially contributed to the interpretations in section 5.2.

Disentangling the Variation Structure of Seismic Ground Velocities - Registration for Incomplete Non-Gaussian Functional Data

Alexander Bauer †

Department of Statistics, LMU Munich, Germany.

E-mail: alexander.bauer@stat.uni-muenchen.de

Fabian Scheipl

Department of Statistics, LMU Munich, Germany.

Helmut Küchenhoff

Department of Statistics, LMU Munich, Germany.

Alice-Agnes Gabriel

Department of Earth and Environmental Sciences, LMU Munich, Germany.

Summary. We analyze seismological data comprising spatially indexed, incomplete ground velocity time series with a highly volatile Gamma structure. These functional data are used to assess critical geophysical parameters associated with high ground motion in the event of an earthquake. To separate phase from amplitude variation, functional data are registered, i.e., their observed domains are deformed elastically to align the curves with template functions. Most available registration approaches are limited to complete, densely measured curves with Gaussian noise and cannot handle incomplete curves which are not recorded over their entire domain. We develop a method for joint likelihood-based registration and latent Gaussian process-based generalized functional principal component analysis to handle incomplete curves, provide sophisticated open-source software and compare the approach to existing routines.

Keywords: amplitude variability; curve alignment; functional data analysis; partially observed curves; phase variability; seismology.

1. Introduction

Dealing with phase variability is crucial in functional data analysis. Many different approaches exist for registering curves (see e.g. Marron et al., 2015), but their application to diverse real world data settings remains challenging. Most existing approaches target small to intermediate datasets of completely observed curves with small amounts of Gaussian

†A. Bauer's research was partially funded by the German Research Foundation (DFG) under Grant KU 1359/4-1; F. Scheipl's research was partially funded by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A. The authors of this work take full responsibilities for its content.

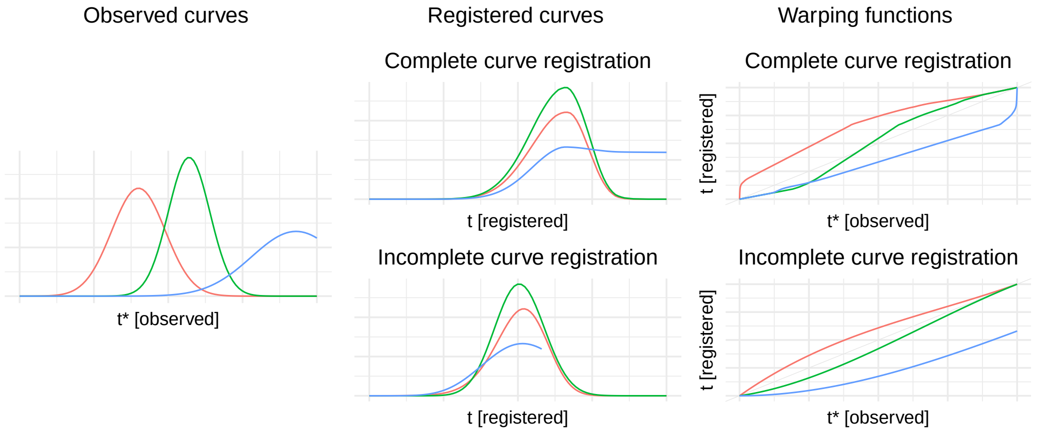

Fig. 1. Simulated observed curves with trailing incompleteness (left column), registered curves only comprising amplitude variation (center), and inverse warping functions visualizing phase variation (right). Registration was performed with the complete curve SRVF approach of Srivastava et al. (2011) with function `time_warping()` from the R package `fdasrvf` (top row; Tucker, 2020) and our incomplete curve approach (bottom row). Note the extreme time dilation around the blue curve's peak in the top row, which yields a highly implausible registered curve.

noise, evaluated on dense, regular grids. Especially the registration of *incomplete* curves, i.e., curves whose underlying process is not observed from its natural starting point all the way to its natural endpoint, has received only limited attention so far (see e.g. Matuk et al., 2019; Bryner and Srivastava, 2021). However, such data arise in many fields. Missing information about the initial development of some processes, i.e. *leading incompleteness*, can be caused by different starting conditions of subjects at the beginning of a study. *Trailing incompleteness* towards the end of the underlying processes is present in experiments and studies with a fixed endpoint that causes right-censoring, or in panel studies with relevant dropout rates. If both types of incompleteness are present, we use the term *full incompleteness*.

— **Notation**
In this functional data setting, we observe discretized incomplete curves $Y_i(t_i^*)$, $i = 1, \dots, N$ over individual *chronological time domains* T_i^*, where the observed grids $\boldsymbol{t}_i^* = [t_{ij}^*]_{j=1,\dots,n_i}$ may be irregular or sparse, and chronological domains $T_i^* = [t^*_{\min,i}, t^*_{\max,i}]$ are defined by the individual observation periods. W.l.o.g. we assume that all observed curves are realizations of stochastic processes over a common underlying *internal time domain* $T = [0,1]$ and $T_i^* \subseteq T \ \forall i$. Registering the curves requires estimating inverse warping functions $h_i^{-1} : T_i^* \mapsto T$ that account for the data's phase variation and map individual chronological times to the internal time of the underlying process. The resulting registered curves $Y_i(t) = Y_i(h_i^{-1}(t_i^*))$ only contain amplitude variation.

Applying conventional "complete curve" registration methods to incomplete curves of-

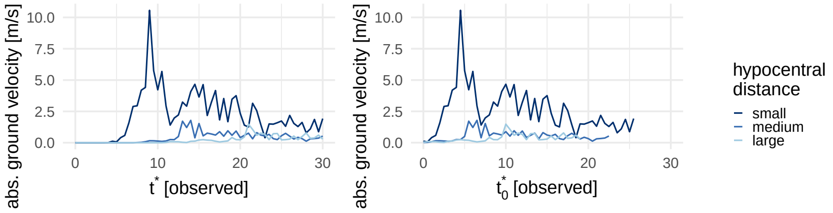

Fig. 2. Typical seismic observations recorded at different hypocentral distances. As a preprocessing step, the leading zero measurements of the raw curves (left pane) are cut off and absolute ground velocities are analyzed on the *time since the first relevant absolute ground velocity measurement* t_0^* (right).

ten leads to nonsensical results (see Figure 1). This is caused by the implicit, unwarranted assumption that the endpoints of the individual observed chronological domains T_i^* are identical to those of the global internal time domain T.

— **Data setting**
Our approach is motivated by synthetic seismic data originating from large-scale numerical *in silico* experiments based on the 1994 magnitude 6.7 earthquake in Northridge (California). The experiments were performed by the Department of Earth and Environmental Sciences (LMU Munich, Germany) using the software SeisSol (Pelties et al., 2014; Uphoff et al., 2017github.com/SeisSol/SeisSol) and are used to assess critical geophysical parameters associated with high ground motion in the event of an earthquake. The simulated ground motion curves contain highly relevant phase variation due to different propagation velocities of the seismic waves and their varying distance to the hypocenter of the earthquake. The complete data comprise 800 000 curves, each recorded over 30 seconds with a frequency of 2Hz. Seismic activity has not subsided after 30 seconds in many cases, so the curves are mostly incomplete. For more information on the data setting see Bauer et al. (2017), Bauer et al. (2018) and Happ et al. (2019). In order to investigate the structure of phase and amplitude variability of these highly variable and spatially indexed data, we register them to spatially varying template functions learnt from the data and represent the registered curves in a lower-dimensional space. Note that these data do not have a simple structure with additive Gaussian noise since absolute ground motion velocities are nonnegative and higher values entail higher variability (see Figure 2).

We also apply our method to a version of the well-known Berkeley child growth study data (Ramsay and Silverman, 2005). The data contain annual measurements of the body heights of 39 boys and 54 girls from ages 1 to 18. The focus lies on the first derivatives of the data, i.e., the speed of growth in different stages of childhood and adolescence. We simulate full incompleteness in these data by drawing an artificial initial age for every child in the first quarter of the time domain as well as an individual drop-out year in the second half of the domain. The observed curves with simulated incompleteness are visualized in Figure 6.

— **Study aim and contributions**
Motivated by the challenges inherent in the seismic data, we require a curve registration method that (i) is able to handle incomplete curves, (ii) is applicable to non-Gaussian data and (iii) includes a lower-dimensional representation of the registered curves. The latter is especially of interest for further analyses of the estimated phase and amplitude variation structure. To achieve these goals, we (i) derive and implement a novel penalized approach for incomplete data registration and (ii, iii) derive and implement extensions of the methods introduced by Wrobel et al. (2019) for exponential family distributions beyond Binomial and Gaussian data. Furthermore, we implement multiple computational improvements in the underlying software stack to accelerate and stabilize the algorithm. Our implementation is available in the `registr` package (Wrobel and Bauer, 2021) for the open-source software R (R Core Team, 2020). All analyses in this paper can be reproduced based on the code and data in our public GitHub repository (Bauer, 2021).

Before introducing our method in Section 3 we give an overview of prior work in Section 2. Sections 4 and 5 comprise an extensive simulation study and the applications. We end with a discussion in Section 6.

2. Related Work

2.1. Registration

Accounting for phase variation is often critical when analyzing functional data. We refer to Marron et al. (2015) and the references therein for an introduction to the general issue and a detailed overview of available registration approaches. Many early approaches focused on the alignment of curves towards given (salient) structures, i.e., landmark registration (e.g. Kneip and Gasser, 1992). More recent proposals mostly perform alignment towards *template functions* that are either based on domain knowledge or estimated based on some measure of centrality, with such estimates often iteratively refined over the course of the registration procedure. Warping functions are commonly estimated purely nonparametrically (e.g. Tucker, 2020; Chakraborty and Panaretos, 2021), as (piecewise) linear functions (e.g. Sangalli et al., 2010; Vitelli, 2019; McDonnell et al., 2021) or in a basis expansion. Common examples for the latter are the use of (penalized) B-spline (Telesca and Inoue, 2008; Wrobel et al., 2019) or Fourier (Mattar et al., 2009) bases or of warplets (Claeskens et al., 2010).

— **SRVF-based registration approaches**
One of the most popular approaches is the square-root velocity function (SRVF) framework introduced by Srivastava et al. (2011), who showed that the warping-invariant Fisher-Rao metric quantifies pure amplitude distances and is equivalent to the simple L_2-metric in the SRVF space. Recently, Guo et al. (2020) have extended this framework to jointly analyze amplitude, phase and spatial variation.

Cheng et al. (2016), Kurtek (2017) and Lu et al. (2017) adapted the SRVF approach to perform registration in a Bayesian setting. Bayesian approaches were also introduced for data settings with stronger noise under informative priors (Matuk et al., 2019; Tucker et al., 2021). These Bayesian approaches can provide a full representation of the joint phase and amplitude uncertainty, but are computationally very demanding. The method of Matuk et al. (2019) handles sparse and fragmented Gaussian functional data where measurements are only available over some parts of the observation domain, but no software implementation was publicly available at the time of writing.

More recently, Nunez et al. (2021) and Chen and Srivastava (2021) introduced the neural network-based frameworks SrvfNet and SrvfRegNet, respectively. Both approaches build on the SRVF framework and enable a highly efficient estimation and prediction of warping functions in large-scale data settings for registering curves to their Karcher mean or fixed and pre-specified template functions.

— **Other registration approaches**

All the above approaches are limited to continuous functional data, mostly under the assumption of Gaussian errors. Some extensions to binary functional data such as Wu and Srivastava (2014) and Panaretos and Zemel (2016) rely on a pre-smoothing step to obtain a continuous representation of the curves. The *congealing* approach of Learned-Miller (2005) (adapted to functional data by Mattar et al., 2009), which iteratively optimizes measures of alignment like the integrated point-wise differential entropy via gradient descent, is applicable in diverse data situations and computationally efficient. Wrobel et al. (2019) utilizes a likelihood-based optimization approach for exponential family data which is able to practically handle moderate-to-large scale datasets (Wrobel et al., 2021).

— **Joint registration and low-rank representations**

For analyzing the main modes of amplitude and phase variation, it is common to estimate a low-rank representation of the registered curves or, at least, the template functions that serve as registration targets, and potentially also of the estimated warping functions. Tucker et al. (2013), Hadjipantelis et al. (2015), Lee and Jung (2016) and Happ et al. (2019) use (joint) functional principal component analysis (FPCA) for finding compact representations of both phase and amplitude modes of variation.

Tucker (2014) (utilizing the SRVF approach of Srivastava et al., 2011, for registration), Wagner and Kneip (2019) and Kneip and Ramsay (2008) (optimizing a least squares criterion) and Wrobel et al. (2019) (optimizing an exponential family likelihood) all utilize iterative algorithms to successively refine warping functions that lead to registered curves whose amplitude variation can be represented in terms of a low-rank FPC basis.

— **Incomplete curve registration**

Comparatively few approaches have been developed so far for the registration of incomplete curves. Some heuristic approaches are available in the field of dynamic time warping (DTW) for time series analysis. Subsequence DTW offers a framework to find a subsequence of a (fully observed) template curve to which a partially observed curve can be matched (see Müller, 2015, 7.2). Tormene et al. (2009) introduce an algorithm for "open-begin and open-end" DTW (OBE-DTW) that is also able to handle full incompleteness.

More sophisticated registration approaches were introduced only recently. Sangalli et al. (2010) and Vitelli (2019) make use of linear warping functions with free starting points and endpoints. The observed curve domains are constrained so that they dilate or extend the domain by a factor $0.9 - 1.1$ to ensure reasonable warpings. As noted above, Matuk et al. (2019) allows to analyze fragmented functional data, but their approach is feasible only for small to intermediate sets of Gaussian data.

Bryner and Srivastava (2021) introduced an approach for *elastic partial matching* to tackle trailing incompleteness. Before registering each curve to its template using the complete curve SRVF approach of Srivastava et al. (2011) they estimate the time scaling necessary to (partially) match the observed domain of a specific curve to the domain of the template curve and perform the registration only on the intersection of the curves' domains. Both steps are combined in a joint, gradient-based algorithm. At the time of writing, no implementation of their method was available on request.

— **Software implementations**

Multiple packages for the statistical open-source software R (R Core Team, 2020) exist that implement registration approaches. Basic approaches outlined in Ramsay and Silverman (2005) are implemented in package `fda` (Ramsay et al., 2020). The OBE-DTW approach of Tormene et al. (2009) is available in `dtw` (Giorgino, 2009). Package `fdasrvf` (Tucker, 2020) implements the SRVF registration of Srivastava et al. (2011) and the iterative procedure of Tucker (2014) for finding overall similar registered curves with a low-rank FPCA representation. Code for Wagner and Kneip (2019) is available on GitHub (Wagner, 2020). R package `registr` (Wrobel and Bauer, 2021) implements the likelihood-based approach of Wrobel et al. (2019) and the incomplete curve extensions presented in this work for various exponential family distributions.

2.2. Generalized Functional Principal Component Analysis

Functional principal component analysis (FPCA) is a technique to analyze and represent functional data in terms of their main modes of variation (Ramsay and Silverman, 2005). To purely represent amplitude variation, FPCA is most commonly applied to curves without phase variation, potentially after an initial registration step. As noted above, the concept of FPCA was also extended to separately (Tucker et al., 2013) or simultaneously (Happ et al., 2019) analyze amplitude and phase variation. Multiple approaches exist for FPCA on sparse or partially observed functional data, but mostly assume Gaussianity (c.f. Stefanucci et al., 2018). Adaptations to the non-Gaussian case for performing generalized FPCA (GFPCA) do exist, but have to be assessed with care since marginal estimation of the overall mean can introduce bias (Gertheiss et al., 2017).

— **Probabilistic GFPCA**

A popular method for multivariate non-Gaussian data is *probabilistic FPCA* (Tipping and Bishop, 1999), based on likelihood optimization. For the case of Gaussian functional data, James et al. (2000) and Rice and Wu (2001) use a similar approach based on mixed-effects regression. Zhou et al. (2008) adapted these methods in a paired-curve setting. Huang et al. (2014) extended the ideas of James et al. (2000) and Zhou et al. (2008) to non-Gaussian functional data and combined them with a clustering approach. Recently, Wrobel et al. (2019) further adapted the mixed model-based approach by introducing a link function and estimating the GFPCA based on computationally efficient variational approximations. To the best of our knowledge, efficient approximations are available only for Gaussian and binary data and are not directly adaptable to further exponential family distributions.

Bayesian adaptations of probabilistic FPCA to non-Gaussian settings were introduced by van der Linde (2009) for binary and count data, and Goldsmith et al. (2015) with an extension to multilevel data. While these Bayesian approaches can provide a full representation of the underlying uncertainty and show good performance in sparse data situations (c.f. Gertheiss et al., 2017), they are computationally demanding.

— **Two-step GFPCA**
A nonparametric approach to Gaussian FPCA was introduced by Yao et al. (2005) and adapted by Hall et al. (2008) for the non-Gaussian case. Serban et al. (2013) extended this method to further handle multilevel binary data with potentially rare events. Li and Guan (2014) used a similar approach to model point processes with a spatio-temporal correlation structure. Gertheiss et al. (2017) showed that the marginal mean estimates proposed by Hall et al. (2008) can introduce bias in non-Gaussian data settings and tackled this issue by plugging the eigenfunction estimates into a generalized additive mixed model to achieve an estimation of the mean structure conditional on FPC scores represented as random effects.

— **Some notes on consistency**
Consistent estimators for the covariance operator are crucial for FPCA. The consistency and convergence rates of estimators for covariance operators were thoroughly studied for differently dense data settings (c.f. Wang et al., 2016; Cao et al., 2016), their properties in the presence of stronger, potentially non-Gaussian noise, however, remain an area of active research. Standard techniques for covariance estimation quickly become computationally infeasible in high-dimensional (Li et al., 2020) or irregular (Cederbaum et al., 2018) data settings and algorithmic innovations are required. Recently, Sarkar and Panaretos (2021) introduced promising neural network architectures for the efficient, nonparametric approximation of (multidimensional) covariance operators and their eigen-decomposition.

Several studies evaluated the consistency of covariance and FPCA estimators for incomplete curve settings. When incompleteness originates from a missing completely at random (MCAR) process and measurements are dense, established estimators for the mean, the covariance and for eigenfunctions and eigenvalues are consistent (Kraus, 2015). For the subject-specific functional principal component (FPC) scores, Kraus (2015) introduced a consistent estimator. His comparison to the PACE approach of Yao et al. (2005) indicates that the bias of comparable conditional methods for estimating the FPC scores is likely to be small.

Substantial bias can be caused by systematic missingness in the data. Liebl and Rameseder (2019) review certain violations of the MCAR assumption and motivate novel estimators for the mean and covariance structure for dense incomplete curve settings. While the classical estimators for the mean and covariance are consistent in regions of the domain with (virtually) no missingness, they are prone to (severe) bias the stronger the violation from the MCAR assumption and the fewer observations are available.

Estimation accuracy is also crucially affected by the coverage of the overall domain, especially so for estimating the covariance operator. Only if a sufficient number of observed curves overlap on the respective parts of their observed domains can the corresponding regions of the covariance surface be estimated reliably. For the setting of short observed domains (”functional fragments”), Delaigle et al. (2020), Descary and Panaretos (2019) and Zhang and Chen (2017) introduce conditions and approaches for consistently estimating (parts of) the covariance surface.

— **Software implementations**
Some general FPCA methods are implemented in R packages `fda` (Ramsay et al., 2020) and `refund` (Goldsmith et al., 2020). The multivariate FPCA approach of Happ et al. (2019) is implemented in package `MFPCA` (Happ-Kurz, 2020); the PACE algorithm of Yao et al. (2005) in `fdapace` (Carroll et al., 2020). The methods outlined in Gertheiss et al. (2017) are available for the binary curves setting in `gfpca` (Goldsmith, 2016), where the mixed regression in the two-step approach is estimated with package `gamm4` (Wood and Scheipl, 2020). Our accompanying package `registr` (Wrobel and Bauer, 2021) implements the Gaussian and binary curve GFPCA of Wrobel et al. (2019) as well as the two-step approach of Gertheiss et al. (2017) for various exponential family distributions.

3. Methods

As outlined, curves can have missing information at the beginning of their domain (i.e., *leading incompleteness*), at the end of their domain (*trailing incompleteness*), or both (*full incompleteness*). Our approach is able to handle all three types of incompleteness for curves observed on potentially irregular individual grids of evaluation points, without assuming Gaussianity of the observed data.

We first introduce our registration approach and the approach for generalized FPCA in full detail, and then present the main iterative algorithm to obtain a solution where the registered curves are well represented by a low-rank GFPCA basis. Potential identifiability issues and practical implications are discussed at the end of this section. Computational details are given in Appendix A1.

3.1. Registration for incomplete curves

We extend the likelihood-based framework for registering complete curves from exponential family distributions of Wrobel et al. (2019). In the registration step, the individual *chronological time domains* T_i^* are mapped onto the registered *internal time domain* T. This is achieved by estimating inverse warping functions h_i^{-1} that deform an unregistered curve $Y_i(t_i^*)$ toward a suitable template function $\mu_i(t)$ so that

$$\begin{aligned} \mathsf{E}\left[Y_i\left(h_i^{-1}(t_i^*)\right)|h_i^{-1}\right] &= \mu_i\left(t\right), \\ \text{with} \quad h_i^{-1}(t_i^*) &= \boldsymbol{\Theta}_h(t_i^*)\boldsymbol{\beta}_i, \end{aligned} \tag{1}$$

with $Y_i(t)$ the registered curve. The inverse warping functions are represented through a B-spline basis with design matrix $\boldsymbol{\Theta}_h \in \mathbb{R}_{D_i \times K_h}$, K_h basis functions and a corresponding coefficient vector $\boldsymbol{\beta}_i$.

Given some distribution from the exponential family this yields the log-likelihood for curve i:

$$\ell\left(h_i^{-1}|y_i,\mu_i\right) = \log\left(\prod_{j=1}^{D_i} f_{i,j}\left[y_i(t_{i,j}^*)\right]\right), \tag{2}$$

with $f_{i,j}(\cdot)$ the corresponding density with expected value $\mu_i\left(h_i^{-1}(t^*_{i,j})\right)$ and observed vectors of function evaluations $y_i(t^*_{ij})$, $j = 1, \ldots, D_i$. We impose working assumptions of mutual conditional independence across functions $[Y_i \perp Y_{i'}]\,|\mu_i, \mu_{i'}$ as well as within functions $[Y_i(t_{ij}) \perp Y_i(t_{ik})]\,|\mu_i$.

— **Constrained optimization**
Warping functions must follow certain constraints so that they yield reasonable transformations of the time domain. First, all warping functions have to be strictly increasing to preserve the temporal order of a curve's measurements. Second, warping functions have to be domain-preserving with regard to the maximal domain of the underlying process. We ensure both by using a constrained optimization algorithm for the warping functions' basis coefficients (see Appendix A1).

— **Circumventing the constraint of fixed time intervals**
If all curves are observed over an identical time interval, domain preservation requires that all warping functions map t^*_{min} and t^*_{max} to themselves so that they begin and end on the diagonal line. This assumption is made in most currently available registration procedures, based on an implicit assumption that the process of interest (e.g., the growth process of children) was observed from its very beginning up to its very end for all subjects. For incomplete curves, however, forcing observed and registered domain lengths to be identical is clearly unsuitable. We drop these hard constraints on the warping functions' basis coefficients and allow warping functions to start and / or end at any point inside the overall time domain. To avoid large deformations of the time domain that are not strongly supported by the data, we penalize the total amount by which the registration changes the duration of the (observed) time domain. In a setting with full incompleteness, we use the following penalized log-likelihood for the registration step:

$$\begin{aligned} \ell_{\text{pen}}\left(h_i^{-1}|y_i, \mu_i\right) &= \ell\left(h_i^{-1}|y_i, \mu_i\right) - \lambda \cdot n_i \cdot \text{pen}\left(h_i^{-1}\right), \\ \text{with} \qquad \text{pen}\left(h_i^{-1}\right) &= \left(\left[h_i^{-1}(t^*_{\text{max},i}) - h_i^{-1}(t^*_{\text{min},i})\right] - \left[t^*_{\text{max},i} - t^*_{\text{min},i}\right]\right)^2. \end{aligned} \tag{3}$$

For leading incompleteness with $h_i^{-1}(t^*_{\text{max},i}) = t^*_{\text{max},i}\ \forall\, i$, this simplifies to $\text{pen}\left(h_i^{-1}\right) = \left[h_i^{-1}(t^*_{\text{min},i}) - t^*_{\text{min},i}\right]^2$ and for trailing incompleteness with $h_i^{-1}(t^*_{\text{min},i}) = t^*_{\text{min},i}\ \forall\, i$ to $\text{pen}\left(h_i^{-1}\right) = \left[h_i^{-1}(t^*_{\text{max},i}) - t^*_{\text{max},i}\right]^2$. In other words, the penalty for one-sided incompleteness represents the squared distance of the respective endpoint of h_i^{-1} to the diagonal. The penalization parameter λ controls how much overall time dilation or compression the registration can perform and is scaled by the number of measurements n_i of curve i to ensure that the impact of the penalization relative to the likelihood is not affected by the number of measurement points per function. Details on the choice of λ are given in Section 3.4.

3.2. Generalized Functional PCA for incomplete curves

We adapt the *two-step approach* of Gertheiss et al. (2017) to estimate a low-rank GFPCA representation of the registered curves $Y_i\,(t) = Y_i\left(h_i^{-1}(t^*_i)\right)$. Following Hall et al. (2008) and the groundwork of Yao et al. (2005), functional principal components (FPCs) are estimated using a marginal, semiparametric method based on assuming a latent Gaussian process $X_i(t)$ so that

$$\begin{aligned} \mathsf{E}[Y_i(t)] = \mu_i(t) = g[X_i(t)], \\ X_i(t) \approx \alpha(t) + \sum_{k=1}^{K} c_{i,k} \cdot \psi_k(t), \end{aligned} \tag{4}$$

where each observed $Y_i(t)$ corresponds to the transformation of the latent process $X_i(t)$ with some fixed response function $g(\cdot)$, and the latent process itself can be decomposed into a smooth global mean $\alpha(t)$, FPCs $\psi_k(t)$ with respective eigenvalues $\tau_k > 0$, and FPC scores $c_{i,k} \sim N(0, \tau_k)$. With known $\psi_k(t)$, model (4) is a generalized functional additive mixed model along the lines of Scheipl et al. (2016a) with a smooth conditional latent mean function in a P-spline representation, functional random effects in an FPC basis representation, and functional random effect scores $c_{i,k} \sim N(0, \tau_k)$. While the derivation of the covariance approximation below assumes that $X_i(t)$ shows only small variation around its mean, stronger variation only has "a modest effect on the errors in individual predictions" (Hall et al., 2008, 4.2).

To derive the GFPCA solution, we first center the observed $Y_i(t)$ based on their marginal mean $\mu_Y(t) = \mathsf{E}[Y_i(t)]$, estimated through a simple smoother of all $(t_{ij}, Y_i(t_{ij}))$-pairs via a generalized additive model (GAM, Fahrmeir et al., 2013) with response function $g(\cdot)$ and the appropriate exponential family for the response. The covariance of the latent process can then be approximated by

$$\widehat{\mathsf{Cov}}\left[X_i(s), X_i(t)\right] \approx \frac{\hat{\sigma}_Y(s,t)}{g^{(1)}[\mu_X(s)] \cdot g^{(1)}[\mu_X(t)]}, \tag{5}$$

with $\sigma_Y(s,t) = \mathsf{E}[Y_{c,i}(s) \cdot Y_{c,i}(t)]$ based on the centered curves $Y_{c,i}(t)$, the marginal mean $\mu_X(t)$ estimated accordingly to $\mu_Y(t)$ and $g^{(1)}(\cdot)$ the first derivative of the response function. For given time points s_1 and s_2, $\sigma_Y(s_1, s_2)$ is estimated as the mean of all pairwise products $y_{c,i}(s_1) \cdot y_{c,i}(s_2)$. The estimated surface $\hat{\sigma}_Y(s,t)$ is a smoothed version of $\sigma_Y(s,t)$ using a bivariate tensor product P-spline basis (Fahrmeir et al., 2013). Since the surface is expected to show some discontinuity along the diagonal this smoothing step is performed under exclusion of the diagonal elements (c.f. Yao et al., 2005). The FPCs $\psi_k(t)$ and their respective eigenvalues τ_k are then estimated from the spectral decomposition of $\widehat{\mathsf{Cov}}\left[X_i(t), X_i(s)\right]$. Our approach deviates slightly from the method of Hall et al. (2008) it is based on: We mean-center the data before taking their crossproducts instead of subtracting the crossproduct of the estimated mean from the crossproducts of the data. In our experience, this yields smoother estimates of the covariance surface which are more amenable to a low-rank FPC representation.

3.3. Joint approach

We utilize the iterative algorithm of Wrobel et al. (2019) to combine the outlined approaches for registration and GFPCA. Our aims are twofold: (i) register all observed curves $Y_i(t_i^*)$ to suitable template functions and (ii) adequately represent the registered curves

$Y_i(t) = Y_i(h_i^{-1}(t_i^*))$ through a low-rank GFPCA basis. We solve this problem by alternating the registration step (conditional on the current GFPCA representations $\mu_i(t)$) and the GFPCA step (conditional on the current estimates of the warping functions h_i^{-1}). The initial registration step is performed with respect to a fixed common template function $\mu(t)^{[0]}$ which has to be set by the user. Subsequent iterations then use the low-rank GFPCA representations $\mu_i(t)$ as curve-specific template functions. Full details on the iterative estimation are given in Algorithm 1.

The number of FPCs in each iteration can be chosen based on the explained proportion of variance. We adapt this criterion to account for peculiarities of the covariance structure estimated with the two-step approach. Full details are discussed at the end of Section 3.4.

Algorithm 1 Joint Registration & GFPCA

```
Require: Observed curves $y_i(\boldsymbol{t}_i^*)$; starting template $\mu(t)^{[0]}$; explained share of variance $\kappa_{\text{var}}$
    of GFPCA solution; convergence tolerance $\Delta_h$, iteration counter $q = 0$.
Initial registration of observed curves $y_i(\boldsymbol{t}_i^*)$ to global initial template $\mu(t)^{[0]}$ to initialize
    $\hat{h}_i^{-1}(t^\star)^{[0]}$;
while $\sum_{i=1}^{N}\left(\sum_{j=1}^{D_i}\left[\hat{h}_i^{-1}(t^*_{i,j})^{[q]} - \hat{h}_i^{-1}(t^*_{i,j})^{[q-1]}\right]^2\right) > \Delta_h$ do
    $q \to q + 1$
    Update GFPCA using registered curves $y_i\left(\hat{h}_i^{-1}(\boldsymbol{t}_i^*)^{[q-1]}\right)$ (Section 3.2).
    Re-estimate GFPCA representations $\mu_i(t)^{[q]}$ based on the first $K^{[q]}$ FPCs that explain
        at least a share $\kappa_{\text{var}}$ of the total variance (Model (4))
    Update warping function estimates $\hat{h}_i^{-1}(t^\star)^{[q]}$ by re-registering observed curves $y_i(\boldsymbol{t}_i^*)$
        to $\mu_i(t)^{[q]}$.
end while
Final GFPCA estimation based on the registered curves $y_i\left(\hat{h}_i^{-1}(\boldsymbol{t}_i^*)^{[q]}\right)$
    to obtain GFPCA representations $\mu_i(t)$ based on the first $K$ FPCs that explain at least
    share $\kappa_{\text{var}}$ of the total variance.
```

3.4. Pitfalls and Practical Considerations

— **Identifiability**
A common issue in the separation of amplitude and phase variation is that disentangling the two types of variation is an ill-posed problem in most realistic settings. Structured variability in the curves can almost always be attributed to either warpings of the time domain or superpositions of principal components, or any combination of the two. Both Wagner and Kneip (2019) and Chakraborty and Panaretos (2021) have shown that the general registration problem has a unique solution only if the amplitude variation is of rank 1, i.e. for FPC rank $K = 1$. In practice, this non-identifiability can be removed by introducing suitable inductive biases for estimates of the warping and template functions through priors, penalties and/or limiting the expressivity of model components, e.g. by choosing

low-rank basis representations. We assess the severity of this identifiability problem for our method in a simulation study in Section 4. Note that the low-rank basis representations of the warping functions we employ also seem to successfully avoid the “pinching” problem (see e.g. Ramsay and Li, 1998, 4.2).

— **Choice of the template function**
As outlined above, the template function $\mu(t)^{[0]}$ for the initial registration step in the joint estimation has to be set by the user. The choice should be based on subject knowledge and can be crucial for obtaining reasonable results (compare Appendix A7) and quick convergence in subsequent iterations.

— **Choice of the penalization parameter λ**
Our registration approach controls the overall amount of compression or dilation through the penalization parameter λ. The choice of λ should be based on substantive knowledge so that estimated warping functions represent realistic accelerations and/or decelerations of the observed processes.

— **Choice of the number of FPCs**
The number of FPCs can be chosen based on the explained share of variance of the low-dimensional FPC basis. In this regard, the two-step approach faces two issues: First, since the spectral decomposition is applied to a smoothed covariance surface and not the raw covariance of the data itself, the “explained” share of variance is relative to this “structured” part of the total observed variance. Second, based on our practical experience, spectral decompositions of covariance surfaces often yield a large number of subordinate FPCs which each explain only a very small amount of overall variation, but jointly explain a relevant share. As we show in Appendix A6, it can be argued that these subordinate FPCs often represent phase variation rather than amplitude variation.

We deliberately avoid including such subordinate FPCs in the FPCA solution since the FPCs in the joint approach should only represent the main amplitude variation. Our goal is to find suitable template functions to register against which don’t include modes of phase variation, i.e., the template functions do not need to represent each individual registered curve with high fidelity. Accordingly, we suggest a two-fold criterion for choosing the number of FPCs based on our two-step approach: Choose as many FPCs as are needed to explain a large portion (90%, by default) of the overall structured variation. However, do not include such FPCs in the final solution that account for very little variation (less than 2%, by default). In this way, we define a low-rank FPCA representation for the template functions which might explain less than 90% of the overall variation but does not include a multitude of subordinate modes of (phase) variation.

4. Simulation Study

To assess the performance of our method and compare it to other established approaches we perform a simulation study on both Gaussian and Gamma data, motivated by our seismic application. We focus on the comparison of approaches that jointly perform registration and FPCA and assess (i) their ability to recover de-noised underlying curves, (ii) their performance in disentangling and estimating the underlying amplitude and phase variation, and (iii) their computational efficiency.

We compare our proposal (called “FGAMM” in the following) – combining two-step FPCA with our (in)complete curve registration – with the earlier approach of Wrobel et al. (2019) (“varEM”) – using an identical registration approach combined with a variational EM-based FPCA – and the joint SRVF approach of Tucker (2014) (Algorithm 4.1, “SRVF”) which combines the SRVF registration of Srivastava et al. (2011) with the vertical fPCA introduced in Tucker et al. (2013). The latter approach is only applied to complete curve settings since the software implementation available at the time of writing (Tucker, 2020, R-package `fdasrvf`) is not able to handle incomplete curves.

4.1. Simulation design

In each simulation setting, we first simulate $N = 100$ complete curves on a regular time grid on $[0, 1]$ with length $D_i = 50 \; \forall i$ from model (4) with FPC rank $K \in \{1, 3, 4\}$, with

- mean function $\alpha(t)$ a Gaussian density function with $\mu = 0.45$ and $\sigma = 0.2$,
- eigenfunctions $\psi_k(t)$ as the $(k + 1)$th orthonormal polynomial on $[0, 1]$,
- mutually independent FPC scores $c_{i,k} \sim N(0, \tau_k)$ and $\boldsymbol{\tau} = 1$ for $K = 1$, $\boldsymbol{\tau} = (0.7, 0.25, 0.05)$ for $K = 3$, $\boldsymbol{\tau} = (0.4, 0.3, 0.2, 0.1)$ for $K = 4$,
- Gaussian setting: $Y_i(t_j) \sim N\left(X_i(t_j), \sigma^2 = 0.03\right)$,
- Gamma setting: $Y_i(t_j) \sim \Gamma\left(k = 5, \theta = \frac{1}{5} \exp\left(X_i(t_j)\right)\right)$,

with $X_i(t_j) = \alpha(t) + \sum^K \psi_k(t)c_{i,k}$ the simulated (latent) process.

Warping functions are simulated utilizing a B-spline basis using cubic splines and three degrees of freedom. Their basis coefficients are drawn from a uniform distribution over $[0, 1]$ and cumulatively summed up to ensure monotony. Three settings of (in)completeness are analyzed: Complete curves, weak incompleteness and strong incompleteness. The latter two settings only comprise trailing incompleteness. Weak incompleteness and strong incompleteness are simulated by randomly drawing a cut-off time from a uniform distribution over the last 40% and 70% of the time domain, respectively.

Regarding the correlation structure between the extents of (i) amplitude variation, (ii) phase variation and (iii) incompleteness we analyze three different settings. In the first, the three dimensions are mutually uncorrelated. The second setting comprises a strong negative correlation between amplitude and phase variation, shifting the peaks of curves with larger amplitudes towards the beginning of the domain. The third setting comprises a stronger positive correlation between amplitude and the amount of incompleteness, resulting in stronger incompleteness for curves with lower amplitudes.

Visualizations of the simulated data can be found in Appendix A3.1. For the methods FGAMM and varEM, we use eight and four basis functions for the estimation of the mean curve and the inverse warping functions, respectively. In the Gaussian setting, the penalization parameter is set to $\lambda = 0.025$. In the Gamma setting, it is set to 1 and 0.5 for FGAMM (assuming a Gamma distribution) and varEM (assuming a Gaussian distribution), respectively. The observed overall mean curve is used as the initial template function.

In the FGAMM approach, the covariance surface is smoothed with ten marginal P-spline basis functions. Since the implementation of the SRVF approach relies on the curves being observed on a regular grid, the simulated curves are linearly interpolated onto a regular grid. We perform 100 replications for each simulation setting and method. The following results only cover the simulation settings without correlation between phase, amplitude and incompleteness. Unless noted otherwise, the results for the other simulation settings are structurally similar (see Appendices A3.2 and A3.3).

— **Adaptive estimation of the number of FPCs**
The number of estimated FPCs was pre-set to the respective true simulated amplitude rank. While our method includes adaptive, data-based estimation of the number of FPCs (see Algorithm 1), we did not pursue this approach here since this would jeopardize our ability to differentiate (i) its ability to recover FPCs and their scores accurately and (ii) its ability to select a suitable number of FPCs based on the data. Additional results based on the more realistic use-case with adaptive estimation of the FPCs are given in Appendices A3.4 and A3.5. The methods' performances on the Gaussian simulation settings with adaptively estimated FPC rank K are structurally similar to the ones with pre-specified rank. In the Gamma settings, while this is the case for the estimated phase components, all methods struggle to recover the correct number of FPCs and specifically the varEM approach performs worse in terms of the estimation of amplitude variation.

4.2. Results

— **Performance metrics**
We base our method comparisons in Figures 3 and 4 on different performance metrics, most based on the mean (integrated) squared error (MISE, MSE) for functional and scalar estimates, respectively. Overall performance is quantified using the difference between the simulated individual mean structures (before adding random noise) and the respective representations based on the final FPCA solution (measure MISE_y). This metric indicates how well the complete structured variation, i.e, the combined phase and amplitude variation, of the observed data is recovered. The performance regarding the separation and estimation of amplitude and phase variation is quantified by (i) comparing the spans of the true and estimated FPC bases with a measure introduced by Larsson and Villani (2001) and adapted by Scheipl et al. (2016b) (amplitude variation, LV_ψ) and by (ii) comparing the true and estimated warping functions (phase variation, MISE_h). Following Scheipl et al. (2016b), the measure LV_ψ quantifies the overlap of the spans of two matrices $\boldsymbol{A} \in \mathbb{R}^{n \times p_A}$ and $\boldsymbol{B} \in \mathbb{R}^{n \times p_B}$, $n > p_A, p_B$:

$$\text{LV}_\psi(\boldsymbol{A}, \boldsymbol{B}) = \frac{1}{p_A} \cdot \text{trace}\left(\boldsymbol{V}_B^T \boldsymbol{V}_A \boldsymbol{V}_A^T \boldsymbol{V}_B\right),$$

with $\boldsymbol{V}_Z$, $Z \in \{A, B\}$, a matrix of the left singular vectors of matrix $\boldsymbol{Z}$. We scale the measure by the dimension of the true FPC basis p_A to obtain a codomain of $[0, 1]$ where value 1 encodes perfect representation of the true amplitude variation space and 0 represents completely orthogonal spans. In accordance with the other performance measures we report $1 - \text{LV}_\psi$ so that smaller values encode better performance. Note that LV_ψ cannot be computed for the SRVF approach since that method is based on an FPCA of the SRVF transforms of the original functions and does not yield orthonormal eigenfunctions in the original function space. Finally, we compute the estimation performance of the overall amount of time dilation or compression by comparing the true and estimated domain lengths of the registered curves (MSE$_d$).

— **Results Gaussian settings**
The results for the Gaussian settings are visualized in Figure 3. While methods varEM and FGAMM do a good job in representing the underlying structured variation (MISE$_y$) and in estimating both warping functions (MISE$_h$) and original domain lengths (MSE$_d$), amplitude variation (LV$_\psi$) is only estimated with higher accuracy for amplitude rank 1. Both FPCs and warping functions are estimated more accurately if amplitude variation has smaller rank.

Comparing the methods and focusing only on the complete curve settings (left panels) for which it is applicable, the joint SRVF approach performs consistently worse than the other two approaches. For MISE$_y$ and MISE$_h$ the median performance of FGAMM for amplitude rank 2–3 is better by 89% and 79% compared to SRVF, respectively. The varEM approach performs slightly better than FGAMM for the complete curve settings in terms of representing the overall variation, and slightly worse in terms of recovering the space of amplitude variation. Regarding the incomplete curve settings, the incomplete curve approaches perform consistently best with respect to MSE$_d$ and MISE$_y$. The estimated curve representations contain a much higher share of the originally observed variation than is represented by methods with assumed completeness. While the incomplete curve approaches mostly perform better in terms of phase variation (MISE$_h$ and MSE$_d$), this is not consistently the case for the estimation of amplitude variation (LV$_\psi$). In summary, among the evaluated incomplete curve methods, varEM performs somewhat better than FGAMM, especially for representing the observed variation. We do not observe a large drop in estimation performance between the settings with weak and strong incompleteness.

— **Results Gamma settings**
For the Gamma settings, the varEM and SRVF approachesfall back on a misspecified Gaussian or "least squares" approach since neither are implemented for Gamma data. FGAMM utilizes the appropriate Gamma likelihood for both registration and GFPCA steps. The results are displayed in Figure 4. All in all, for the setting without correlation between amplitude, phase and incompleteness, the performance with regard to MISE$_h$ and MSE$_d$ is similar to the Gaussian case. All methods show consistently worse performance than in the Gaussian setting in terms of MISE$_y$ and LV$_\psi$, also for small amplitude ranks.

On complete data, the SRVF approach again performs worst in terms overall representation and warping function estimation, with the FGAMM median performance for amplitude rank 2–3 being better by 83% and 82%, respectively. Regarding the estimation of the overall representation and the warping functions, the (incomplete) FGAMM approach assuming the Gamma structure performs consistently better than varEM. However, FGAMM performs worse in recovering the FPC space, especially for the largest amplitude rank. For the estimation of the amplitude structure and the domain lengths, varEM leads to consistently better results. Comparing these results to the settings with weak and strong incompleteness,

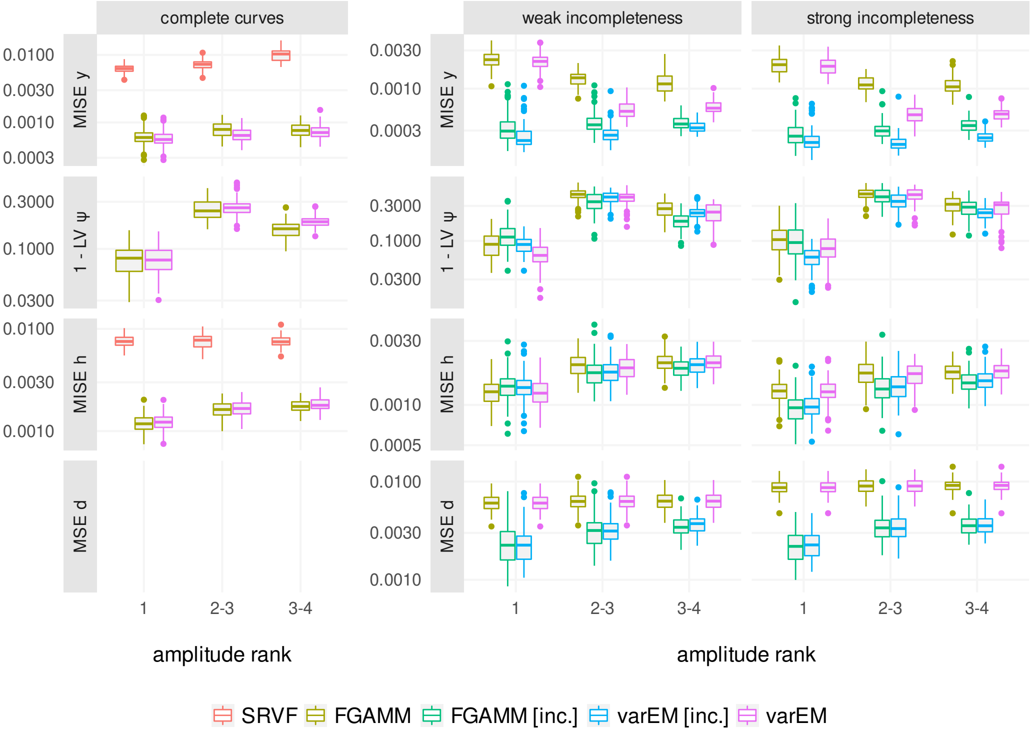

Fig. 3. Results for the simulation setting with Gaussian data and mutually uncorrelated amplitude, phase and amount of incompleteness. All y scales are $\log_{10}$ transformed.

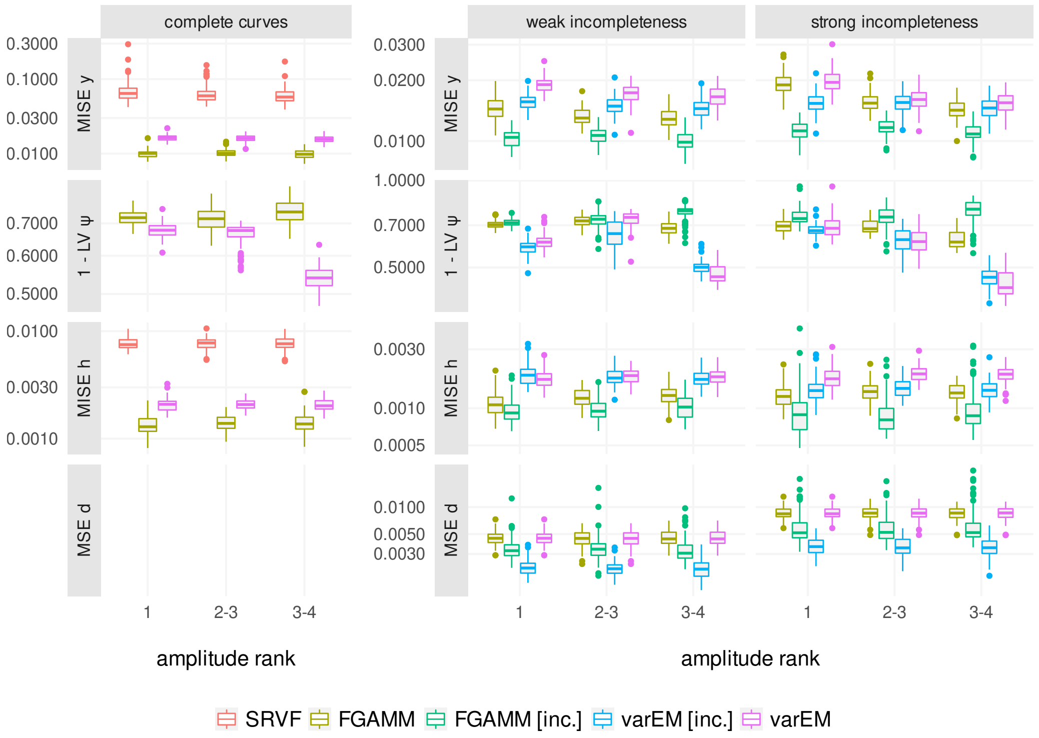

Fig. 4. Results for the simulation setting with Gamma data and uncorrelated amplitude, phase and amount of incompleteness. All y scales are $\log_{10}$ transformed.

the latter only show a structurally worse estimation performance for MSE$_d$.

4.3. Runtime analysis

We evaluate the efficiency of the approaches on one simulation setting with a Gaussian structure, amplitude rank 2–3 and complete curves. Only FGAMM is additionally applied to the respective setting with Gamma data. The median runtimes of each method, based on 20 runs, are visualized in Figure 5.

For the comparison, methods FGAMM and varEM ("varEM 2.1") are based on function `register_fpca` in version 2.1.5 of the `registr` package, which uses methods from packages `gamm4` (Wood and Scheipl, 2020) (v0.2.7) and `lme4` (Bates et al., 2015) (v1.1.26). These methods are compared to the old version of `registr` (v1.0.0, based on `gamm4` v0.2.6 and `lme4` v1.1.23), which does not contain the algorithmic improvements outlined in Appendix A1. We also compare our methods to function `align_fPCA` of package `fdasrvf` (Tucker, 2020) (v1.9.4) which implements the joint SRVF approach. All methods except version 1.0 of

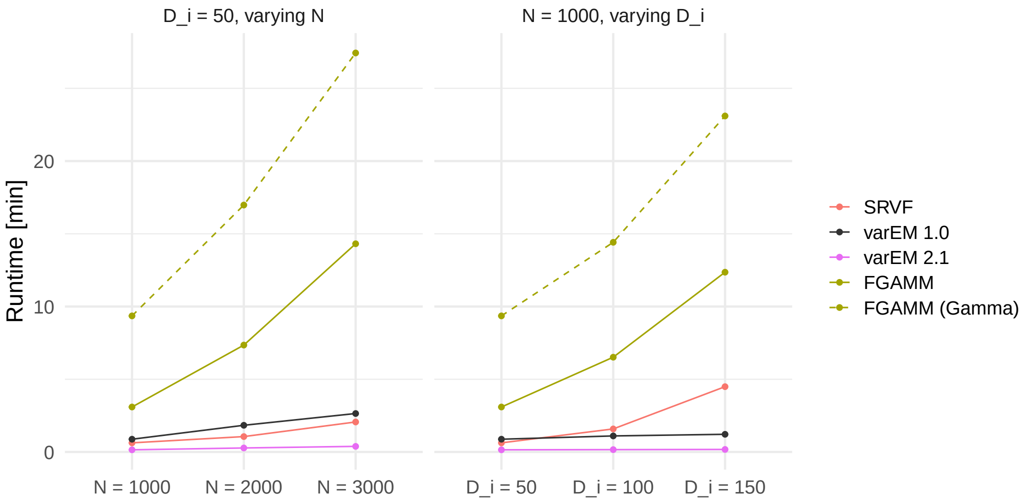

Fig. 5. Median runtimes for one setting of the simulation study with amplitude rank 2-3 and no incompleteness, based on 20 runs for each parameter combination. For the analysis, we vary the number of curves N and the number of measurements per curve D_i. Dashed curves are runtimes for FGAMM (Gamma).

package `registr` ("varEM 1.0") were run in parallel mode using ten cores.

— **Main findings**
As can be seen in Figure 5, the optimized algorithm in varEM 2.1 is clearly the most efficient method. For the setting with 50 measurements per curve and 3 000 curves ("$D_i = 50, N = 3000$") varEM 2.1 (runtime 23 seconds) is on average 86% faster than varEM 1.0 (159 seconds). The estimation of FGAMM is computationally much more expensive. For the setting "$D_i = 50, N = 3000$" it takes about 14 min, i.e., 37 times longer than varEM 2.1. Also, the runtime of FGAMM scales quadratically in both the number of curves and the number of measurements per curve. The efficiency of the SRVF approach lies between the other methods for smaller samples. However, it becomes computationally demanding for densely observed datasets with higher numbers of measurements per curve.

5. Application

5.1. Berkeley growth study

We compare FGAMM and varEM results on the well-known Berkeley growth data with simulated strong full incompleteness as outlined in Section 1 and visualized in Appendix A4.1. That is, we randomly remove both leading and trailing segments of the curves, with starting points and endpoints drawn at random in the first quarter and the last half of the time domain, respectively. Both methods are then applied with and without the assumption of completely observed curves, using a Gaussian likelihood and the same hyperparameters as used for the simulation study. The number of FPCs to be used in each iteration of the joint registration and FPCA algorithm was estimated adaptively, based on the criterion outlined at the end of Section 3.

While the FGAMM approach chose 5 (assuming completeness) and 4 (incomplete) FPCs, the varEM method chose 7 and 6 FPCs, respectively. In the comparison in Figure 6, we focus on the first two FPCs estimated by each method. Results in full detail are given in Appendix A4.2. Appendix A4.3 shows how the results of the incomplete curve FGAMM approach changes when different values for the penalization parameter λ are used.

While the first two FPCs estimated by the methods with assumed incompleteness show some differences, they represent similar main modes of amplitude variation. The first FPC mainly represents variation at the very beginning of the domain along with the information that the peak in growth in adolescent age appears earlier on if the growth rate in the very first year was stronger. The second FPC represents the information that if the initial growth rate was higher, the peak in adolescent age is more attenuated.

These first two FPCs as estimated by the incomplete curve approaches differ from the first FPCs estimated with assumed completeness. This is mainly due to the fact that the "completeness-assumed" approaches are not able to adequately align the structures observed in the last third of the domain (c.f. top row of Figure 6). In this data setting, the incomplete curve approaches are clearly better able to recover the underlying phase variation in the curves. In terms of computation time, both FGAMM variants and the incomplete varEM take about a minute, while varEM assuming completeness takes almost 2 minutes.

5.2. *Seismic ground motion propagation*

We analyze a subset of seismological interest of the seismic data outlined in Section 1, comprising 2 484 curves from various earthquakes simulated with different physical parameters. Sepcifically, we use data from simulated quakes characterized by (i) a sedimentary subsurface structure amplifying ground shaking, (ii) a geologically well-oriented direction of the tectonic background stress between 27° and 35°, and (iii) the friction parameter of critical linear slip weakening distance between 1.1m and 1.5m. We also restrict the data to those seismograms most relevant for seismic hazard assessment, which (i) lie in forward directivity direction (between cardinal directions 280° and 342°) to focus on wave propagations to the northwest in the direction of the main rupture pulse, and (ii) with a hypocentral distance shorter than 35km. Previous analyses show that the ground velocity curves we study are primarily shaped by the hypocentral distance of the measurement station and the *dynamic coefficient of friction* which resembles the frictional resistance of the geological fault during earthquake propagation (Bauer et al., 2017). For our analysis, we focus on how these two parameters and the topography of the evaluated region are associated with phase and amplitude variation.

As shown in Figure 2, all curves are pre-processed by cutting off any leading zero measurements below 0.01, leading to the observed time domain t_0^* which – being the *time since the first relevant absolute ground velocity measurement* – begins with the arrival time of seismic P-waves and comprises trailing incompleteness only towards the end of the domain after $t_0^* = 23.5$ seconds. Since this induces a MAR structure, where short observed domain

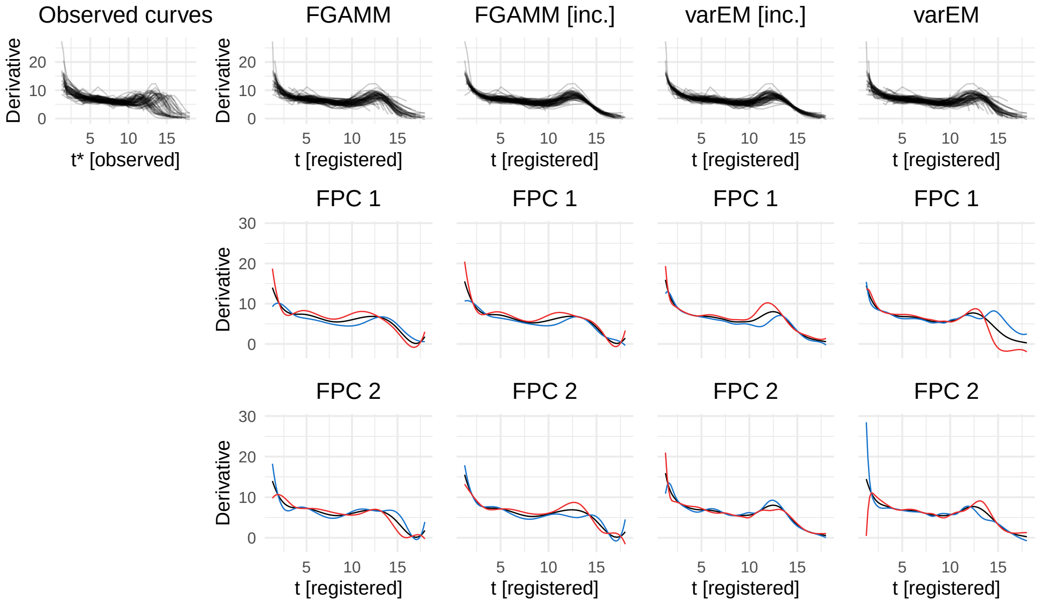

Fig. 6. Observed curves with simulated incompleteness (top left pane), registered curves (top row) and the first two estimated FPCs based on the different approaches. The FPCs $\psi_k(t)$ are visualized by the overall mean curve (solid line) plus (blue line) and minus (red line) $2 \cdot \sqrt{\hat{\tau}_k} \cdot \psi_k(t)$, with $\sqrt{\hat{\tau}_k}$ the standard deviation of the estimated scores for the k'th FPC.

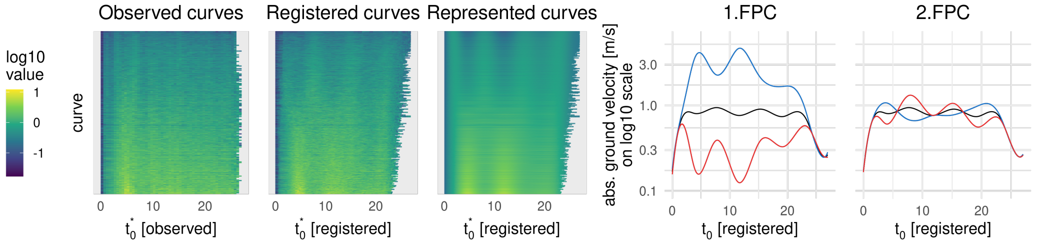

Fig. 7. Lasagna plots of observed and registered curves and of the curves as represented by the final GFPCA solution based on the FGAMM approach, on log10 scale (left pane). Curves are sorted by their maximum observed value. The FPCs $\psi_k(t)$ are visualized by the overall mean curve (black line) plus (blue line) and minus (red line) $2 \cdot \sqrt{\hat{\tau}_k} \cdot \psi_k(t)$, with $\sqrt{\hat{\tau}_k}$ the standard deviation of the estimated scores for the k'th FPC (right pane).

lengths are caused by higher hypocentral distances (causing later P-wave arrival times and smaller amplitudes), the results towards the end of the domain must be interpreted with great care.

We apply the FGAMM approach assuming a Gamma structure and trailing incompleteness, and using a similar parametrization as in the simulation study. The mean curve of all observed curves was used as the template function for the initial registration step. We use a penalization parameter of $\lambda = 0.004$ to discourage extreme distortions of the time domain. Estimation of the joint approach took ten joint iterations and a runtime of 3:31h using a parallelized call for the registration steps with 5 cores. Two FPCs were chosen based on the selection criterion outlined in Section 3 when aiming to explain 95% of amplitude variation.

The two estimated FPCs along with the observed, registered and represented curves are visualized in Figure 7. The full estimated warping functions are shown in Appendix A5.1. The first FPC as the main mode of amplitude variation represents the overall magnitude of the ground velocities, shaped by two salient peaks that resemble the shaking caused by surface wave phase arrivals. The second FPC represents a subsequent mode of variation and mainly shapes how pronounced the initial peak is.

The associations of phase and amplitude variation with the hypocentral distance and the dynamic coefficient of friction are visualized in Figure 8. Amplitude variation shows a very pronounced association structure with both parameters. Focusing on the first FPC, ground velocities are overall stronger the closer the measurement was taken to the hypocenter and the smaller the dynamic friction. The strongest ground motion is observed at hypocentral distances between 20 and 25km, caused by the nonlinear interaction of rupture propagation and the radiated seismic wavefield with topography and the subsurface structure. We find that in this region source effects (rupture directivity) and seismic wave path effects (surface waves) unleash the most energy. The second FPC's scores show a somewhat similar association structure but are more strongly shaped by the hypocentral distance. The highest scores were estimated at around 25km of distance, representing the most pronounced initial peak structure, especially in simulations with low dynamic friction values.

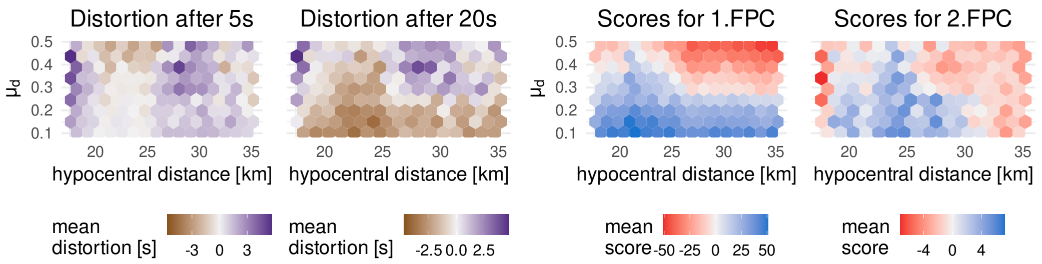

Fig. 8. Estimated phase and amplitude variation conditional on the hypocentral distance of the virtual seismometer and the dynamic coefficient of friction μ_d of the simulation. Phase variation and amplitude variation are shown by displaying the mean of the overall time distortion after 5 and 20 seconds (left pane, with positive and negative values representing time dilation and compression, respectively) and of the curves' mean scores for the FPCs shown in Figure 7, respectively.

Phase variation is also strongly associated with both evaluated parameters. The estimated time distortions at time t given by $\hat{h}_i^{-1}(t) - t$ for $t \in \{5, 20\}$ show somewhat similar patterns to the association structures of the FPC scores. This corroborates the structure displayed in Figure 2 which indicates a strong coupling between amplitude and phase variation since the initial peak is generally observed later for smaller overall observed ground velocities (an effect known in seismology as geometrical spreading). Stronger time distortion of the initial five seconds was mostly estimated for medium-to-large friction values, which are accompanied by smaller ground velocities. While these initial five seconds for curves observed between 20 and 25km of hypocentral distance (shaped by a more pronounced structure of the initial peak around $t_0 = 5$, see Figure 7) were mainly compressed, for curves closer to and farther away from the hypocenter (shaped by a less salient initial peak around $t_0 = 2$) they were mainly dilated. The time distortion of the initial 20 seconds shows a very similar structure to the scores for the first FPC. For curves with higher ground velocities, these initial 20 seconds are mainly compressed, and mainly dilated for lower ground velocities. Finally, the estimated inverse warping functions more often tend to more extreme distortions for higher hypocentral distance and higher dynamic friction values (see Appendix A5, Figure 20). This is due to curves under these conditions showing very small ground velocities with a less salient structure that is hard to align to the estimated template functions.

As expected due to the dominant effects of source directivity and surface waves in the evaluated region around the hypocenter, no structural association of amplitude or phase variation with the local topography was detected (see Appendix A5.2). The obtained results are geophysically plausible and in line with previous analyses of the seismic experiments (Bauer et al., 2017).

6. Discussion

Incomplete data are very common in longitudinal settings but remain under-discussed in many fields of functional data analysis. Our likelihood-based approach for joint registration and generalized FPCA allows for analyzing curves with leading, trailing or full incompleteness in the presence of substantial phase variation and is able to handle non-Gaussian data. All methods are implemented in the open-source R package `registr`.

Our simulation study results indicate that accounting for incompleteness improves the performance in different data settings. While the FGAMM approach shows some bias in the estimation of the underlying FPC structure in the Gamma settings, its substantially better estimation of the warping functions leads to improved overall performance in terms of the representation of the joint phase and amplitude variation structure of the individual curves. Stronger incompleteness does not seem to structurally harm the overall performance. Applications to incomplete Berkeley growth curves and a seismic data setting showcase the practical utility of our new approach.

— **Comparison to SRVF-based approaches**
In contrast to methods based on the SRVF framework, we do not utilize the warping-invariant Fisher-Rao metric. Instead, our flexible penalized likelihood-based approach allows for representing more complex structures of variation in diverse non-Gaussian data situations and is backed by robust optimization algorithms. While SRVF approaches rely on the availability of functional derivatives evaluated on a common, regular grid and may struggle in the presence of stronger (non-Gaussian) noise, this is generally not the case for our method. We utilize a low-dimensional B-spline basis for the inverse warping functions. In our applications, this seemed sufficient to avoid the pinching problem. Extreme time distortions were only estimated for few seismic curve outliers without a pronounced shape.

— **Covariance estimation**
One central topic for future research on GFPCA is a thorough evaluation of the consistency and robustness of different covariance estimators. This comprises questions like at what point in the estimation procedure smoothing and centering (of the raw curves or the final covariance surface) should be performed to obtain the best estimator. Covariance estimators should be evaluated for common practical data settings entailing relevant non-Gaussian noise in combination with small numbers of curves and measurements per curve and different levels of their respective density over the domain.

— **Computational efficiency**
A practical constraint for the application of the evaluated methods remains their computational efficiency in large-scale data settings. In this regard, a promising strain of research are recently proposed neural network based frameworks like Nunez et al. (2021) and Chen and Srivastava (2021) for registration and Sarkar and Panaretos (2021) for covariance estimation.

References

Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**, 1–48.

Bauer, A. (2021) bauer-alex/incfundatregistration_supp: Full code and data supplement.

Bauer, A., Scheipl, F., Küchenhoff, H. and Gabriel, A.-A. (2017) Modeling spatio-temporal earthquake dynamics using generalized functional additive regressions. In *International Workshop on Statistical Modelling 2017*. URL: `https://tinyurl.com/bauer-2017-poster`.

— (2018) An introduction to semiparametric function-on-scalar regression. *Statistical Modelling*, **18**, 346–364.

Bryner, D. and Srivastava, A. (2021) Shape analysis of functional data with elastic partial matching. *arXiv preprint arXiv:2105.08604*.

Cao, G., Wang, L., Li, Y. and Yang, L. (2016) Oracle-efficient confidence envelopes for covariance functions in dense functional data. *Statistica Sinica*, **26**, 359–383.

Carroll, C., Gajardo, A., Chen, Y., Dai, X., Fan, J., Hadjipantelis, P. Z., Han, K., Ji, H., Mueller, H.-G. and Wang, J.-L. (2020) *fdapace: Functional Data Analysis and Empirical Dynamics*. URL: `https://CRAN.R-project.org/package=fdapace`. R package version 0.5.5.

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

Chakraborty, A. and Panaretos, V. M. (2021) Functional registration and local variations: Identifiability, rank, and tuning. *Bernoulli*, **27**, 1103–1130.

Chen, C. and Srivastava, A. (2021) Srvfregnet: Elastic function registration using deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4462–4471.

Cheng, W., Dryden, I. L., Huang, X. et al. (2016) Bayesian registration of functions and curves. *Bayesian Analysis*, **11**, 447–475.

Claeskens, G., Silverman, B. W. and Slaets, L. (2010) A multiresolution approach to time warping achieved by a bayesian prior–posterior transfer fitting strategy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**, 673–694.

Delaigle, A., Hall, P., Huang, W. and Kneip, A. (2020) Estimating the covariance of fragmented and other related types of functional data. *Journal of the American Statistical Association*, 1–19.

Descary, M.-H. and Panaretos, V. M. (2019) Recovering covariance from functional fragments. *Biometrika*, **106**, 145–160.

Fahrmeir, L., Kneib, T., Lang, S. and Marx, B. (2013) *Regression: models, methods and applications.* Springer Science & Business Media.

Gertheiss, J., Goldsmith, J. and Staicu, A.-M. (2017) A note on modeling sparse exponential-family functional response curves. *Computational statistics & data analysis*, **105**, 46–52.

Giorgino, T. (2009) Computing and visualizing dynamic time warping alignments in R: The dtw package. *Journal of Statistical Software*, **31**, 1–24.

Goldsmith, J. (2016) *gfpca: Generalized Functional Principal Components Analysis.* URL: `https://github.com/jeff-goldsmith/gfpca`. R package version 1.1.0. URL https://github.com/jeff-goldsmith/gfpca.

Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J., Di, C., Gellar, J., Harezlak, J., McLean, M. W., Swihart, B., Xiao, L., Crainiceanu, C. and Reiss, P. T. (2020) *refund: Regression with Functional Data.* URL: `https://CRAN.R-project.org/package=refund`. R package version 0.1-23.

Goldsmith, J., Zipunnikov, V. and Schrack, J. (2015) Generalized multilevel function-on-scalar regression and principal component analysis. *Biometrics*, **71**, 344–353.

Guo, X., Kurtek, S. and Bharath, K. (2020) Variograms for spatial functional data with phase variation. *arXiv preprint arXiv:2010.09578.*

Hadjipantelis, P. Z., Aston, J. A., Müller, H.-G. and Evans, J. P. (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**, 545–559.

Hall, P., Müller, H.-G. and Yao, F. (2008) Modelling sparse generalized longitudinal observations with latent gaussian processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**, 703–723.

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**, e220.

Happ-Kurz, C. (2020) *MFPCA: Multivariate Functional Principal Component Analysis for Data Observed on Different Dimensional Domains.* URL: `https://github.com/ClaraHapp/MFPCA`. R package version 1.3-6.

Huang, H., Li, Y. and Guan, Y. (2014) Joint modeling and clustering paired generalized longitudinal trajectories with application to cocaine abuse treatment data. *Journal of the American Statistical Association*, **109**, 1412–1424.

James, G. M., Hastie, T. J. and Sugar, C. A. (2000) Principal component models for sparse functional data. *Biometrika*, **87**, 587–602.

Kneip, A. and Gasser, T. (1992) Statistical tools to analyze data representing a sample of curves. *The Annals of Statistics*, 1266–1305.

Kneip, A. and Ramsay, J. O. (2008) Combining registration and fitting for functional models. *Journal of the American Statistical Association*, **103**, 1155–1165.

Kraus, D. (2015) Components and completion of partially observed functional data. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 777–801.

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

Larsson, R. and Villani, M. (2001) A distance measure between cointegration spaces. *Economics Letters*, **70**, 21–27.

Learned-Miller, E. G. (2005) Data driven image models through continuous joint alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28**, 236–250.

Lee, S. and Jung, S. (2016) Combined analysis of amplitude and phase variations in functional data. *arXiv preprint arXiv:1603.01775.*

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

Li, Y. and Guan, Y. (2014) Functional principal component analysis of spatiotemporal point processes with applications in disease surveillance. *Journal of the American Statistical Association*, **109**, 1205–1215.

Liebl, D. and Rameseder, S. (2019) Partially observed functional data: The case of systematically missing parts. *Computational Statistics & Data Analysis*, **131**, 104–115.

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**, 307–333.

Lu, Y., Herbei, R. and Kurtek, S. (2017) Bayesian registration of functions with a gaussian process prior. *Journal of Computational and Graphical Statistics*, **26**, 894–904.

Marron, J. S., Ramsay, J. O., Sangalli, L. M. and Srivastava, A. (2015) Functional data analysis of amplitude and phase variation. *Statistical Science*, 468–484.

Mattar, M. A., Ross, M. G. and Learned-Miller, E. G. (2009) Nonparametric curve alignment. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 3457–3460. IEEE.

Matuk, J., Bharath, K., Chkrebtii, O. and Kurtek, S. (2019) Bayesian framework for simultaneous registration and estimation of noisy, sparse and fragmented functional data. *arXiv preprint arXiv:1912.05125.*

McDonnell, E. I., Zipunnikov, V., Schrack, J. A., Goldsmith, J. and Wrobel, J. (2021) Registration of 24-hour accelerometric rest-activity profiles and its application to human chronotypes. *Biological Rhythm Research*, 1–21.

Müller, M. (2015) *Fundamentals of music processing: Audio, analysis, algorithms, applications*. Springer.

Nunez, E., Lizarraga, A. and Joshi, S. H. (2021) Srvfnet: A generative network for unsupervised multiple diffeomorphic functional alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4481–4489.

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

Pelties, C., Gabriel, A.-A. and Ampuero, J.-P. (2014) Verification of an ader-dg method for complex dynamic rupture problems. *Geoscientific Model Development*, 847–866.

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/`. URL https://www.R-project.org/.

Ramsay, J. O., Graves, S. and Hooker, G. (2020) *fda: Functional Data Analysis*. URL: `https://CRAN.R-project.org/package=fda`. R package version 5.1.9.

Ramsay, J. O. and Li, X. (1998) Curve registration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **60**, 351–363.

Ramsay, J. O. and Silverman, B. W. (2005) *Functional Data Analysis*. Springer.

Rice, J. A. and Wu, C. O. (2001) Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, **57**, 253–259.

Sangalli, L. M., Secchi, P., Vantini, S. and Vitelli, V. (2010) K-mean alignment for curve clustering. *Computational Statistics & Data Analysis*, **54**, 1219–1233.

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

Scheipl, F., Gertheiss, J., Greven, S. et al. (2016a) Generalized functional additive mixed models. *Electronic Journal of Statistics*, **10**, 1455–1492.

Scheipl, F., Greven, S. et al. (2016b) Identifiability in penalized function-on-function regression models. *Electronic Journal of Statistics*, **10**, 495–526.

Serban, N., Staicu, A.-M. and Carroll, R. J. (2013) Multilevel cross-dependent binary longitudinal data. *Biometrics*, **69**, 903–913.

Srivastava, A., Wu, W., Kurtek, S., Klassen, E. and Marron, J. S. (2011) Registration of functional data using fisher-rao metric. *arXiv preprint arXiv:1103.3817*.

Stefanucci, M., Sangalli, L. M. and Brutti, P. (2018) Pca-based discrimination of partially observed functional data, with an application to aneurisk65 data set. *Statistica Neerlandica*, **72**, 246–264.

Telesca, D. and Inoue, L. Y. T. (2008) Bayesian hierarchical curve registration. *Journal of the American Statistical Association*, **103**, 328–339.

Tipping, M. E. and Bishop, C. M. (1999) Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **61**, 611–622.

Tormene, P., Giorgino, T., Quaglini, S. and Stefanelli, M. (2009) Matching incomplete time series with dynamic time warping: an algorithm and an application to post-stroke rehabilitation. *Artificial intelligence in medicine*, **45**, 11–34.

Tucker, J. D. (2014) *Functional component analysis and regression using elastic methods.* Ph.D. thesis, The Florida State University.

— (2020) *fdasrvf: Elastic Functional Data Analysis.* URL: `https://CRAN.R-project.org/package=fdasrvf`. R package version 1.9.4.

Tucker, J. D., Shand, L. and Chowdhary, K. (2021) Multimodal bayesian registration of noisy functions using hamiltonian monte carlo. *Computational Statistics & Data Analysis*, 107298.

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

Uphoff, C., Rettenberger, S., Bader, M., Madden, E. H., Ulrich, T., Wollherr, S. and Gabriel, A.-A. (2017) Extreme scale multi-physics simulations of the tsunamigenic 2004 sumatra megathrust earthquake. In *Proceedings of the international conference for high performance computing, networking, storage and analysis*, 1–16.

Vitelli, V. (2019) A novel framework for joint sparse clustering and alignment of functional data. *arXiv preprint arXiv:1912.00687.*

Wagner, H. (2020) *Nonparametric Registration to Low Dimensional Function Spaces.* URL: `https://github.com/heikowagner/Nonparametric-Registration-to-Low-Dimensional-Function-Spaces`.

Wagner, H. and Kneip, A. (2019) Nonparametric registration to low-dimensional function spaces. *Computational Statistics & Data Analysis*, **138**, 49–63.

Wang, J.-L., Chiou, J.-M. and Müller, H.-G. (2016) Functional data analysis. *Annual Review of Statistics and Its Application*, **3**, 257–295.

Wood, S. and Scheipl, F. (2020) *gamm4: Generalized Additive Mixed Models using 'mgcv' and 'lme4'.* URL: `https://CRAN.R-project.org/package=gamm4`. R package version 0.2-6. URL https://CRAN.R-project.org/package=gamm4.

Wrobel, J. and Bauer, A. (2021) registr 2.0: Incomplete curve registration for exponential family functional data. *Journal of Open Source Software*, **6**, 2964.

Wrobel, J., Muschelli, J. and Leroux, A. (2021) Diurnal physical activity patterns across ages in a large uk based cohort: The uk biobank study. *Sensors*, **21**, 1545.

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

Wu, W. and Srivastava, A. (2014) Analysis of spike train data: Alignment and comparisons using the extended fisher-rao metric. *Electronic Journal of Statistics*, **8**, 1776–1785.

Yao, F., Müller, H.-G. and Wang, J.-L. (2005) Functional data analysis for sparse longitudinal data. *Journal of the American statistical association*, **100**, 577–590.

Zhang, A. and Chen, K. (2017) Nonparametric covariance estimation for mixed longitudinal studies, with applications in midlife women's health. *arXiv preprint arXiv:1711.00101.*

Zhou, L., Huang, J. Z. and Carroll, R. J. (2008) Joint modelling of paired sparse functional data using principal components. *Biometrika*, **95**, 601–619.

Appendix to
Registration for Incomplete Non-Gaussian Functional Data - Disentangling the variation structure of seismic ground velocities

Alexander Bauer †

Department of Statistics, LMU Munich, Germany.

E-mail: alexander.bauer@stat.uni-muenchen.de

Fabian Scheipl

Department of Statistics, LMU Munich, Germany.

Helmut Küchenhoff

Department of Statistics, LMU Munich, Germany.

Alice-Agnes Gabriel

Department of Earth and Environmental Sciences, LMU Munich, Germany.

A1. Computational Details

We implemented our approach in the package `registr` (Wrobel and Bauer, 2021) for the statistical open-source software R (R Core Team, 2020). The `registr` package allows for the estimation of the joint registration and GFPCA approach both for complete and incomplete curves. All three types of incompleteness (leading, trailing and full incompleteness) and irregular grids are supported. Additional to the methods outlined in this work the package comprises the methods of Wrobel et al. (2019). Several exponential family distributions are available. In the following, we give details on some computational aspects of our method.

A1.1. Registration

The registration codebase builds on the implementation outlined in Wrobel et al. (2019) and Wrobel (2018). We extended the methods by allowing the observed curves to be incomplete. Since the estimation of warping functions in the registration step is performed separately for each curve, we added the option of a parallelized call over the individual curves.

Constrained optimization for the spline coefficients representing the warpings is performed with function `constrOptim()` by inducing linear inequality constraints of the form

$$\boldsymbol{u}_i \cdot \boldsymbol{\beta}_i - \boldsymbol{c}_i \geq 0,$$

†A. Bauer's research was partially funded by the German Research Foundation (DFG) under Grant KU 1359/4-1; F. Scheipl's research was partially funded by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A. The authors of this work take full responsibilities for its content.

with parameter vector $\boldsymbol{\beta}_i$ and constraints given by matrix $\boldsymbol{u}_i$ and vector $\boldsymbol{c}_i$. Further details on the constraint matrices are given in Appendix A2. Alternative optimization algorithms from the NLopt library (Johnson, 2020) and made available by package `nloptr` (Ypma and Johnson, 2020) were evaluated as well, but did not improve the overall results or lead to a more efficient estimation.

A1.2. Generalized Functional Principal Component Analysis

As outlined in Section 3.2, our adaptation of the two-step GFPCA approach of Gertheiss et al. (2017) is based on an additive regression model with random intercept terms for the individual FPCs. We build on robust and highly efficient software for these kinds of models, available in packages `gamm4` (Wood and Scheipl, 2020) and `lme4` (Bates et al., 2015). The algorithms of `lme4` are highly efficient for estimating models with random intercept terms with several thousand individual categories. The estimation of the marginal mean of the process $X_i(t)$ in the case of very large data with $> 100\,000$ rows is performed with the discretization-based estimation algorithm of function `mgcv::bam` (Wood et al., 2017) rather than the estimation algorithm of `mgcv::gam` (Wood, 2017).

Our implementation of the two-step GFPCA approach of Gertheiss et al. (2017) is based on their accompanying package `gfpca` (Goldsmith, 2016). Additionally we made several adjustments to their codebase to improve overall efficiency: First, while functions `lmer()` and `glmer()` from the `lme4` package default to the optimization routine implemented in function `bobyqa` (package `minqa`, Bates et al., 2014), we make use of the more efficient optimizer `NLOPT_LN_BOBYQA` from the NLopt library (Johnson, 2020) as described in Powell (2009).

Second, we tackle one major issue in the building of the covariance structure. In principle, the covariance matrix comprises the pairwise covariances between all unique observed time points per functional datum $y_i(t)$. In real data situations with highly irregular grids, the number of unique combinations of time points can explode in size even for settings with a relatively low number of curves. We utilize a binning strategy to handle this problem. Before building the covariance matrix, we round the vector of observed time points to k significant digits. E.g., $k = 3$ then leads to at most 1000^2 unique combinations and a covariance matrix with maximal size 1000×1000. Similar to the estimation of the marginal mean of $X_i(t)$, the smoothing of the covariance surface is performed with the discretization-based estimation algorithm `mgcv::bam` rather than `mgcv::gam` if the crossproduct matrix comprises $> 100\,000$ elements.

Third, we updated the codebase of `gamm4` to make the initial construction of the random effect model matrices much more efficient by fully exploiting their sparse structure. Our patched version is currently available on GitHub (`https://github.com/r-gam/gamm4`) and will in future be integrated into the main codebase of the `gamm4` package.

A1.3. Joint approach

To make the overall algorithm more efficient, we introduce two major changes. First, all *intermediate iterations* regarding the GFPCA step, apart from the very first and the very last one, are performed with less accuracy. Above all else, we use larger tolerance values and a simple Laplace approximation to the GLMM likelihood (i.e., `nAGQ = 0` and `nAGQinitStep = FALSE` in function `gamm4::gamm4`) for these iterations. Secondly, we use the solution of the previous GFPCA step as starting values for the subsequent GFPCA step.

A2. Constraint Matrices for `constrOptim()`

As outlined in Appendix A1, we estimate the warping functions using function `constrOptim()`. In the estimation step for one warping function, the parameter vector is constrained s.t. the resulting warping function is monotone and does not exceed the overall time domain $[t_{min}, t_{max}]$.

In the following the constraint matrices are listed for the different settings of (in)completeness and assuming a parameter vector of length p:

$$\boldsymbol{\beta}_i = \begin{pmatrix} \beta_{i1} \\ \beta_{i2} \\ \vdots \\ \beta_{ip} \end{pmatrix} \in \mathbb{R}_{p \times 1}$$

A2.1. Complete curve setting

When all curves were observed completely – i.e. the underlying processes of interest were all observed from the beginning until the end – warping functions can typically be assumed to start and end on the diagonal, since each process is completely observed in its observation interval $[t^*_{min,i}, t^*_{max,i}] \subset [t_{min}, t_{max}]$.

Assuming that both the starting point and the endpoint lie on the diagonal, we set $\beta_{i1} = t^*_{min,i}$ and $\beta_{ip} = t^*_{max,i}$ and only perform the estimation for

$$\begin{pmatrix} \beta_{i2} \\ \beta_{i3} \\ \vdots \\ \beta_{i(p-1)} \end{pmatrix} \in \mathbb{R}_{(p-2) \times 1}$$

This results in the following constraint matrices, that allow a mapping from the observed domain $[t^*_{min,i}, t^*_{max,i}]$ to the domain itself $[t^*_{min,i}, t^*_{max,i}] \subset [t_{min}, t_{max}]$:

$$
\boldsymbol{u}_i = \begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & \dots & 0 & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & 0 & \dots & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & -1 \end{pmatrix} \in \mathbb{R}_{(p-1)\times(p-2)}
$$

$$
\boldsymbol{c}_i = \begin{pmatrix} t^*_{min,i} \\ 0 \\ 0 \\ \vdots \\ 0 \\ -1 \cdot t^*_{max,i} \end{pmatrix} \in \mathbb{R}_{(p-1)\times 1}
$$

A2.2. Leading incompleteness only

In the case of leading incompleteness – i.e. the underlying processes of interest were all observed until their very end but not necessarily starting from their beginning – warping functions can typically be assumed to end on the diagonal, s.t. one assumes $\beta_{ip} = t^*_{max,i}$ to let the warping functions end at the last observed time point $t^*_{max,i}$. The estimation is then performed for the remaining parameter vector

$$
\begin{pmatrix} \beta_{i1} \\ \beta_{i3} \\ \vdots \\ \beta_{i(p-1)} \end{pmatrix} \in \mathbb{R}_{(p-1)\times 1}
$$

This results in the following constraint matrices, that allow a mapping from the observed domain $[t^*_{min,i}, t^*_{max,i}]$ to the domain $[t_{min}, t^*_{max,i}] \subset [t_{min}, t_{max}]$:

$$
\boldsymbol{u}_i = \begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & \dots & 0 & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & 0 & \dots & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & -1 \end{pmatrix} \in \mathbb{R}_{p\times(p-1)}
$$

$$
\boldsymbol{c}_i = \begin{pmatrix} t_{min} \\ 0 \\ 0 \\ \vdots \\ 0 \\ -1 \cdot t^*_{max,i} \end{pmatrix} \in \mathbb{R}_{p\times 1}
$$

A2.3. Trailing incompleteness only

In the case of trailing incompleteness – i.e. the underlying processes of interest were all observed from the beginning but not necessarily until their very end – warping functions can typically be assumed to start on the diagonal, s.t. one assumes $\beta_{i1} = t^*_{min,i}$ to let the warping functions start at the first observed time point $t^*_{min,i}$. The estimation is then performed for the remaining parameter vector

$$\begin{pmatrix} \beta_{i2} \\ \beta_{i3} \\ \vdots \\ \beta_{ip} \end{pmatrix} \in \mathbb{R}_{(p-1)\times 1}$$

This results in the following constraint matrices, that allow a mapping from the observed domain $[t^*_{min,i}, t^*_{max,i}]$ to the domain $[t^*_{min,i}, t_{max}] \subset [t_{min}, t_{max}]$:

$$\boldsymbol{u}_i \text{ identical to the version for leading incompleteness}$$

$$\boldsymbol{c}_i = \begin{pmatrix} t^*_{min,i} \\ 0 \\ 0 \\ \vdots \\ 0 \\ -1 \cdot t_{max} \end{pmatrix} \in \mathbb{R}_{p\times 1}$$

A2.4. Leading and trailing incompleteness

In the case of both leading and trailing incompleteness – i.e. the underlying processes of interest were neither necessarily observed from their very beginnings nor to their very ends – warping functions can typically only be assumed to map the observed domains $[t^*_{min,i}, t^*_{max,i}]$ to the overall domain $[t_{min}, t_{max}]$.

This results in the following constraint matrices:

$$\boldsymbol{u}_i = \begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & \dots & 0 & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & 0 & \dots & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & -1 \end{pmatrix} \in \mathbb{R}_{(p+1)\times p}$$

$$\boldsymbol{c}_i = \begin{pmatrix} t_{min} \\ 0 \\ 0 \\ \vdots \\ 0 \\ -1 \cdot t_{max} \end{pmatrix} \in \mathbb{R}_{(p+1)\times 1}$$

6 *Appendix: Bauer et al. (in revision)*

A3. Simulation study

A3.1. Simulation setting

This subsection contains figures for all relevant components of the curves simulated in the simulation study.

A3.1.1. Distribution of the data

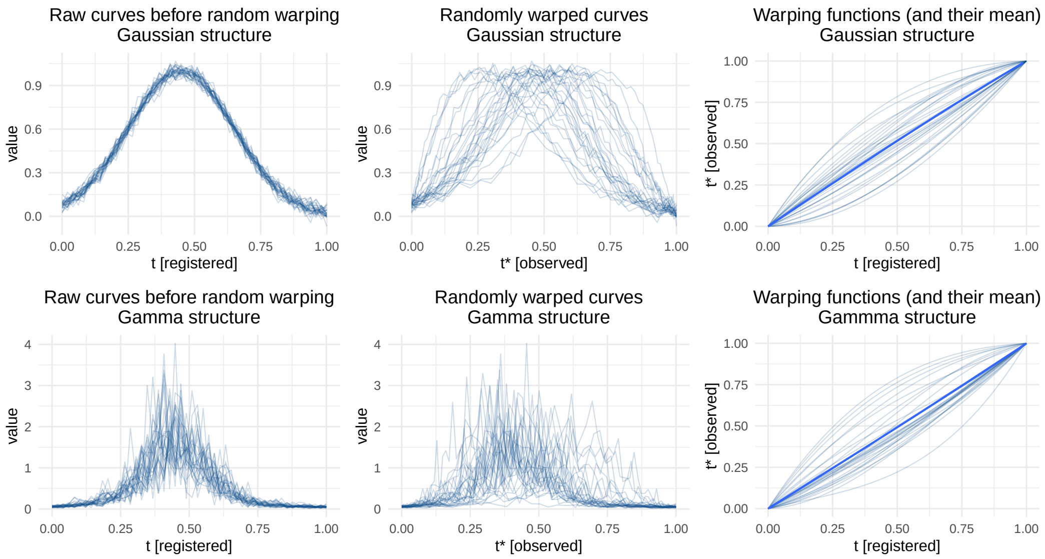

Fig. 1. Structure of simulated Gaussian and Gamma data (top and bottom row, respectively), before adding amplitude variation

A3.1.2. Rank of amplitude variation

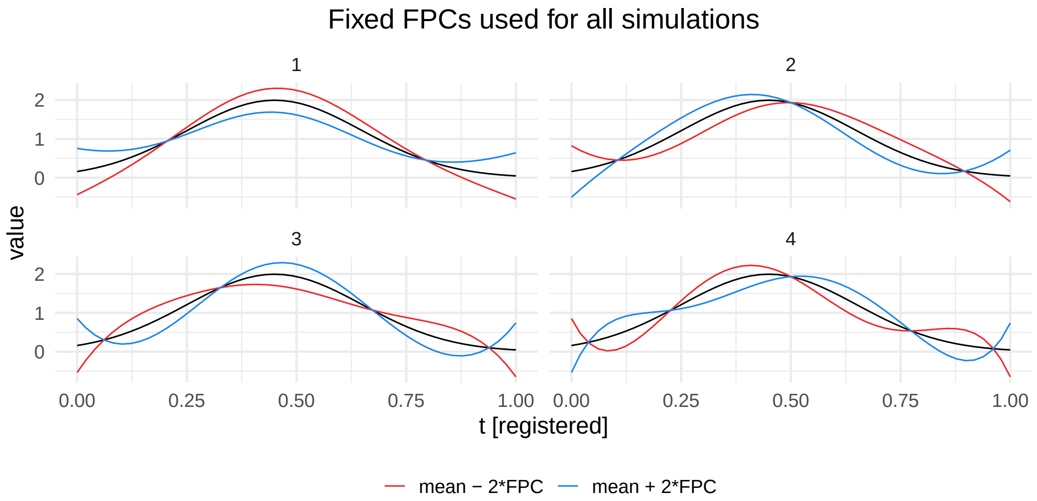

Fig. 2. Simulated eigenfunctions / functional principal components (FPCs), visualized by adding and subtracting them from a some mean curve (black line)

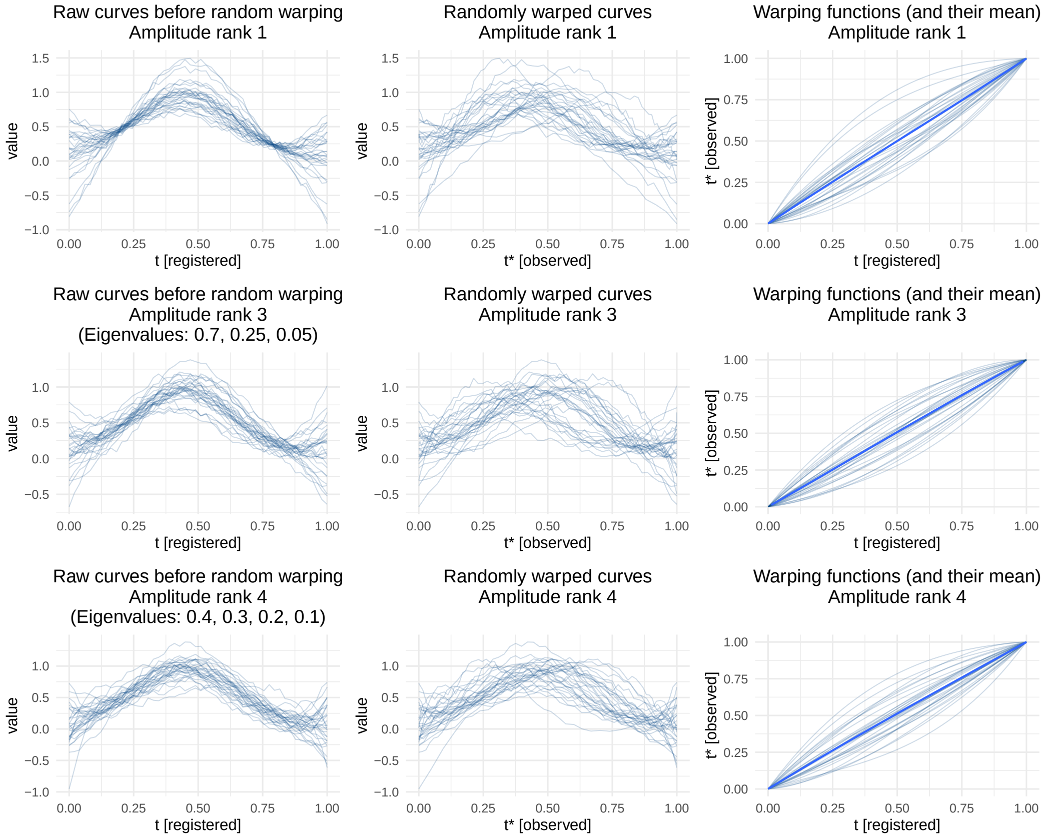

Fig. 3. Simulated curves with Gaussian structure, including amplitude variation and random warping.

A3.1.3. Strength of incompleteness

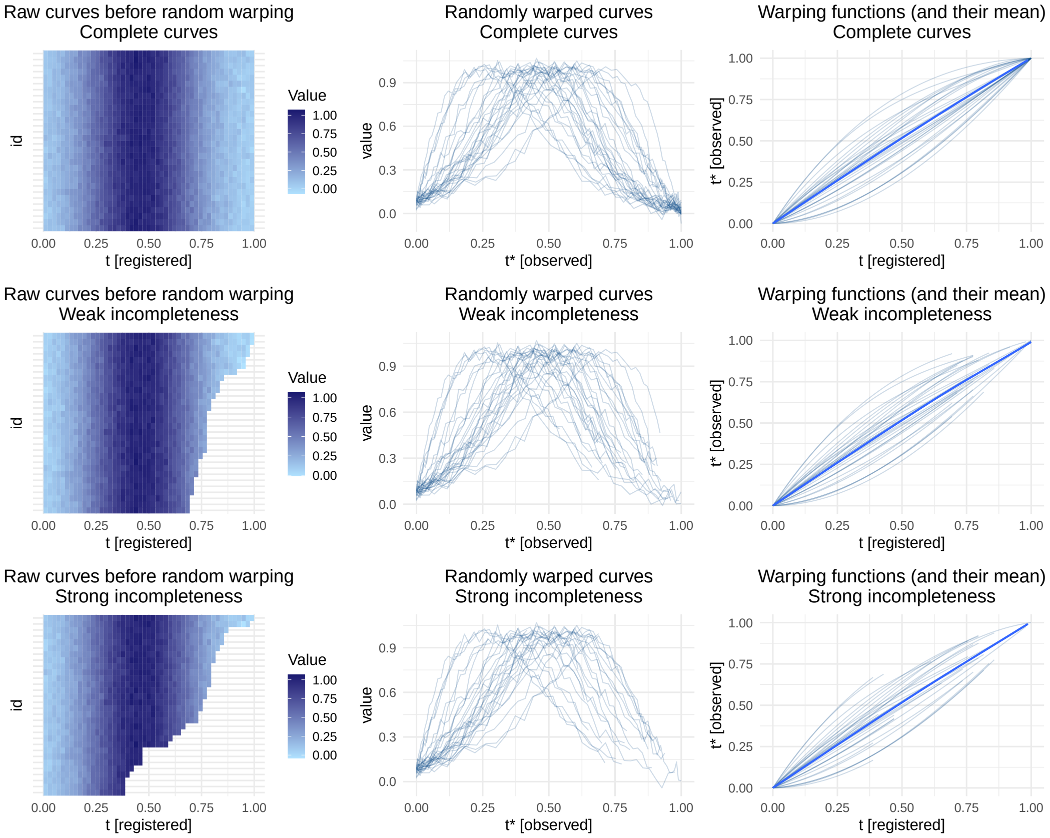

Fig. 4. Simulated curves with Gaussian structure and different strengths of incompleteness.

A3.1.4. Correlation structure

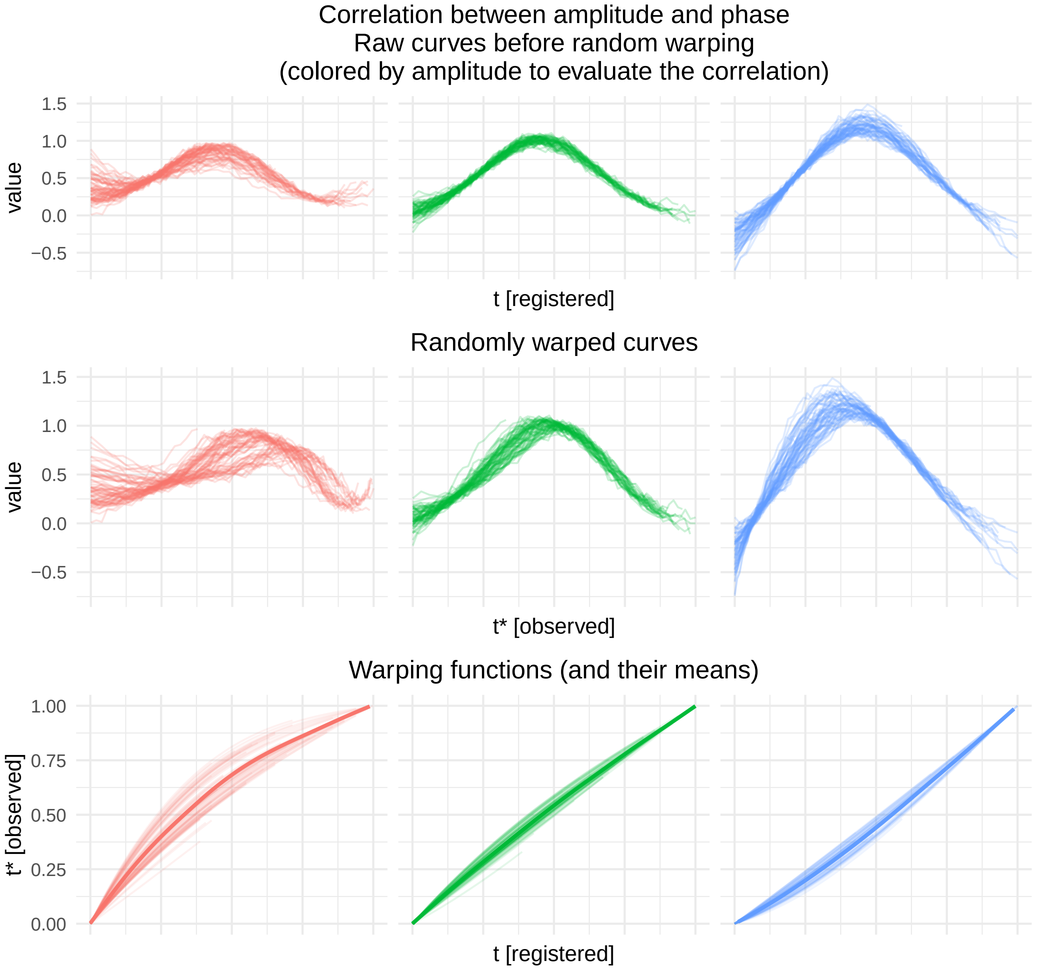

Fig. 5. Simulated curves with Gaussian structure and correlated amplitude and phase variation.

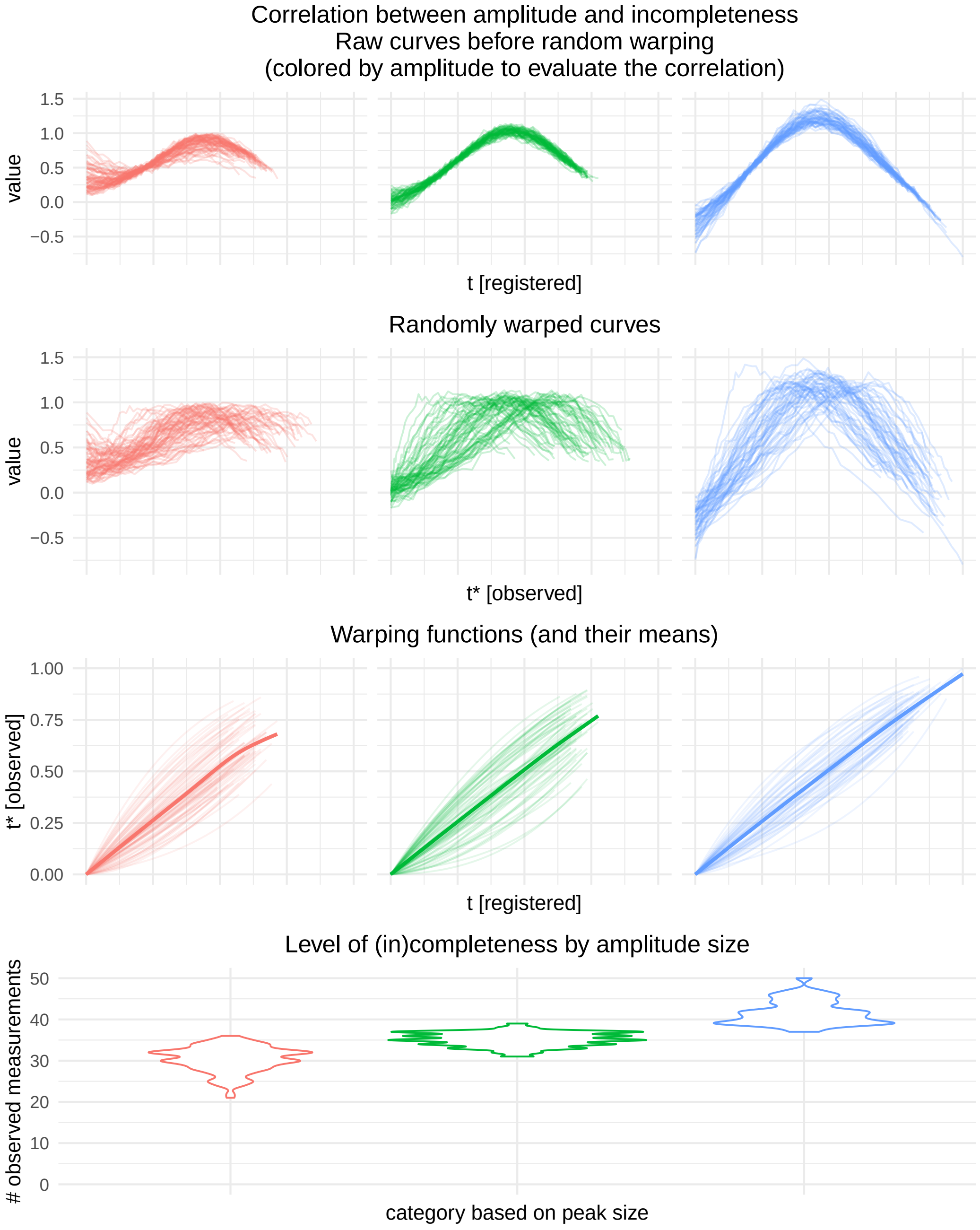

Fig. 6. Simulated curves with Gaussian structure and correlated amplitude variation and amount of incompleteness.

A3.2. Simulation results – Gaussian with correlation structure

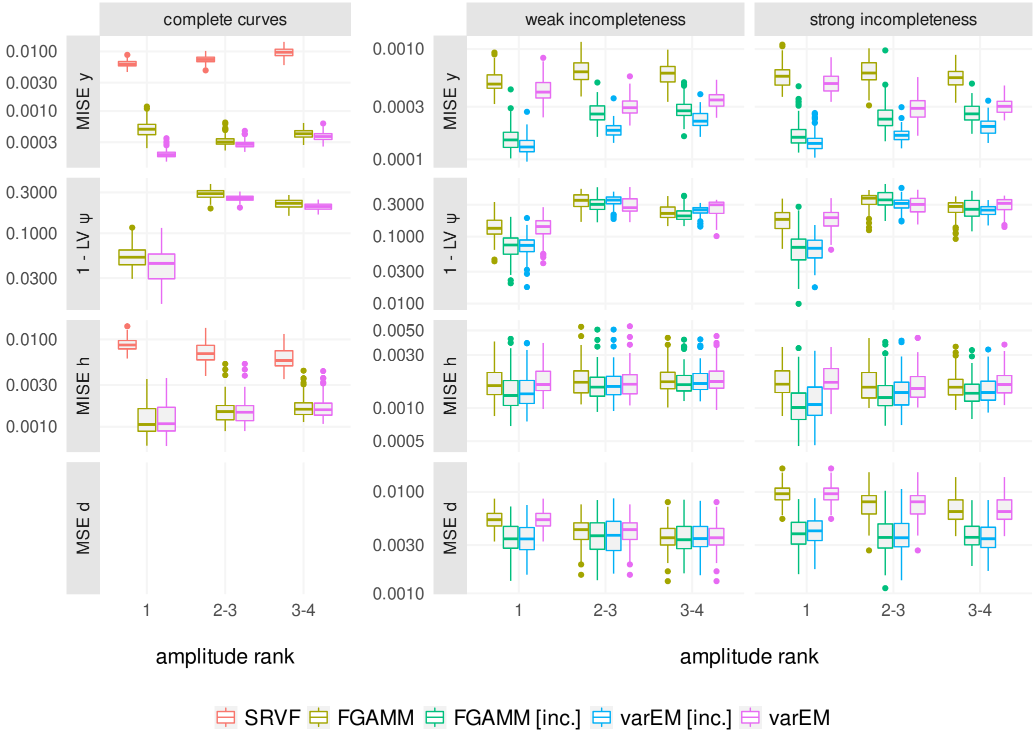

Fig. 7. Results for the simulation setting with Gaussian data and a correlation between amplitude and phase.

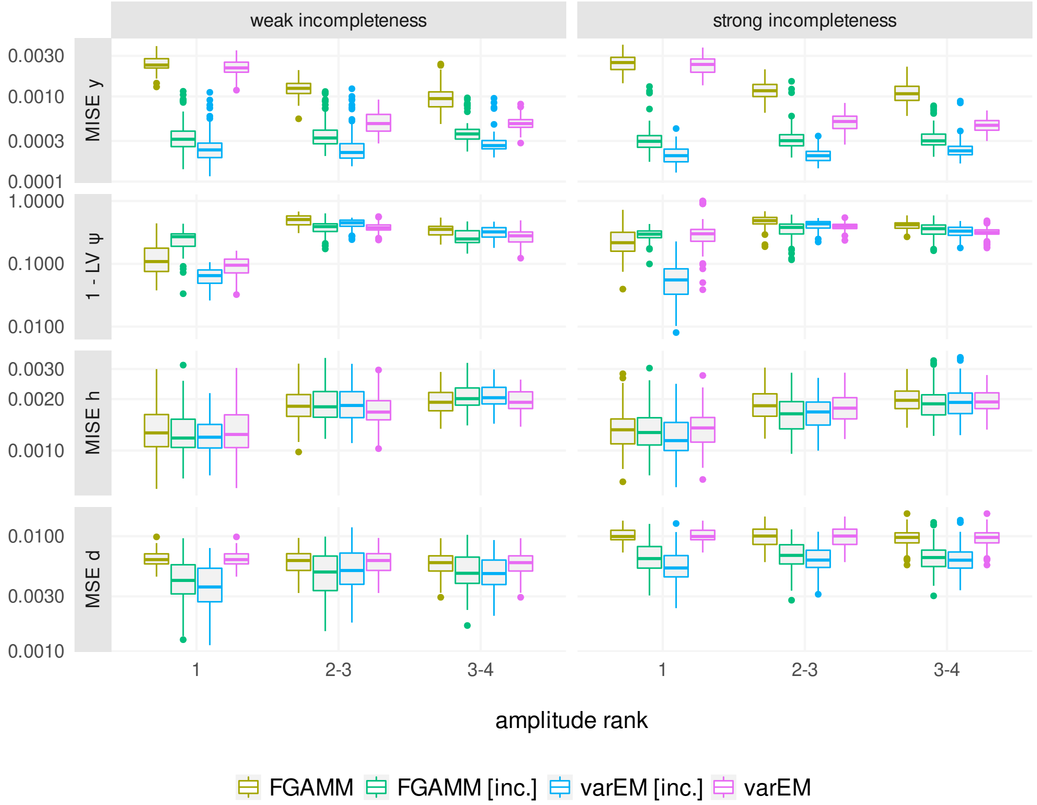

Fig. 8. Results for the simulation setting with Gaussian data and a correlation between amplitude and the amount of incompleteness.

14 *Appendix: Bauer et al. (in revision)*

A3.3. Simulation results – Gamma with correlation structure

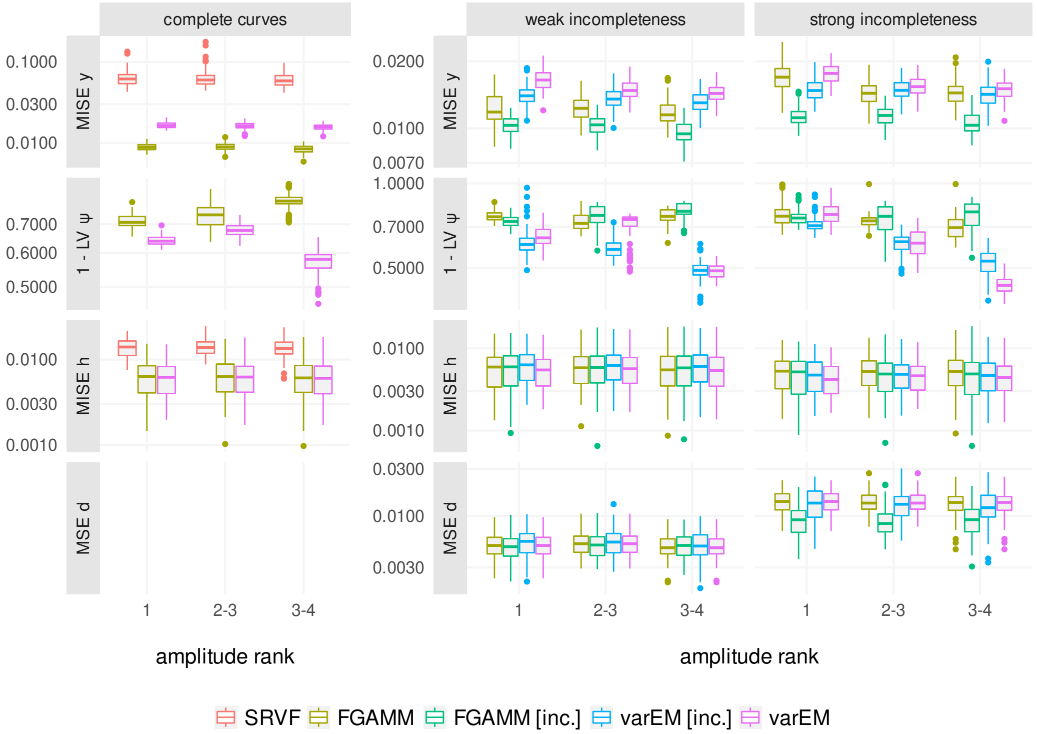

Fig. 9. Results for the simulation setting with Gamma data and a correlation between amplitude and phase.

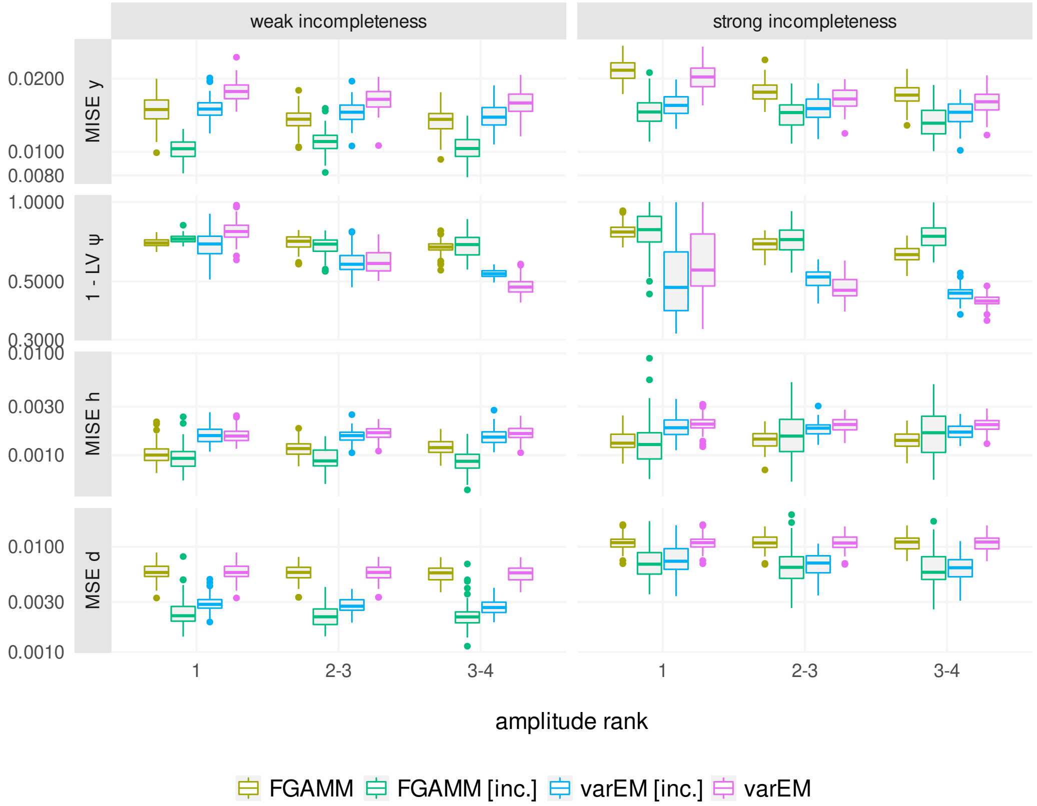

Fig. 10. Results for the simulation setting with Gamma data and a correlation between amplitude and the amount of incompleteness.

A3.4. Simulation results – Gaussian with adaptive FPC estimation

In contrast to the previous settings of the simulation study, the number of Functional Principal Components (FPCs) in the following settings is not fixed to the simulated rank of amplitude variation. Instead, in each iterative FPCA step (i) the varEM approach uses as many FPCs as are needed to explain 90% of the overall amplitude variation, and (ii) the FGAMM approach uses as many FPCs as are needed to explain 90% of the overall amplitude variation, while dropping such FPCs that explain $< 2\%$ of the variation (see criterion outlined at the end of Section 3).

For the varEM approach, the explained share of variance and accordingly the number of FPCs in each iteration is estimated before the main iteration's estimation step by once running the FPCA with 20 FPCs and correspondingly 20 B-spline basis functions to represent the FPC basis. Doing so, we approximate the overall variance in the varEM approach with the variance represented by this FPC basis with 20 FPCs. In contrast to the simulation results in the main part of our paper, we accordingly use 20 instead of eight basis functions for the estimation of the FPC basis in the varEM approach.

Note that the third and fourth FPC in the simulation settings with amplitude rank 2–3 and 3–4 only explain 5% and 10% of the overall amplitude variation, respectively (see Section 4.1). Accordingly, it is not unreasonable if fewer than 3 and 4 FPCs are chosen based on the $\geq 90\%$ criterion, respectively.

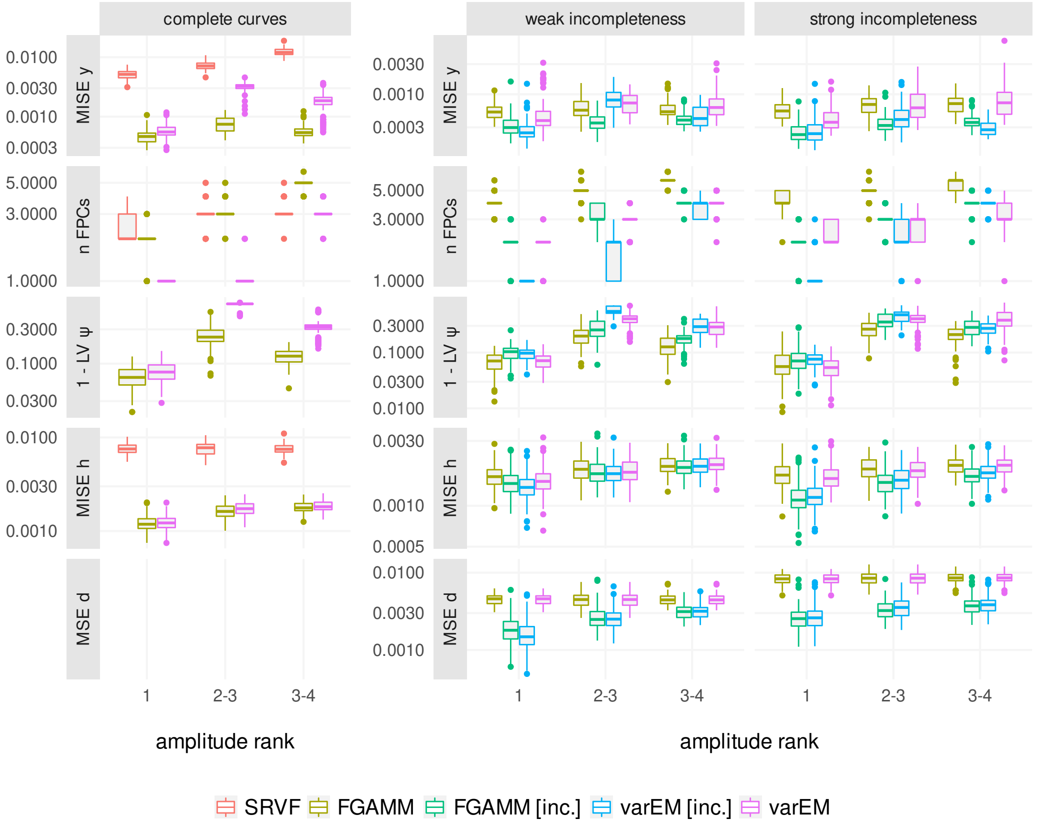

Fig. 11. Results for the simulation setting with Gaussian data and mutually uncorrelated amplitude, phase and amount of incompleteness, where the number of FPCs was adaptively estimated. All y scales are $\log_{10}$ transformed.

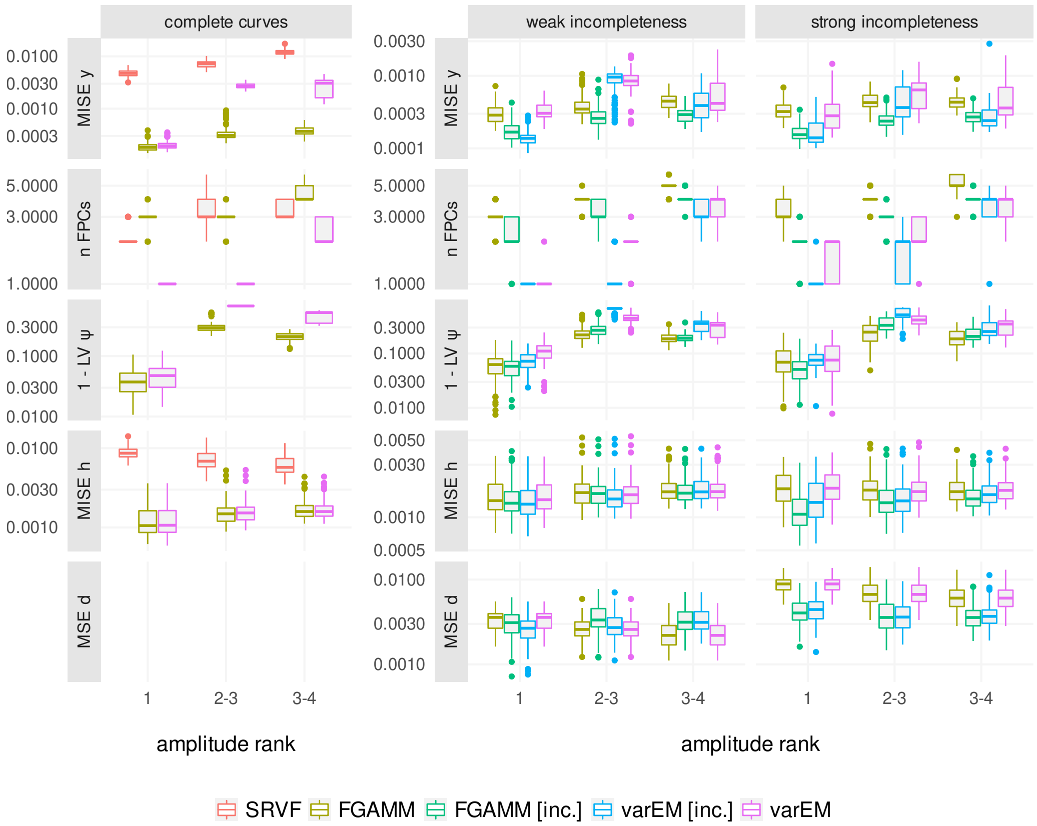

Fig. 12. Results for the simulation setting with Gaussian data and a correlation between amplitude and phase, where the number of FPCs was adaptively estimated.

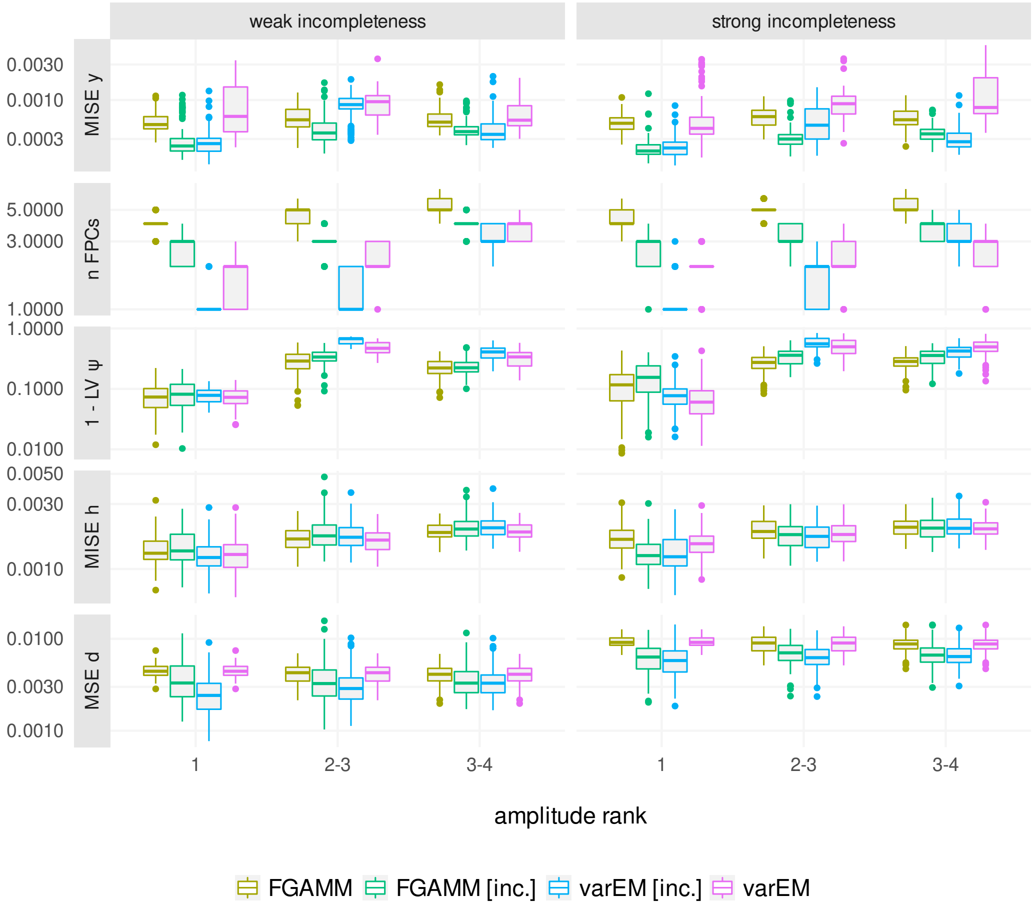

Fig. 13. Results for the simulation setting with Gaussian data and a correlation between amplitude and the amount of incompleteness, where the number of FPCs was adaptively estimated.

A3.5. Simulation results – Gamma with adaptive FPC estimation

Note our remarks at the beginning of Appendix A3.4. The only difference to the Gaussian setting is that the FGAMM approach assumes a Gamma distribution instead of a Gaussian structure.

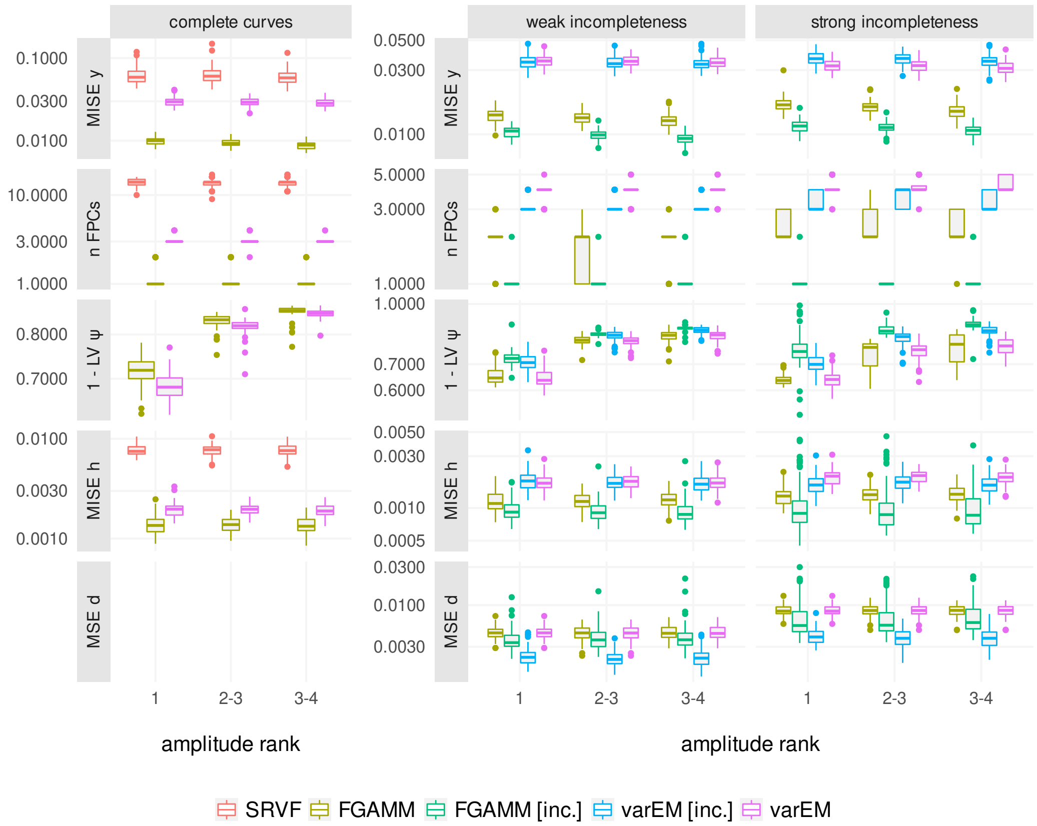

Fig. 14. Results for the simulation setting with Gamma data and mutually uncorrelated amplitude, phase and amount of incompleteness, where the number of FPCs was adaptively estimated.

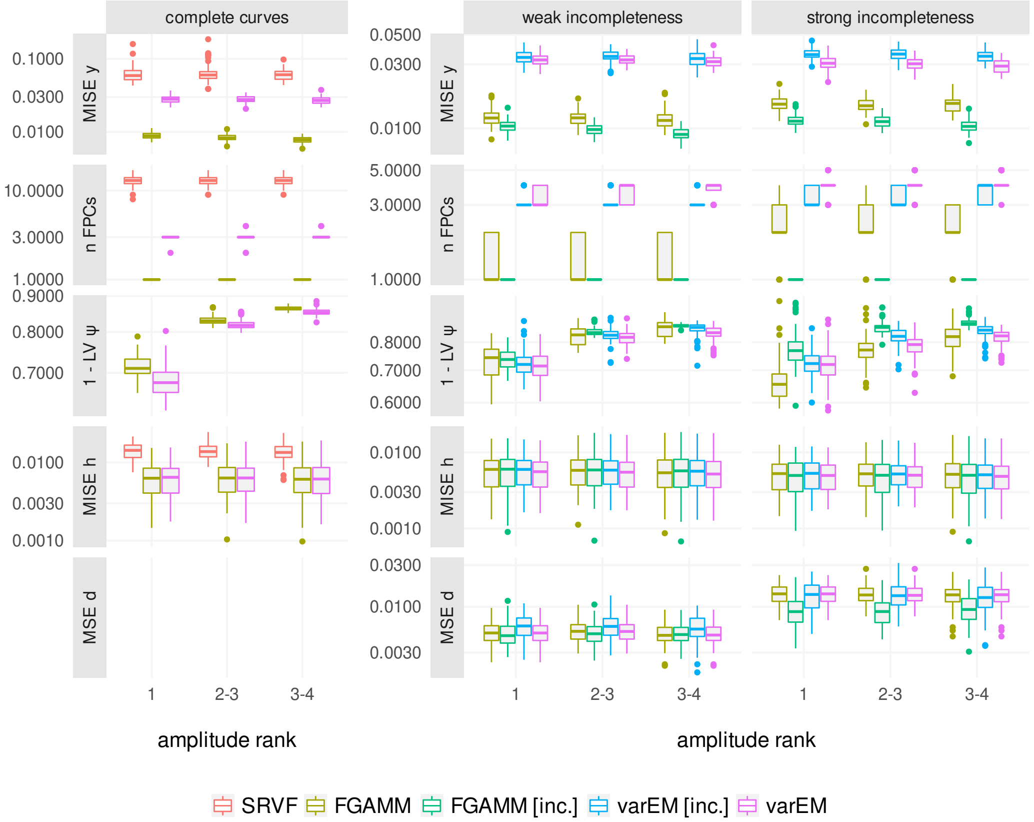

Fig. 15. Results for the simulation setting with Gamma data and a correlation between amplitude and phase, where the number of FPCs was adaptively estimated.

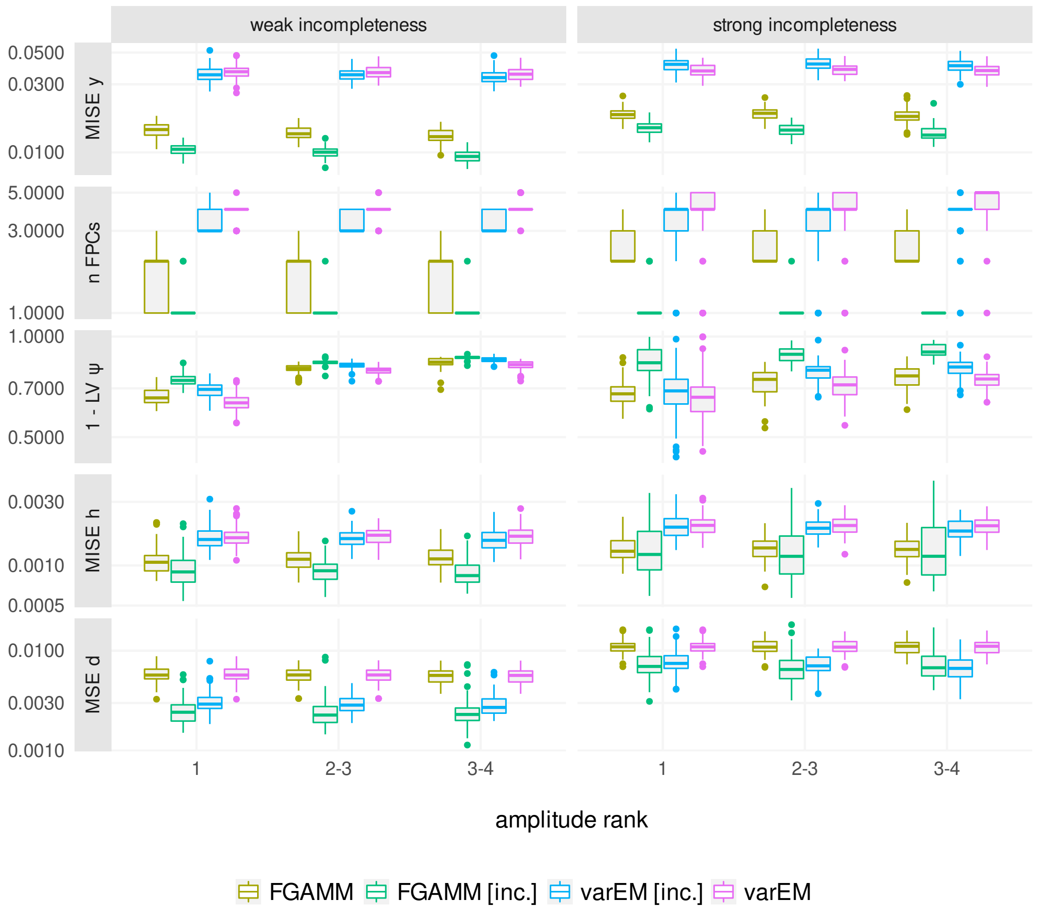

Fig. 16. Results for the simulation setting with Gamma data and a correlation between amplitude and the amount of incompleteness, where the number of FPCs was adaptively estimated.

A4. Berkeley application

A4.1. Curves with simulated incompleteness

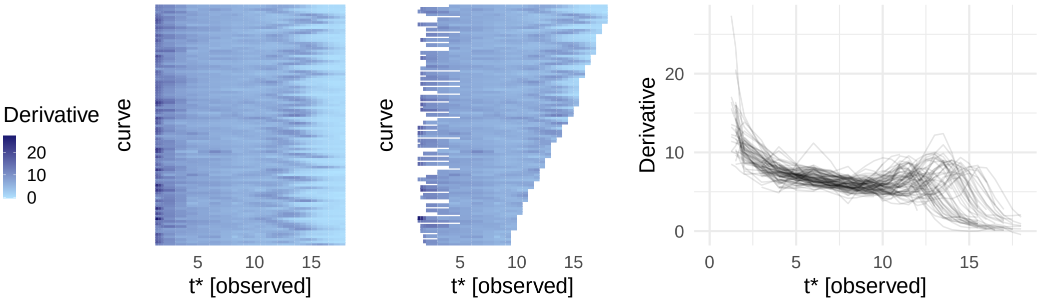

Fig. 17. Lasagna plot of observed curves (left pane) and curves with simulated incompleteness (right) for the first derivative of the Berkeley child growth data.

A4.2. Detailed results

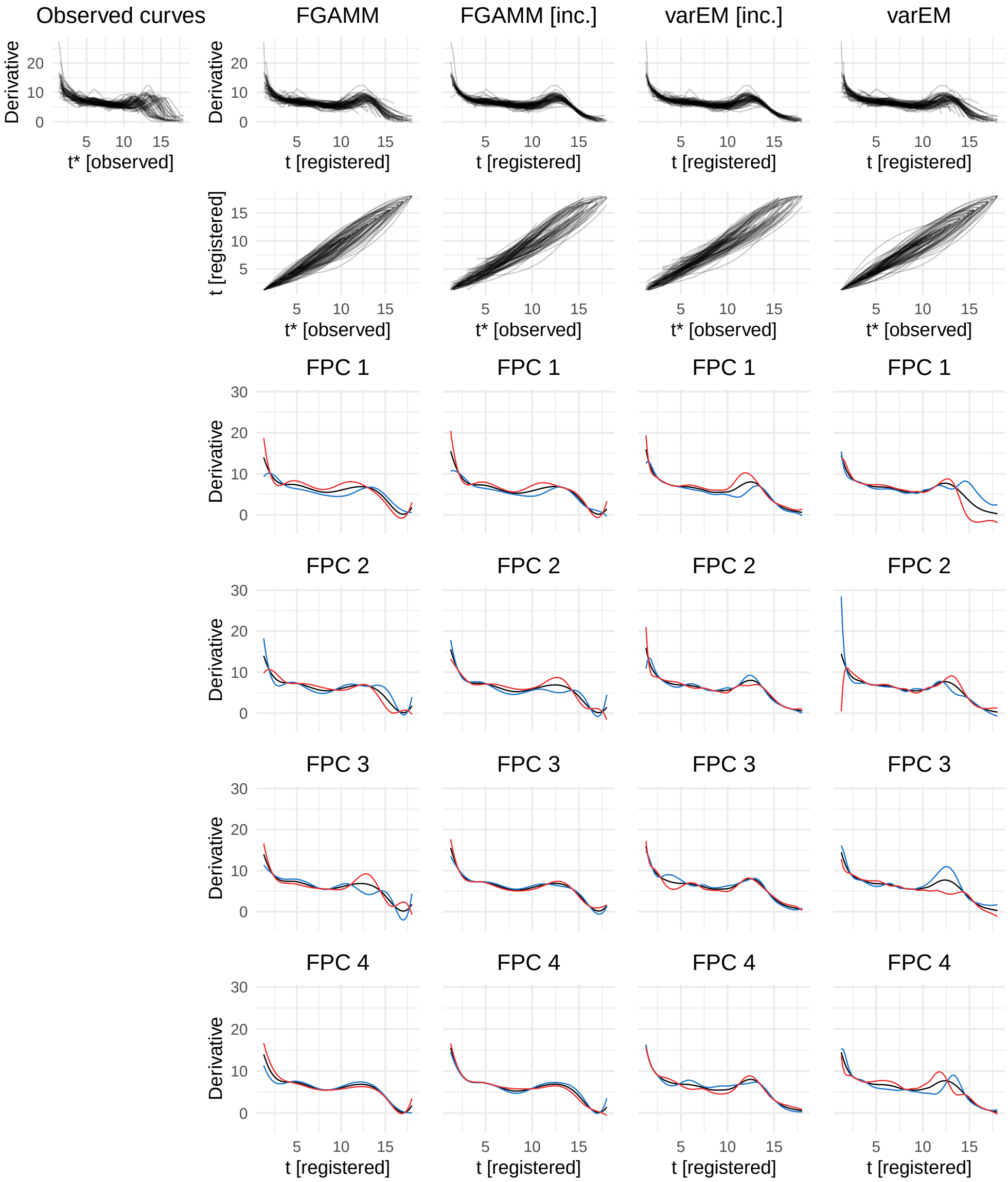

Fig. 18. Observed curves (top left pane), registered curves (top row), estimated inverse warping functions (second row) and the first four estimated FPCs based on the different approaches. The FPCs $\psi_k(t)$ are visualized by displaying the overall mean curve (solid line) plus (dashed line, $+$) and minus (dotted line, $-$) $x \cdot \psi_k(t)$, with x twice the standard deviation of the individual FPC's scores.

A4.3. Varying the penalization parameter λ

The results based on different λ values are shown in Figure 19. The example is based on the Berkeley data discussed in Section 5.

While the overall domain dilation of the warping functions is not penalized with $\lambda = 0$, this is the case the higher the penalization parameter λ is chosen. With value $\lambda = 1$ the penalization is strong enough to cause all registered domain lengths to be (quasi) identical to the observed domain lengths.

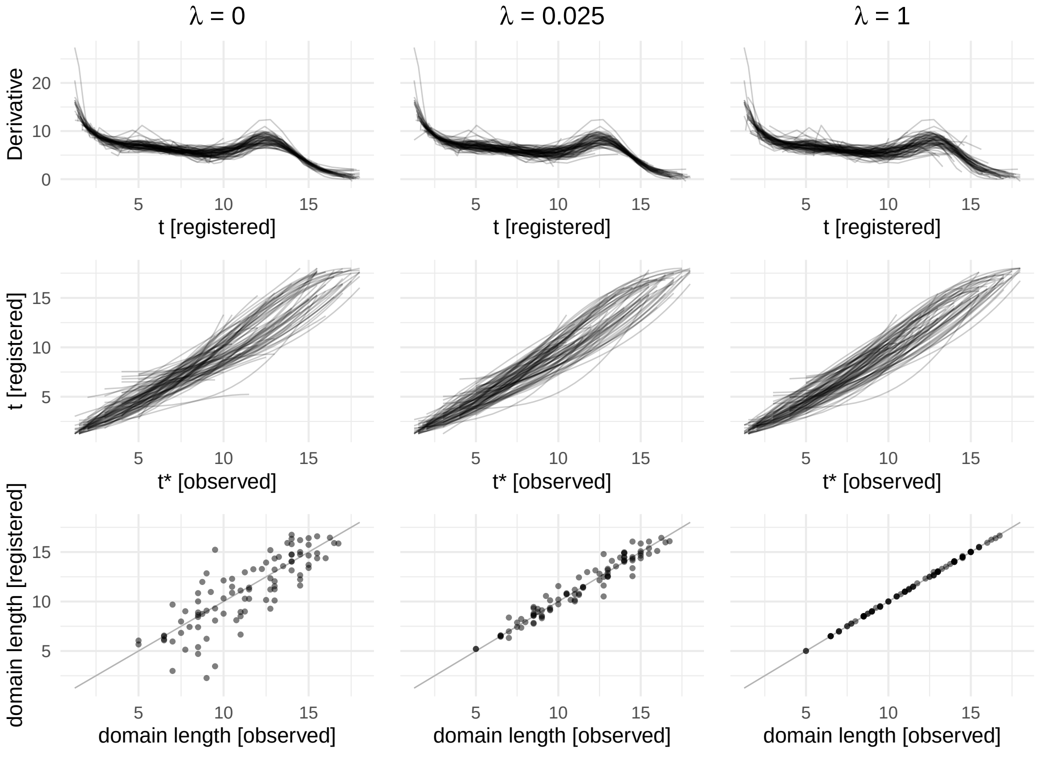

Fig. 19. Results for varying values of the penalization parameter λ after joint registration and Gaussian FPCA with the FGAMM approach. The graphic shows spaghetti plots of the registered curves (first row), estimated warping functions (second row) and the difference between the observed domain lengths and the registered domain lengths (bottom row).

A5. Seismic application

A5.1. Estimated inverse warping functions

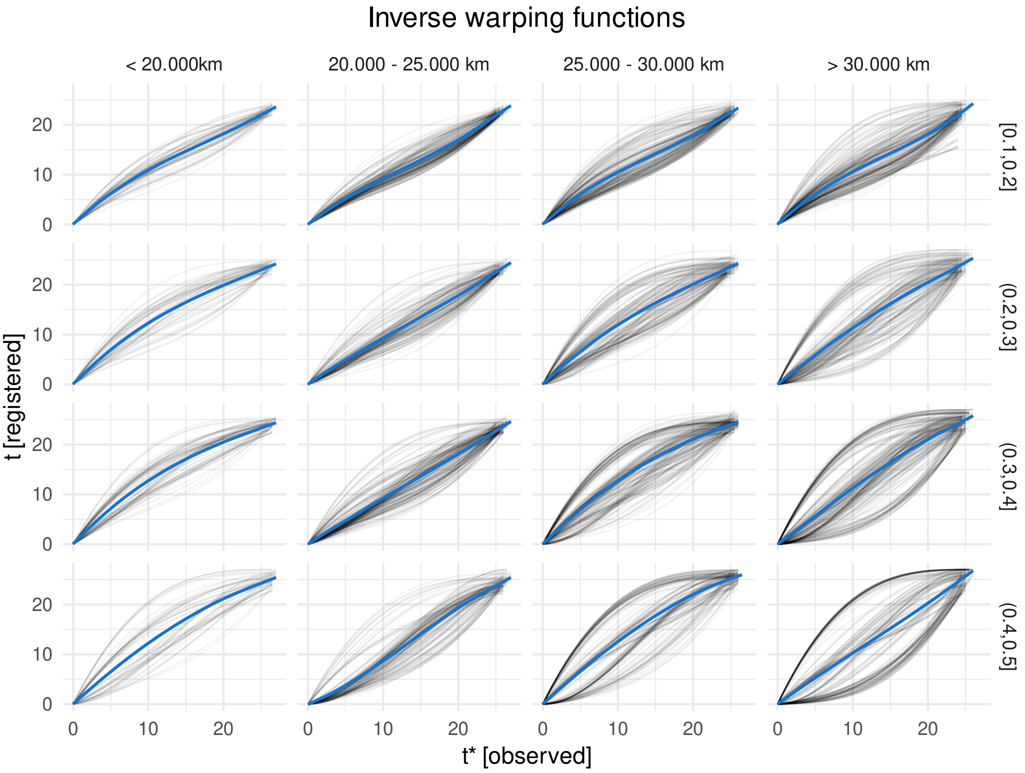

Fig. 20. Estimated inverse warping functions displayed against the hypocentral distance of the seismometers (x-axis) and the dynamic coefficient of friction μ_d (y-axis). In each panel, a solid blue curve marks the mean curve based on all respective warping functions.

A5.2. Estimated phase and amplitude variation over space

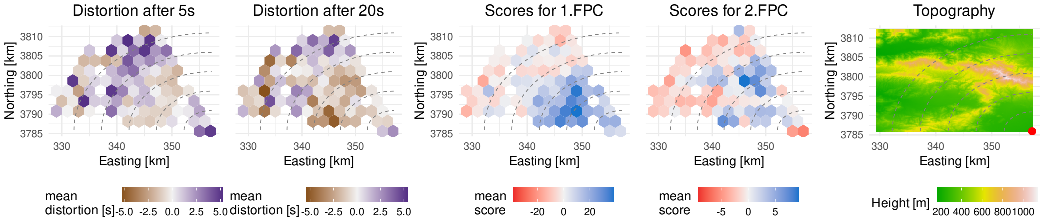

Fig. 21. Estimated phase and amplitude variation visualized over the evaluated region. Phase variation and amplitude variation are shown by displaying the mean of the overall time distortion after 5 and 20 seconds (left pane, with positive and negative values representing time dilation and compression, respectively) and of the curves' mean scores for the FPCs, respectively. The right plot shows the topography of the region, with the epicenter marked as red dot. The grey dashed lines mark the distance to the epicenter in 5km steps.

A6. GFPCA: Structure of subordinate FPCs

This section evaluates one Gaussian data setting from the simulation study to showcase the issue of subordinate functional principal components (FPCs) which

(a) individually explain a very small share of the overall amplitude variation, but jointly explain a relevant share, and

(b) often tend to represent phase variation rather than amplitude variation.

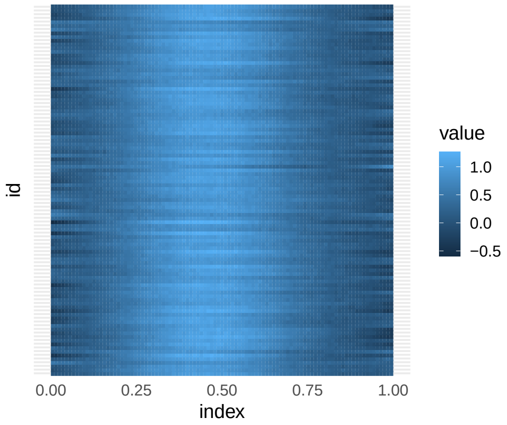

Fig. 22. Lasagna plot of the simulated curves.

For this evaluation, 100 curves are simulated similarly to the simulation setting with complete curves, Gaussian noise, amplitude rank 2–3 and no correlation between amplitude variation and phase variation. The only differences to the simulation study are the following:

- the curves are not randomly warped,
- a regular time grid with length 100 is used.

The simulated curves are visualized in Figure 22. We estimate a solution with 20 FPCs with the two-step approach. These first 20 FPCs and their explained shares of variance are visualized in Figure 23.

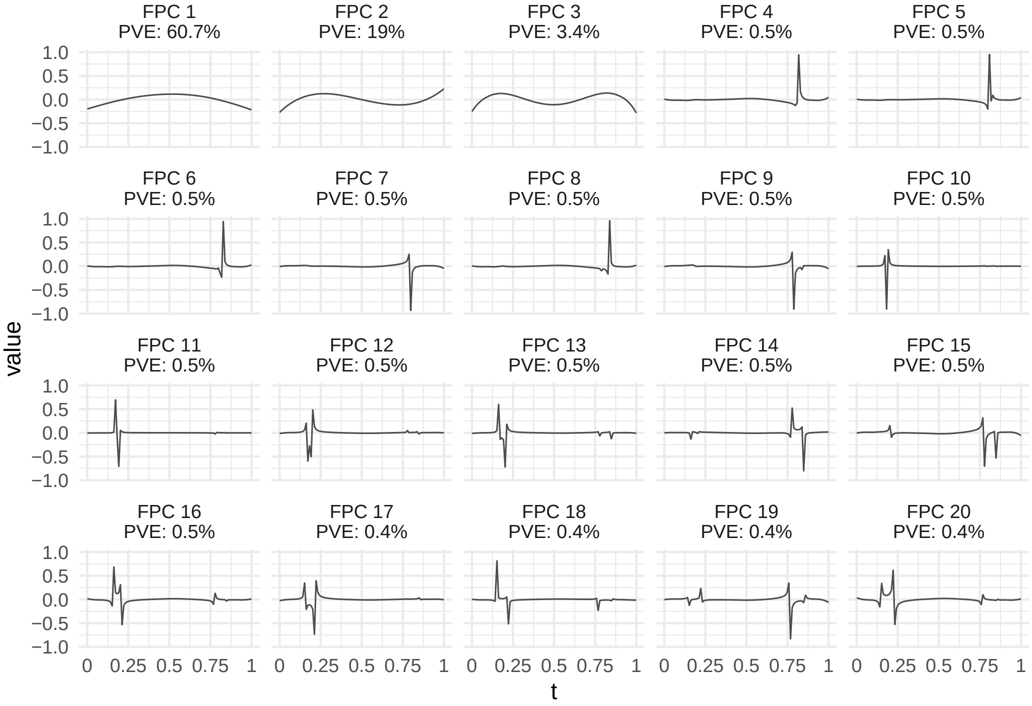

Fig. 23. Visualization of the first 20 FPCs including their percentage of explained variance (PVE).

A7. Choosing the initial template function

To check how much the results of the joint registration and GFPCA approach vary based on different template functions for the initial registration step, we run the application on the Berkeley data (from Section 5.1) with four different template functions:

- Template 1: Overall mean curve (similar to the application in the main paper)
- Template 2: Curve where the main peak in the second half of the domain is not very salient and occurrs quite early on
- Template 3: Curve where the main peak occurs a bit later on and is a bit more salient
- Template 4: Curve where the main peak is even more salient

The template functions and the results of the application of the FGAMM approach to the data can be found in the following Figure.

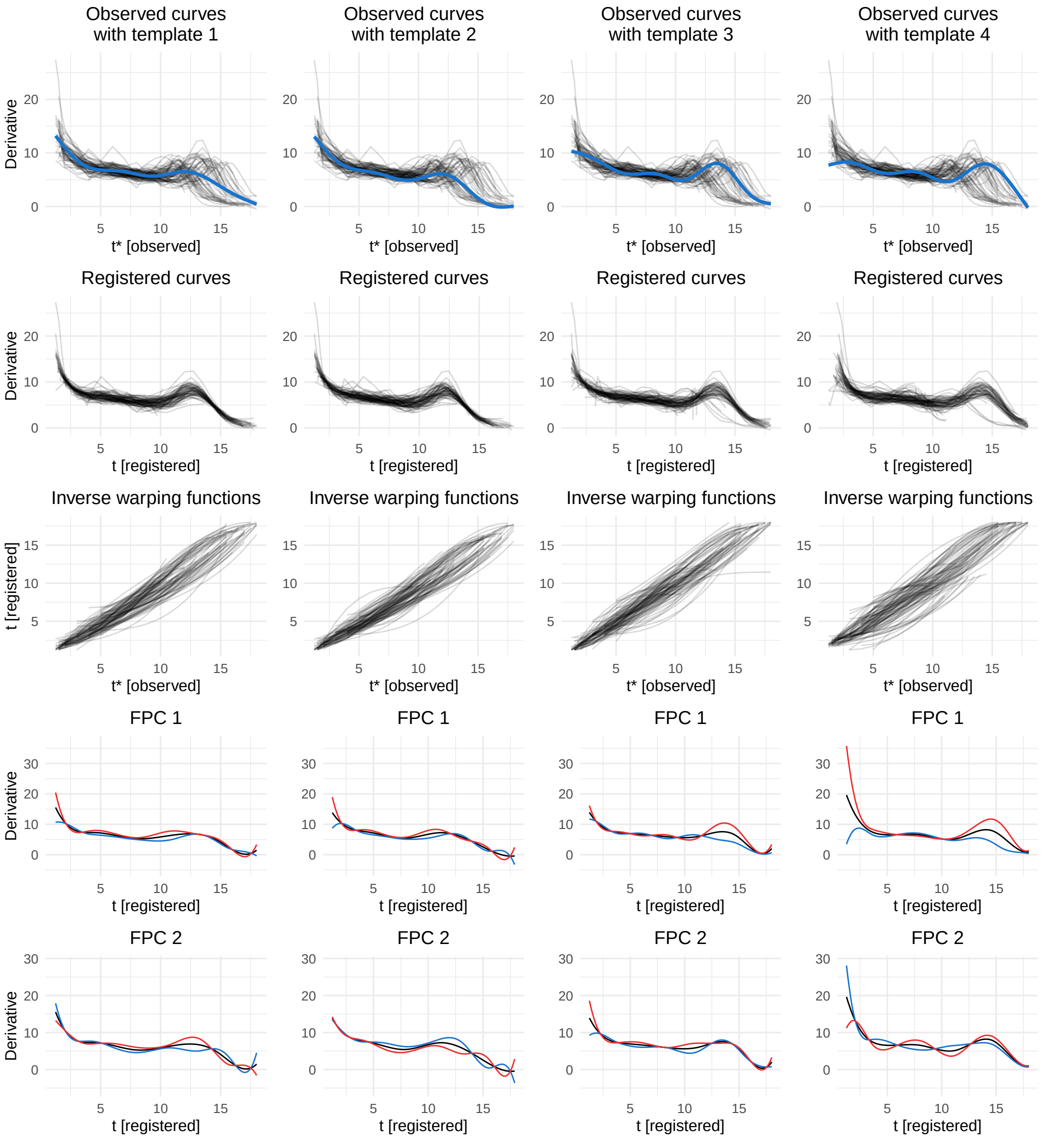

Fig. 24. Results of the FGAMM approach based on the different initial template functions (one column per template function). The rows contain the observed curves with the template function in blue (first row), the registered curves (second row), the estimated inverse warping functions (third row) and the first two estimated FPCs (last two rows). The FPCs $\psi_k(t)$ are visualized by the overall mean curve (solid line) plus (blue line) and minus (red line) $2 \cdot \sqrt{\hat{\tau}_k} \cdot \psi_k(t)$, with $\sqrt{\hat{\tau}_k}$ the standard deviation of the estimated scores for the k'th FPC.

References

Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**, 1–48.

Bates, D., Mullen, K. M., Nash, J. C. and Varadhan, R. (2014) *minqa: Derivative-free optimization algorithms by quadratic approximation.* URL: `https://CRAN.R-project.org/package=minqa`. R package version 1.2.4.

Gertheiss, J., Goldsmith, J. and Staicu, A.-M. (2017) A note on modeling sparse exponential-family functional response curves. *Computational statistics & data analysis*, **105**, 46–52.

Goldsmith, J. (2016) *gfpca: Generalized Functional Principal Components Analysis.* URL: `https://github.com/jeff-goldsmith/gfpca`. R package version 1.1.0. URL https://github.com/jeff-goldsmith/gfpca.

Johnson, S. G. (2020) *The NLopt nonlinear-optimization package.* URL: `https://github.com/stevengj/nlopt`. URL https://github.com/stevengj/nlopt.

Powell, M. J. (2009) The bobyqa algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge*, 26–46.

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/`. URL https://www.R-project.org/.

Wood, S. and Scheipl, F. (2020) *gamm4: Generalized Additive Mixed Models using 'mgcv' and 'lme4'.* URL: `https://CRAN.R-project.org/package=gamm4`. R package version 0.2-6. URL https://CRAN.R-project.org/package=gamm4.

Wood, S. N. (2017) *Generalized additive models: an introduction with R.* CRC press.

Wood, S. N., Li, Z., Shaddick, G. and Augustin, N. H. (2017) Generalized additive models for gigadata: modeling the uk black smoke network daily data. *Journal of the American Statistical Association*, **112**, 1199–1210.

Wrobel, J. (2018) register: Registration for exponential family functional data. *Journal of Open Source Software*, **3**, 557.

Wrobel, J. and Bauer, A. (2021) registr 2.0: Incomplete curve registration for exponential family functional data. *Journal of Open Source Software*, **6**, 2964.

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

Ypma, J. and Johnson, S. G. (2020) *nloptr.* R package version 1.2.2.2.

6. registr 2.0: Incomplete Curve Registration for Exponential Family Functional Data

Contributing article

Wrobel, J. and Bauer, A. (2021). registr 2.0: Incomplete Curve Registration for Exponential Family Functional Data. *Journal of Open Source Software*, 6(61): 2964.

Code repository

`https://github.com/julia-wrobel/registr`

Copyright information

Author contributions

The idea for the publication was brought up by Alexander Bauer. The paper is the successor of the JOSS publication for version 1.0 of the `registr` package (see Wrobel, 2018, as cited in this publication), in which Julia Wrobel outlined the basic functionalities of the `registr` package created by her. The methodological contribution of incomplete curve registration was introduced by Alexander Bauer, Fabian Scheipl and Helmut Küchenhoff, as outlined in the *Author contributions* section for the first contribution of this dissertation. Alexander Bauer wrote the codebase, extensive documentation and unit tests for all newly introduced methods and wrote the initial draft of the publication. Fabian Scheipl partly contributed to the codebase by making the covariance estimation with function `cov_hall` more efficient. Julia Wrobel extensively proofread the paper.

The Journal of Open Source Software

registr 2.0: Incomplete Curve Registration for Exponential Family Functional Data

Julia Wrobel[1] **and Alexander Bauer**[2]

1 Department of Biostatistics and Informatics, University of Colorado Denver, USA **2** Department of Statistics, LMU Munich, Germany

DOI: 10.21105/joss.02964

Software

- Review
- Repository
- Archive

Editor: Monica Bobra

Reviewers:

- @qpmnguyen
- @g4brielvs

Submitted: 23 December 2020
Published: 06 May 2021

Introduction

Functional data are observed in many different fields. Typical examples are longer-term panel studies where a sequence of measurements is observed for each subject. Compared to classical longitudinal studies, functional data analysis focuses more on the shapes of the (time-dependent) processes by analyzing the observed curve per subject. E.g., one can analyze the speed of growth of children until adulthood in the Berkeley child growth study (see left pane of Figure 1).

Functional data comprise different modes of variation. In the Berkeley study, not only can growth spurts be more or less pronounced regarding the actual growth (i.e., *amplitude variation* along the y-axis), but each spurt can also be shifted for some months / years for individual subjects (i.e., *phase variation* along the x-axis). Observed curves often have to be preprocessed with a *registration method* in order to separate phase and amplitude variation before analysis.

Most registration methods can only handle continuous data or data with a Gaussian structure. However, functional data are often non-Gaussian or even categorical. E.g., function values could be binary indicators representing physical (in)activity of patients over time (Wrobel et al., 2019). Moreover, most registration approaches are only applicable to completely observed curves that comprise the underlying process from its very start to its very end.

Basic routines for registering (Gaussian) data are implemented in R package Ramsay et al. (2020). Performing joint registration and clustering is possible with Parodi et al. (2015). The popular square-root velocity function (SRVF) framework for curve registration is implemented in Tucker (2020) for completely observed curves on a regular grid. Similar to our approach the package allows for registering all curves to similar shapes which can be well represented by some low-rank basis.

Exponential Family-based Registration

The `registr` package is based on the methods outlined in Wrobel et al. (2019). Registration is performed using a likelihood-based approach and estimates *inverse warping functions* $h_i^{-1}: t_i^* \mapsto t$ that map the observed time domain t_i^* for subject i to the common time domain t. The overall model is

$$E\left[Y_i\left(h_i^{-1}(t_i^*)\right)|h_i^{-1},\alpha(t),\mathbf{c}_i,\boldsymbol{\psi}(t)\right]=\mu_i(t),$$

$$g\left[\mu_i(t)\right]=\alpha(t)+\sum_{k=1}^{K}c_{ik}\psi_k(t),$$

The Journal of Open Source Software

with $Y_i\left(t_i^*\right)$ and $Y_i\left(h_i^{-1}(t_i^*)\right)$ the unregistered and registered curves, respectively, and $\mu_i(t)$ the estimated subject-specific means serving as template functions, i.e., the target for the registration. The assumed distribution with link function $g(\cdot)$ and this conditional expectation allow us to define a log-likelihood $\ell(i)$ for each observed function (see Wrobel et al., 2019). The subject-specific means $\mu_i(t)$ are expressed through a low-rank representation based on a population-level mean $\alpha(t)$ and a linear combination of population-level basis functions $\psi_k(t)$ and subject-specific scores $\boldsymbol{c}_i$, composed with a fixed link function $g(\cdot)$. We estimate this representation using a likelihood-based approach for generalized functional principal component analysis (GFPCA).

The overall model is estimated with function `register_fpca()`, which iterates between the estimation of warping functions (implemented in function `registr()`) and GFPCA estimation (functions `fpca_gauss()` or `bfpca()` for Gaussian or binomial data, respectively). This approach is consistent with earlier versions of the `registr` package (compare Wrobel, 2018).

In version 2.0, the package now includes the *two-step GFPCA* approach of Gertheiss et al. (2017) to handle further exponential family distributions. The respective implementation is based on the `gfpca` package of Goldsmith (2016). New distributions are supported both for registration and GFPCA. Furthermore, for the registration step, the individual template functions (to which each curve is mapped) can now be flexibly defined by the user with the argument `Y_template` in `registr()` and `register_fpca()`. This is of relevance since in many settings the overall mean of the unregistered curves is no reasonable template.

Incomplete Curve Registration

We extend the approach of Wrobel et al. (2019) to incomplete curves where the underlying process was either not observed from its very beginning (i.e., *leading incompleteness*) or until its very end (*trailing incompleteness*), or both (*full incompleteness*).

Since the underlying process is fully contained in the observed interval for complete curves, the first and last value of complete-curve warping functions lie on the diagonal line so that they preserve the overall domain. For incomplete curves, warping functions are estimated without this starting point and / or endpoint constraint.

However, fully removing these constraints can lead to extreme distortions of the time domain. We include a regularization term λ that penalizes the amount of domain dilation or compression performed by the inverse warping functions. Mathematically speaking, we add a penalization term to the log likelihood $\ell(i)$ for curve i. For a setting with full incompleteness this results in

$$\begin{aligned}
\ell_{\text{pen}}(i) &= \ell(i) - \lambda \cdot n_i \cdot \text{pen}(i),\\
\text{with} \quad \text{pen}(i) &= \left([\hat{h}_i^{-1}(t^*_{max,i}) - \hat{h}_i^{-1}(t^*_{min,i})] - [t^*_{max,i} - t^*_{min,i}]\right)^2,
\end{aligned}$$

where $t^*_{min,i}, t^*_{max,i}$ are the minimum / maximum of the observed time domain of curve i and $\hat{h}_i^{-1}(t^*_{min,i}), \hat{h}_i^{-1}(t^*_{max,i})$ the inverse warping function evaluated at this minimum / maximum. For leading incompleteness with $h_i^{-1}(t^*_{max,i}) = t^*_{max,i} \forall i$ this simplifies to $\text{pen}(i) = \left(\hat{h}_i^{-1}(t^*_{min,i}) - t^*_{min,i}\right)^2$, and for trailing incompleteness with $h_i^{-1}(t^*_{min,i}) = t^*_{min,i} \forall i$ to $\text{pen}(i) = \left(\hat{h}_i^{-1}(t^*_{max,i}) - t^*_{max,i}\right)^2$. The penalization term is scaled by the number of measurements n_i of curve i to ensure a similar impact of the penalization for curves with different numbers of measurements. In practical settings, λ has to be set manually to specify which kinds of warpings are deemed unrealistic and should be prevented. The choice of λ should be based on subject knowledge by comparing the registration results given different λ values.

In registr() and register_fpca() the type of incompleteness can be defined by argument incompleteness. Further details are given in the package vignette *incomplete_curves*. When applied to the Berkeley data with simulated full incompleteness, our approach leads to a reasonable registration as shown in Figure 1.

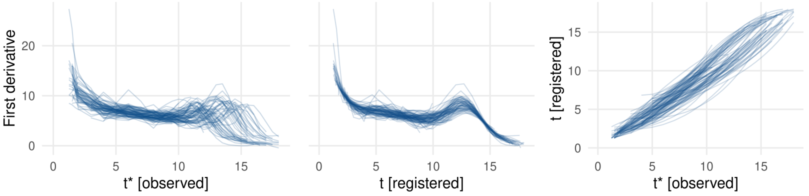

Figure 1: Left pane: Berkeley child growth data with simulated incompleteness; center: curves after registration; right: estimated inverse warping functions.

Acknowledgements

We thank Fabian Scheipl and Helmut Küchenhoff for valuable methodological contributions.

References

Gertheiss, J., Goldsmith, J., & Staicu, A.-M. (2017). A note on modeling sparse exponential-family functional response curves. *Computational Statistics & Data Analysis*, *105*, 46–52. https://doi.org/10.1016/j.csda.2016.07.010

Goldsmith, J. (2016). *gfpca: Generalized functional principal components analysis*. https://github.com/jeff-goldsmith/gfpca

Parodi, A., Patriarca, M., Sangalli, L., Secchi, P., Vantini, S., & Vitelli, V. (2015). *fdakma: Functional data analysis: K-mean alignment*. https://CRAN.R-project.org/package=fdakma

Ramsay, J. O., Graves, S., & Hooker, G. (2020). *fda: Functional data analysis*. https://CRAN.R-project.org/package=fda

Tucker, J. D. (2020). *fdasrvf: Elastic functional data analysis*. https://CRAN.R-project.org/package=fdasrvf

Wrobel, J. (2018). Register: Registration for exponential family functional data. *Journal of Open Source Software*, *3*(22), 557. https://doi.org/10.21105/joss.00557

Wrobel, J., Zipunnikov, V., Schrack, J., & Goldsmith, J. (2019). Registration for exponential family functional data. *Biometrics*, *75*(1), 48–57. https://doi.org/10.1111/biom.12963

Part III.

Function-on-Scalar Regression

7. An Introduction to Semiparametric Function-on-Scalar Regression

Contributing article

Bauer, A., Scheipl, F., Küchenhoff, H., and Gabriel, A.-A. (2018). An introduction to semiparametric function-on-scalar regression. *Statistical Modelling*, 18(3–4): 346–364.

Code repository

https://github.com/bauer-alex/FoSIntro

Copyright information

Author contributions

The concept of the publication was jointly developed by Alexander Bauer, Fabian Scheipl and Helmut Küchenhoff. Alexander Bauer analyzed the seismic data and wrote the complete codebase as well as the initial draft for the major part of the publication. Fabian Scheipl was closely involved in all parts, wrote parts of Chapters 4.3 and 5 himself and heavily revised the overall linguistic style. Helmut Küchenhoff was closely involved in refining the underlying idea and the style of the publication. Fabian Scheipl and Helmut Küchenhoff performed extensive proofreading of the paper. Alice-Agnes Gabriel created the seismic data, was closely involved in their analysis and proofread the respective parts of the paper.

Statistical Modelling 2018; **18(3–4)**: 346–364

An introduction to semiparametric function-on-scalar regression

Alexander Bauer[1], Fabian Scheipl[1], Helmut Küchenhoff[1] and Alice-Agnes Gabriel[2]
[1]Department of Statistics, Ludwig-Maximilians-Universität, Munich, Germany.
[2]Department of Geophysics, Ludwig-Maximilians-Universität, Munich, Germany.

Abstract: Function-on-scalar regression models feature a function over some domain as the response while the regressors are scalars. Collections of time series as well as 2D or 3D images can be considered as functional responses. We provide a hands-on introduction for a flexible semiparametric approach for function-on-scalar regression, using spatially referenced time series of ground velocity measurements from large-scale simulated earthquake data as a running example. We discuss important practical considerations and challenges in the modelling process and outline best practices. The outline of our approach is complemented by comprehensive R code, freely available in the online appendix. This text is aimed at analysts with a working knowledge of generalized regression models and penalized splines.

Key words: Functional regression, functional response, generalized additive model, semiparametric regression, penalized splines, geophysics

Received June 2017; **revised** October 2017; **accepted** October 2017

1 Introduction

Regression models for functional responses try to model structures like time-dependent processes or 2D or 3D images (Ramsay and Silverman, 2005). Functional data are thereby defined as data that vary over a specific domain $\mathcal{T}$, for example, time. Observations typically consist of measurements at individual points over that domain.

One valid alternative to functional response regression for data structured like this is longitudinal data analysis, modelling the separate measurements along each function using scalar regression while explicitly specifying their temporal correlation structure, for example, by including (random) time effects or by assuming autocorrelated residuals over time. However, eliciting an appropriate correlation structure is usually non-trivial. Using functional regression, correlation structures over the functional domain can be modelled flexibly and implicitly.

A functional approach should be the method of choice if the shape of a response over its functional domain is of main interest. Functional regression models enable researchers to quantify how various parameters influence the expected level and shape of the functional responses.

Address for correspondence: Alexander Bauer, Department of Statistics, Ludwig-Maximilians-Universität, Ludwigstr. 33, DE–80539 München, Germany.
E-mail: alexander.bauer@stat.uni-muenchen.de

10.1177/1471082X17748034

If the response is of a functional nature and all predictor variables are constant over the functional responses' domain, the corresponding model is a function-on-scalar regression model. This work gives an introduction to this model class aimed at researchers looking for a pragmatic overview on how to apply this method without having to dive deeper into the technical part of it. As such, our focus is on explaining general concepts rather than providing detailed mathematical explanations of the method. Furthermore, we list important practical considerations and give advice on which methods are needed in which situation. Throughout the text, we show how to apply the methods using real-world data.

Various approaches to model function-on-scalar data exist. Our main focus lies on the flexible framework of Greven and Scheipl (2017a) which covers models of the form

$$\begin{aligned} Y_i(t)|\boldsymbol{\mathcal{X}}_i &\sim F(\mu_{it}, \boldsymbol{\nu}) \\ g(\mu_{it}) &= \beta_0(t) + \sum_{r=1}^{R} f_r(\boldsymbol{\mathcal{X}}_{ri}, t). \end{aligned} \tag{1.1}$$

For all observational units $i = 1, \ldots, n$, the functional response, evaluated at specific points t of the functional domain, is assumed to come from some given distribution F with conditional expectation $\mu_{it} = \mathbb{E}(Y_i(t)|\boldsymbol{\mathcal{X}}_i)$ and dispersion and shape parameters $\boldsymbol{\nu}$. The expectation is connected to an additive predictor with a functional intercept $\beta_0(t)$ and R potentially nonlinear covariate effects $f_r(\cdot)$ by a pre-specified link function $g(\cdot)$. The covariate effects $f_r(\cdot)$ each depend on a subset $\boldsymbol{\mathcal{X}}_r$ of the covariate set $\boldsymbol{\mathcal{X}}$ and can potentially vary over the functional domain $\mathcal{T}$. More specifically, we refer to $\mathcal{T}$ as the *time domain*, as this is the functional domain in our running example. All methods, however, are also applicable for other functional domains.

Well-written introductions to the basic concepts and philosophy of functional data analysis are given in Ramsay and Silverman (2005) and Ramsay et al. (2009). Reviews of current research can be found in Morris (2015) and Wang et al. (2015). Readers interested in an in-depth review of available implementations for function-on-function and scalar-on-function regression models are pointed to Greven and Scheipl (2017a). An alternative approach that is closely related to the approach used here was developed by Reiss et al. (2010).

We perform our analyses in `R` (R Core Team, 2016, v. 3.3.2) using the function `pffr` from the package `refund` (Goldsmith et al., 2016), which is based on the `gam` function for scalar additive regression from the `mgcv` package (Wood, 2006, v. 1.8-15). The `refund` package is a flexible and fully documented package for functional data analysis. This article is accompanied by the open source `R` package `FoSIntro` (Bauer, 2017), available on GitHub, which comprises several convenience functions for the work with function-on-scalar models based on `pffr`. The GitHub repository also contains code showing how to apply all methods shown in this article.

The article is structured as follows: Section 2 introduces the running example for this work. Important statistical aspects of semiparametric regression are sketched in Section 3. Section 4 discusses concepts and challenges of function-on-scalar regression. We finish with a discussion and outlook in Section 5.

If the main interest lies in predicting or analysing specific characteristics of the functional response, alternative approaches are often more adequate. In particular, the function-on-scalar regression approach presented here is not well suited for predicting *peak* ground velocities as the penalized estimation of smooth structures tends to systematically underestimate the maxima.

2 Application to seismic ground motion data

Bauer (2016) used function-on-scalar regression to quantify how frictional failure across an earthquake fault affects ground velocities at different distances from the earthquake's hypocentre over time. Figure 1's left panel shows three typical observations of the functional response *ground velocity* over the functional domain *time*. All covariates in the study were constant over time.

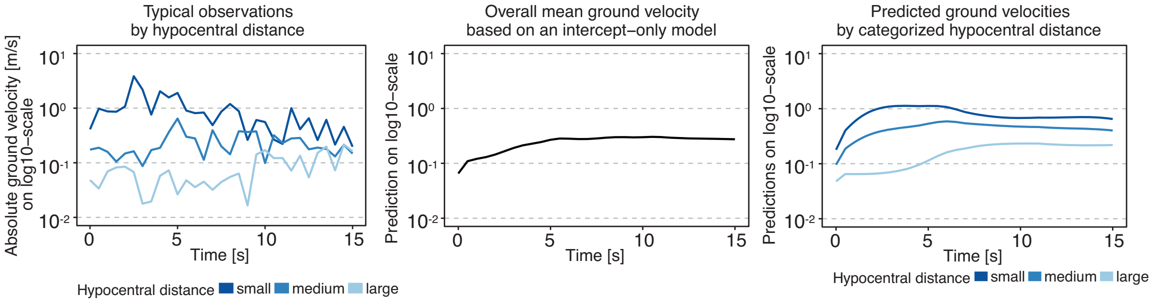

Figure 1 Left: Typical observations of absolute ground velocity over time. Peak ground velocity is delayed and decreases as the hypocentral distance increases. Middle: Overall functional mean of the ground velocities based on model (3.1) which only contains the intercept. Right: Estimated mean ground velocities by categorized hypocentral distance, based on model (3.2)

The aim of statistical modelling is to gain a better understanding of the associations between initial seismic conditions like fault stress and fault strength prior to earthquakes as well as local topography and geology with the temporal and spatial distribution of ground movement caused by an earthquake. The data is derived from large-scale *in silico* earthquake scenario simulations with the open source software SeisSol (Breuer et al., 2014; Pelties et al., 2014 www.seissol.org), based on a real seismic event that took place in Northridge (California) in 1994. Multiple simulations with varying initial conditions are analysed.

Shaking velocity and ground movement was recorded in high temporal resolution at a dense network of virtual seismometers distributed across Southern California. In the notation of (1.1), each response function $Y_i(t)$ represents the first 15s of the absolute ground velocity measurements from one of 75 such simulations for a given virtual seismometer i in a resolution of 2Hz. A subset of 260 seismometers was used for the analysis. Leading zeros were discarded up to the first relevant ground movement ($Y_i(t) \geq 0.01$) in order to remove irrelevant phase variation as described on p. 21. In keeping with the introductory level of this text, we only look at a submodel of Bauer (2016) and omit most seismological details.

The analysis is focused on the effects of five physical parameters on ground velocity: three frictional resistance variables, the direction of the regional tectonic background stress and the soil material of the simulated area: either rock or sediment. These parameters were pre-set in each seismic simulation. As seen in Figure 1, distance from the fault has an important effect on both the shape and the level as well.

3 Basic concepts of semiparametric modelling

The regression framework introduced by Greven and Scheipl (2017a) is based on additive or semiparametric regression models. Such models are one approach for estimating nonlinear effects of variables. In the following, we will introduce the basic modelling concepts of semiparametric functional regression by practically motivating differently complex models, each followed by a brief summary of the most important methodological basics.

3.1 Semiparametric models with one-dimensional smooth effects

In the simplest setting, we estimate the overall functional mean of ground velocities using a model only containing a functional intercept and no covariates:

$$g(\mathbb{E}(Y_i(t))) = \beta_0(t). \tag{3.1}$$

As the response in our application is strictly positive, we assume a Gamma distribution with a log link function $g(\cdot)$ in all examples throughout this article. Figure 1's middle panel shows this overall estimated functional mean for Equation (3.1). It can be seen that the overall mean is increasing over the first few seconds until it reaches a constant level.

As a next step, we want to assess a possible association of ground velocities with hypocentral distance, that is, we also want to quantify just *how different* curves at different hypocentral distances are on average. This can be done by extending Equation (3.1) with a dummy-coded categorical covariate x for grouped hypocentral distance

$$g(\mathbb{E}(Y_i(t)|x_i)) = \beta_0(t) + \beta_1(t)I_{\text{medium}}(x_i) + \beta_2(t)I_{\text{large}}(x_i), \tag{3.2}$$

where $I_{\text{medium}}(x_i)$ is 1 if the hypocentral distance x_i of the station where observation $Y_i(t)$ was recorded is intermediate, and 0 otherwise. Interpretation of categorical effects is equivalent to scalar regression, meaning that each effect $\beta_1(t)$ and $\beta_2(t)$ quantifies a deviation from the reference category 'small distance'. As can be seen in Figure 1's right panel, the estimated effects in Equation (3.2) show relevant differences in their level and shape.

3.2 Estimating one-dimensional smooth effects

Estimation of the functional intercept and the time-varying distance category effects is performed using a spline-based approach, where the effect is represented as the sum of scaled spline basis functions. Readers not familiar with this and other basic concepts regarding *penalized estimation* for *generalized additive models* are pointed to Fahrmeir et al. (2013) or Wood (2006). In a nutshell, penalization is a useful tool in estimating smooth effects as it allows estimation of nonlinear effects simply by defining the *maximally possible wiggliness* of each effect's shape, which is limited by the number of spline basis functions being used for that effect. Overfitting is then prevented by using an estimation criterion that punishes complexity of the effect estimates (i.e., wigglier shapes) while simultaneously rewarding goodness of fit. Parameters that control the relative weights in this trade-off between a good fit of the training data on one hand and a parsimonious model with simple effect shapes that is more likely to generalize well for previously unseen test data on the other hand are estimated from the data automatically.

Many different spline bases are available for one-dimensional smooth effects, cf. the documentation for `mgcv`. P-splines (Eilers and Marx, 1996) with second order difference penalties as well as thin plate regression splines (TPRS, see Wood (2003), based on Duchon (1977)) correspond to a weak prior assumption of linear effects. By default, `pffr` uses cubic P-splines with first order differences over the functional responses' domain. This yields smooth effects and corresponds to a weak prior assumption of effects being constant over $\mathcal{T}$. TPRS bases often perform slightly better than P-splines (Wood, 2003), but also suffer from numerical problems in some situations and are much more computationally expensive to set up.

Some more specialized spline bases are very useful in particular situations and easily available in the software we use here, for example, cyclic splines for periodic effects where boundary values must be equal or soap film smooths for fits with constraints along complex domain boundaries like seashores. Morris (2017) compares a Bayesian wavelet-based approach well suited for spiky data on regular grids to the method described here.

Using the spline-based approach, both the estimation of time-varying effects and of effects that vary nonlinearly over the variable domain itself is possible. An example for the latter is given in Figure 2's left panel, which shows the estimated effect of the variable *slip weakening*, that is, the distance over which initial friction diminishes to its minimum. Higher values in this parameter correspond to bigger overall friction and thus to ground velocity curves that have a lower level overall. Note that this type of time-constant effect does not affect the shape of the functional responses, only their overall level.

Putting all currently described effect shapes together, we are now capable of specifying models of the form

$$g(\mathbb{E}(Y_i(t)|\boldsymbol{\mathcal{X}}_i)) = \beta_0(t) + \sum_{j=1}^{J} \beta_j(t)x_{ji} + \sum_{k=1}^{K} f_k(x_{ki}), \tag{3.3}$$

which include J time-varying linear effects $\beta_j(t)$ as well as K smooth effects $f_k(\cdot)$ which are time-constant, but vary over the respective variables x_k.

3.3 Semiparametric models with multidimensional smooth effects

As a final step, we now include multidimensional smooth effects into the model. Such effects can vary nonlinearly both over the domain of the functional response and the domain of the covariate (or multiple covariate domains in the case of interaction effects). As an example, the three rightmost panels of Figure 2 visualize the estimated nonlinear time-varying effect of the hypocentral distance. To facilitate interpretation of the effect, it is shown using both a heatmap (panel 3) as well as a 3D surface (panel 4). In addition, a comparison of the predictions for specific values is a valuable tool as well (panel 2). One can see (a) that smaller hypocentral distances correspond to higher ground velocities (note the large negative slope of the estimated surface along the distance axis), (b) that the initial peak of ground velocity sets in later the farther away from the earthquake centre the virtual seismometers are located (note that the peak for a given distance is located higher up the time axis as distance increases), (c) that the peak becomes somewhat less pronounced for larger distances and (d) that the effect is almost linear over hypocentral distance.

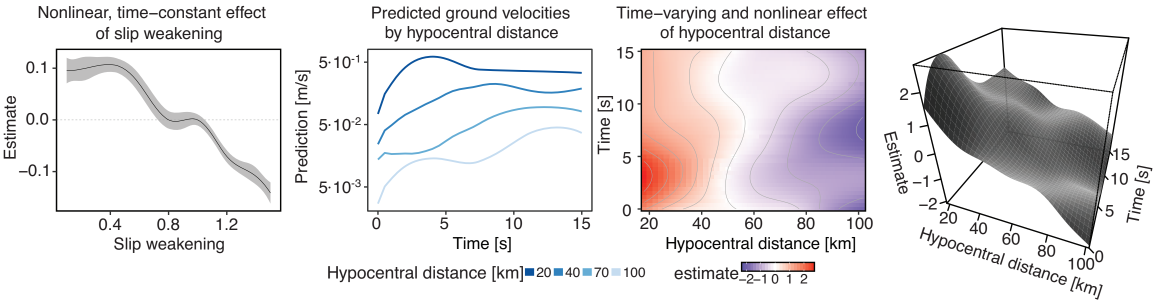

Figure 2 Panel *1*: Estimated time-constant effect of slip weakening $f_{\text{slip}}(x_{\text{slip}})$, which implies a (nonlinear) shift in the average level of $Y_i(t)$ as x_{slip} changes. *2*: Predictions for specific values of hypocentral distance with remaining covariates set to realistic values. *3, 4*: Effect of hypocentral distance visualized using a heatmap and a 3D surface. Note that values in panels 1, 3 and 4 are on the scale of the additive predictor (i.e., $\log_e$([m/s])), while panel 2 is on $\log_{10}$-scale

3.4 Estimating multidimensional smooth effects

Incorporation of multidimensional smooths into Equation (3.3) is easily done by generalizing it to

$$g(\mathbb{E}(Y_i(t)|\boldsymbol{\mathcal{X}}_i)) = \beta_0(t) + \sum_{r=1}^{R} f_r(\boldsymbol{\mathcal{X}}_{ri}, t). \qquad (3.4)$$

In addition to the functional intercept we now have R covariate effects $f_r(\boldsymbol{\mathcal{X}}_{ri}, t)$ which potentially vary over both covariate domains and the functional domain $\mathcal{T}$. We write $\boldsymbol{\mathcal{X}}_{ri}$ instead of x_{ri} to emphasize that each smooth potentially depends on multiple covariates, thereby covering linear interactions terms and multidimensional smooths. Note that $f_r(\cdot)$ can obviously also be a time-constant, linear effect.

We briefly sketch two possibilities for setting up a multidimensional spline basis for representing effects $f_r(\boldsymbol{\mathcal{X}}_{ri}, t)$. *Tensor product spline bases* are created by setting up an adequate marginal one-dimensional basis for each dimension of the effect (e.g., hypocentral distance and the time domain) and then taking the *Kronecker product* of the marginal bases (i.e., multiplying each basis function of each marginal dimension with all basis functions of all other marginal dimensions). This results in a multivariate spline basis defined on the joint domain of all involved covariates (and time). A major advantage of this method is its large flexibility as the appropriate marginal bases and penalties can be chosen freely to suit the problem. Since penalization of such tensor product spline terms is done separately for each dimension, this also allows for different roughnesses of the various marginal dimensions (e.g., an effect $f(x_r, t)$ that is very smooth over some covariate x_r but still wiggly over time t). A disadvantage of tensor product splines is that tensor basis functions are defined on a regular grid over the joint domain and some basis functions may lie in regions where there are not many or no data points at all, leading to computational inefficiencies and badly conditioned model fits. An alternative to tensor product spline bases are multidimensional TPRS, a direct generalization of the one-dimensional TPRS basis. The most important difference is that TPRS basis functions imply identical roughness in all directions. In practice, this only makes sense if marginal variables are on comparable scales, for example, in a 2D spatial effect with longitude and latitude as the marginal covariates.

3.5 Some practical considerations

Since the number of basis functions limits the maximal complexity of the shape of any effect $f_r(\boldsymbol{\mathcal{X}}_{ri}, t)$, it needs to be sufficiently large. Which number to choose initially depends greatly on the data situation and it is very difficult to provide general advice. For most applications, 20-30 basis functions for a one-dimensional effect will typically be sufficient, but this is feasible only if enough observations are available for estimation. In situations with fewer data points or simple effect shapes, however, it can also be appropriate to use only 5 or 10 basis functions initially. After estimating the model, the effective degrees of freedom (edf; see Wood, 2006, Ch. 4.4) of each term give an indication of whether the amount of flexibility was sufficient or not. If the edf are near their maximum, the model should be re-estimated using a larger number of basis functions. In this case, a larger basis that is expressive enough for the effect's true complexity can improve the estimate. An automated approach for checking adequacy of the chosen basis dimension was introduced by Pya and Wood (2016) and is implemented in the `gam.check` function of `R` package `mgcv`.

In some situations, the placement of the basis functions over the effect's domain can be of great importance. If no further information is available an equidistant placement is a valid approach. In contrast, a user-specified placement can make sense if the data are spread very unequally across the domain and the researcher supposes that the effect will vary more strongly in regions where more data points lie or when a few data points lie far beyond the main data cloud. In such cases, it can be more efficient to place more knots in regions with more data, especially in situations with small to moderate sample size.

4 Inference and model checking

This section focuses on setting up and evaluating a function-on-scalar model. As motivated in the last section, the general model of Greven and Scheipl (2017a) can be written as

$$g(\mathbb{E}(Y_i(t)|\boldsymbol{\mathcal{X}}_i, E_i(t))) = \beta_0(t) + \sum_{r=1}^{R} f_r(\boldsymbol{\mathcal{X}}_{ri}, t) + E_i(t), \tag{4.1}$$

where the conditional expectation of the response $Y_i(t)$ is modelled by R potentially nonlinear effects (as defined on p. 10) and a functional intercept $\beta_0(t)$. The newly introduced term $E_i(t)$ specifies functional error terms. Those smooth errors are estimated as curve-specific functional random intercepts and can be used to incorporate possible autocorrelation and variance heterogeneity along the functional domain (Scheipl et al., 2015) as motivated in the next paragraph. The additive predictor is mapped to the domain of the functional responses by a given link function $g(\cdot)$, which for the Gamma-model in our application example is simply the natural logarithm. Note that the interpretation of effects in models including (functional) random effects like $E_i(t)$ is generally conditional, not marginal, similar to conditional GLMMs (Diggle et al., 2002): the estimates quantify the expected change in *individual* conditionally expected values, not in population averages. This distinction is meaningless if the link function $g(\cdot)$ is the identity, that is, for Gaussian models.

In many practical applications, the assumption of independence along t conditional on the additive predictor for each functional response is not borne out and observed residuals are correlated (and frequently heteroscedastic) along t. This can easily be diagnosed by computing the empirical covariance and correlation of the residuals (see panel 4 of Figure 4, p. 24). If residual intra-curve correlations are non-negligible, confidence intervals (CIs) and tests will be overly optimistic. If computationally feasible, models should then include functional smooth residuals $E_i(t)$ to account for such autocorrelation and variance heterogeneity.

Generally speaking, all response distributions from scalar regression are also available for use in models with a functional response. Whether effects are constant or varying over the functional domain should be investigated for all variables (metric and categorical). How appropriate effect types can be determined as part of the modelling process is outlined in Section 4.4.

4.1 Uncertainty quantification

CIs for smooth effects can either be constructed *globally* (or simultaneously), *pointwise* or *intervalwise*, the interpretation being that the CI overlaps the true effect globally, at a specific point or in a specific interval with a given probability, respectively. This is an area of active research; see, for example, Krivobokova et al. (2010) or Marra and Wood (2012). As a generally applicable method, bootstrapping can be used to construct all different types of CIs. However, it can be computationally expensive—often prohibitively so for high-dimensional data or complex models.

Several bootstrap strategies exist that can be used in this context. The most established approach in the context of regression modelling is the conditional or parametric bootstrap (Efron and Tibshirani, 1994), which consists of the following steps for constructing a pointwise CI for the linear coefficient β_1 based on a sample of size n, but is also easily generalizable to compute intervalwise or global intervals:

1. Create B bootstrap samples from the data. In each of the B samples a new response value y_i^b is generated for each observation i by drawing a random value from the conditional response distribution specified by the regression model. In the Gaussian case, new response values y_i^b can be drawn from the distribution

   $$Y_i|\boldsymbol{X}_i \sim N(\hat{y}_i, \hat{\sigma}_\epsilon^2),$$

   where $\hat{y}_i$ is the model-based prediction for observation i and $\hat{\sigma}_\epsilon^2$ is the estimated error variance.
2. Calculate the model on each of the B samples and save β_1 as β_1^b, $b = 1, \ldots, B$.
3. Define the CI using empirical quantiles, for example, the 2.5% and 97.5% quantiles to obtain a 95% CI.

Note that parametric bootstrapping heavily relies on the underlying model being specified correctly. In case of violation of the model assumptions, this approach can lead to overly optimistic intervals and instead nonparametric bootstrapping should be used, where resampling is based on the raw data (Efron, 1979). Because of the exemplaric character of our running example we use nonparametric bootstrapping to estimate CIs.

As an alternative to bootstrapping, the *empirical Bayesian* CIs developed by Marra and Wood (2012), which are an extension of Nychka's (1988) CIs, are computationally efficient and implemented in `mgcv`. However, Marra and Wood show that these intervals do not perfectly fulfil the property of pointwise CIs. Figure 3 shows a comparison of the CIs of Marra and Wood and real pointwise, bootstrap-based CIs, with the latter being ever so slightly wider throughout in this case. Considering, however, that differences between these two are usually small as

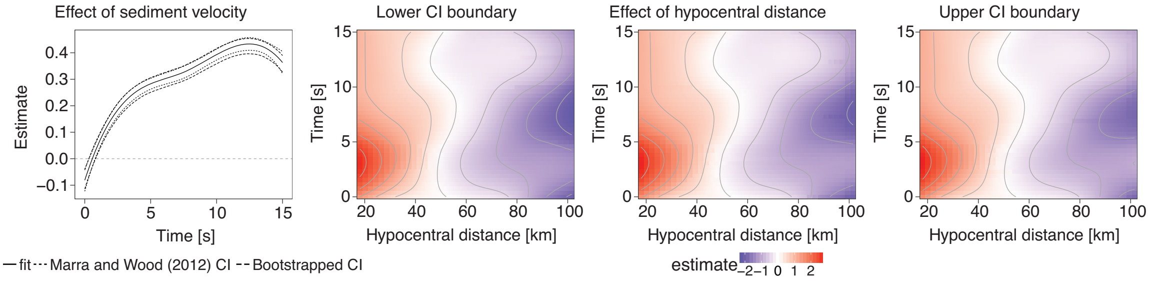

Figure 3 Left: Comparison of the Marra and Wood (2012) CIs and pointwise, nonparametric bootstrap-based CIs (95%) using 1000 Bootstrap samples for the smooth effect of sediment velocity. Right: Pointwise, nonparametric bootstrap-based CIs (95%) and point estimate for the time-varying smooth effect of hypocentral distance using 1000 Bootstrap samples

long as the model is not severely misspecified, the Marra and Wood CIs are a useful tool to compute uncertainty of smooth estimates efficiently.

In contrast to one-dimensional effects, visualization of uncertainty for multidimensional smooth effects is more complex as a 3D surface plot cannot be used to show both the point estimate and CIs. Instead, the best approach is to use separate heatmaps for the point estimate, the lower CI boundary and the upper CI boundary using identical colour legends, as shown in Figure 3. Looking at the estimates, it can be seen that the uncertainty about the effect of hypocentral distance is rather small.

Regarding predictions, both pointwise CIs for the predicted mean values and pointwise prediction intervals can be obtained based on Wood (2006,Ch. 1.3.6). For intervalwise or global versions of both interval types again bootstrap-based methods have to be used, but our practical experience suggests that the differences to pointwise CIs are usually negligible for practical purposes.

4.2 Hypothesis testing

Most of the relevant hypotheses in function-on-scalar regression can be tested using the five test approaches listed in Table 1. All tests apart from the bootstrap are Wald-like tests which are based on the approximate normal distribution of the estimated regression coefficients. For details, see Wood (2013) and Marra and Wood (2012). The appropriate test distribution mostly depends on the question whether the scale or dispersion parameter ϕ has to be estimated or not (Wood, 2006). For a normal response $\phi = \sigma^2$ is generally unknown, whereas the use of Poisson or Binomial responses implies a known value of $\phi = 1$.

Hypotheses for scalar coefficients of the form $\beta_j = 0$ can be tested using a t-test. For testing multiple β's being zero at the same time an F-test can be used (Wood, 2006, Ch. 4.8.5). A different F-test based on the test statistic introduced in Wood (2013) is available to test whether a nonlinear effect is significantly different from zero. Note that this is a test only for the *global* hypothesis. For a pointwise or intervalwise evaluation, a bootstrap-based approach has to be used. As in the previous section,

Table 1 Overview on relevant tests, based on whether the scale or dispersion parameter ϕ has to be estimated or not. $^{(1)}$ test based on Wood (2006, 4.8.5); $^{(2)}$ test based on Wood (2013)

Test		Testable alternative hypotheses
ϕ unknown	ϕ known	
t-test	z-test	Is a linear effect different from zero?
F-test$^{(1)}$	χ^2-test$^{(1)}$	Is at least one of multiple parameters different from zero?
F-test$^{(2)}$	χ^2-test$^{(2)}$	Is a smooth effect different from zero?
LR-test		Is model M_1 better than model M_2?
Bootstrap-based test		All hypotheses

a bootstrap is hereby used to create an appropriate CI and as a second step the null hypothesis is rejected if zero is not (or at no point for an intervalwise test) inside of the CI. Finally, specific hypotheses can also be tested by comparing models using likelihood ratio (LR)-tests (Wood, 2006, Ch. 4.10.1). Be aware that an LR-test can only be used for model comparison if the two models are nested. Some more information on model comparison is given in Section 4.4.

Note that all the tests given earlier are conditional on the estimated penalty parameters that control the effective degrees of freedom of each term. However, neglecting smoothing parameter uncertainty does not seem to have a large negative impact on the validity of p-values and the performance of CIs unless penalty parameters are poorly identified (Marra and Wood, 2012). An approach to account for smoothing parameter uncertainty in p-value calculation is outlined in Wood et al. (2016b) and implemented in `mgcv` as well (see `Vc` in `?gamObject`).

An overview on the most important hypotheses in function-on-scalar regression is given in Table 2, which lists possible research questions together with the appropriate tests. As a special note, testing whether two scalar effects (or two smooth effects) of x_k and x_j are different from one another only arises in situations where, for example, two treatments x_k and x_j (with time-varying effects) should be compared. Thus, this can be translated into another hypotheses by using one treatment as the reference category and then testing the hypothesis ‘Is the linear (or smooth) effect estimating the difference between treatments different from zero?’

Apart from the penalized likelihood-based (or *empirical* Bayesian) framework introduced here, fully Bayesian inference like, for example, the framework of Morris (2017), see Section 4.6, often allows for easier handling of complex or non-standard inferential problems. When relying on the software implementation of the Greven and Scheipl (2017a) framework in the `R` package `refund`, Bayesian estimation of all exponential family models is available using the automatic translation of the model specification and model data into `JAGS` (Plummer, 2016) code using `mgcv`’s `jagam` function (Wood, 2016) for automated, tuning-free, fully Bayesian inference based on Markov Chain Monte Carlo sampling.

Working with generally high-dimensional functional data, researchers should be aware that, all else being equal, large sample sizes lead to smaller p-values in the case of H_0 not being true. In such cases, importance should not be attached primarily to p-values of point hypotheses of ‘no effect’. Instead, best practice in interpreting

Table 2 Overview on possible hypotheses with corresponding tests. [(0)] tests are only reported for the case of unknown scale parameter ϕ. If ϕ is known we refer to Table 1; [(1)] F-test based on Wood (2006, 4.8.5); [(2)] F-test based on Wood (2013)

Research question (alternative hypothesis)	Test$^{(0)}$
Is the linear effect of x_j different from zero?	
↪ Case 1: x_j is metric or binary	t-test
↪ Case 2: x_j is categorical with $>$ 2 categories	LR-test
Is the smooth effect of x_j different from zero?	
↪ Globally	F-test$^{(2)}$
↪ At a specific point	Bootstrap
↪ In a specific interval	Bootstrap
Is at least one of multiple parameters different from zero?	LR-test
Are the linear effects of x_j and x_k different from one another?	see text
Are the smooth effects of x_j and x_k different from one another?	see text
Is the linear effect of x_j different depending on the value of x_k?	
↪ Case 1: both x_j and x_k are metric or binary	
↪ Case 1a: x_k is binary or the effect is varying linearly over the metric x_k	t-test
↪ Case 1b: the effect is varying nonlinearly over the metric x_k	LR-test
↪ Case 2: x_j and/or x_k are categorical with $>$ 2 categories	LR-test
Is the smooth effect of x_j different depending on the value of x_k?	
↪ Case 1: x_k is binary	F-test$^{(2)}$
↪ Case 2: x_k is metric or categorical with $>$ 2 categories	LR-test
Is model M_1 better than M_2?	LR-test

regression results is based on well-founded discussion of the *relevance* of the estimated effect strength and its associated uncertainty while considering whether the sample is appropriate for drawing general conclusions from it. Having quite high-dimensional data ourselves, we do not report specific test results for our running example.

4.3 Some specific challenges

We now list some further challenges that are specific to dealing with functional data. A first comparison of different modelling approaches regarding those problems is given and is complemented by the main discussion in Section 4.6.

If the functional responses have hierarchical, longitudinal or spatio-temporal structure, there may be non-negligible inter-curve correlation that the model has to account for. In the case of grouped data, that is, longitudinal or hierarchical data, functional random intercepts and slopes varying over the functional domain of the response can be incorporated into the model (Greven and Scheipl, 2017a). Spatio-temporal correlation with a pre-specified structure between functional responses can be included explicitly by including smooth effects over space or time. Scheipl et al. (2015, Online Appendix C) contains a worked example and code for spatially correlated curves.

Another common problem when dealing with *time-varying* functional data is misalignment or *phase variation* of functional observations. This means that certain salient features of the functional responses like peaks or plateaus do not occur at the exact same time points. Few functional data analysis frameworks are currently able to incorporate both phase and amplitude variation (cf. `fdasrvf`, Tucker, 2016) and we are not aware of any implementation of functional response regression able to do so. Ignoring misalignment typically results in blurred estimates. Therefore, an appropriate pre-processing of the data is necessary, for example, to align all peaks at the same time points. An overview on methods tackling phase variation in functional data analysis is given by Marron et al. (2015). In our application, ground velocity curves are heavily misaligned since the seismic shock waves take longer to reach seismometers further away from the hypocentre and the corresponding curves thus remain at (close to) zero for longer times. We pragmatically solve this problem by removing leading zeros before model estimation.

Functional data are frequently high-dimensional and estimation of complex models can be very expensive, both in terms of computation time and memory requirements. Pragmatically speaking, analysts facing such a problem should consider downsizing the data, for example, by reducing the resolution of functional measurements over the functional domain or by using only a subset of the data for estimation and the remainder for model validation. Highly efficient estimation algorithms are available for some approaches. For the class of spline-based models we focus on here, one can use the algorithm of Wood et al. (2016a), which is implemented in the function `bam` in `R` package `mgcv` (Wood, 2006), also accessible via `pffr`. The fully Bayesian wavelet-based approach of Morris (2017) and collaborators, implemented in the `WFMM` software (Herrick, 2015) has excellent scaling behaviour for time and memory both in terms of data set size and model complexity.

Finally, users should be aware that some methods for functional data are only applicable if the functional observations contain no missing measurements and were observed on a regular grid, that is, all functional observations are evaluated at the same points of the functional domain. A comparison of the applicability of various function-on-scalar regression frameworks is given in Section 4.6.

4.4 Model selection

Generally speaking, model selection in functional regression models underlies the same principles as in scalar regression (see, e.g., Marra and Wood, 2011; Fahrmeir et al., 2013, Ch. 3.4.3). Using model selection in function-on-scalar regression can be useful for various issues, for example, for deciding which response distribution and link function is optimally suited to the data or whether an effect should be incorporated linearly or as a smooth curve. Additionally, very high-dimensional data often reduces the effectiveness of penalization methods as the information in the observed data overwhelms the penalization prior (Gelman et al., 2014). In such situations it can be necessary to use model selection to optimize the number of basis functions for each smooth effect.

Leeb and Pötscher (2005) propose a test set based approach to prevent overfitting and preserve valid p-values when performing model selection. For smaller datasets, *k*-fold cross validation is a valid alternative (Hastie et al., 2009). Using one of those two approaches, the best model can, for example, be found by using the prediction error as the optimization measure. When model selection is based on training set performance, other criteria like AIC or LR tests should be used (Fahrmeir et al., 2013). Note that if using the semiparametric approach smoothing parameter uncertainty should be accounted for in AIC computation (see Wood et al., 2016b).

For our data, we use a test set based model selection approach with mean square prediction error (MSE) as the criterion for two purposes. First, penalization did not work very well in this setting, probably due to the massive amount of data available. Therefore, we use a pragmatic model selection procedure to select the number of basis functions for each smooth effect and to decide whether individual effects should be incorporated as a smooth effect or linearly. Second, we use model selection to choose between different response distributions and link functions.

4.5 Model evaluation

Model assumptions for functional response regression are mostly the same as in respective scalar models, that is, observations are independent conditional on the additive predictor. Model evaluation is mainly done by visualizing the residual structure.

A selection of useful residual plots is shown in Figure 4. The structure of the residuals plotted against the fitted values (panel 1) is acceptable. Most measurements are predicted approximately correct. The odd structure of negative residuals is based on the ground velocities being non-negative, which results in highly negative residuals not being possible. Plotting the mean residuals over space (panel 2) shows that a substantial spatial struture is remaining in the residuals. Nearly all regions where ground velocities were substantially underestimated are west of the earthquake centre, while seismometer readings in regions to the east and to the south of the epicentre were overestimated. Across time, however, we again observe an acceptable amount of

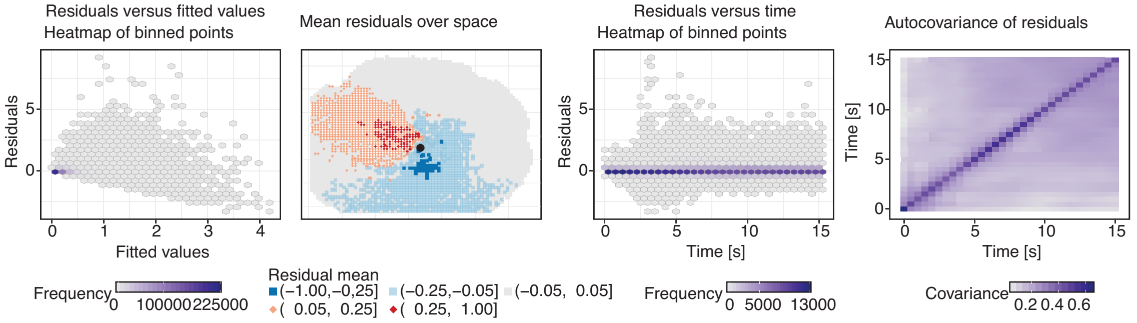

Figure 4 From left to right: Residuals versus fitted values, residuals versus space, residuals versus the functional domain, autocovariance of residuals over the functional domain. The black dot in the second plot marks the epicentre

residual structure—the hexbin plot (panel 3) does not show any systematic deviations from a constant trend at zero, with some extreme peak ground velocities observed at around five seconds. As functional data often are very high-dimensional standard scatterplots of the residuals, having the problem of overplotting should generally be avoided in favour of alternative plots like density plots or hexagonal binning (Carr et al., 2016), as was done in the left and middle plots of Figure 4. The empirical autocovariance of the residuals (panel 4) corresponds well enough to the model assumptions: it is fairly constant along the diagonal (i.e., the variance of the residuals is fairly homogeneous over functional domain t) and drops off quickly towards zero away from it (i.e., the autocorrelation of the residuals along t is rather small and very short-range), slightly less so for $t > 10$.

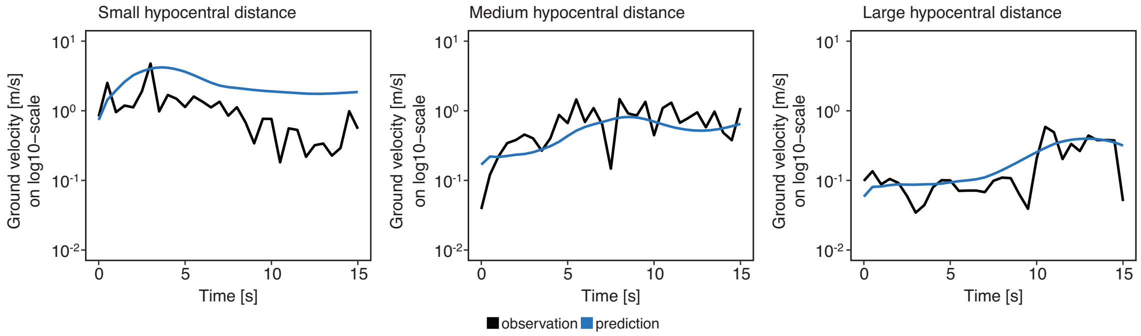

Figure 5 Comparison of model predictions and raw observations for typical observations with different hypocentral distances

For an evaluation of the prediction power of the model, measures like MSE of the predictions can be calculated. We also recommend graphical evaluation of predictions for single functional observations as was done in Figure 5 to get an overview on model performance.

4.6 Alternative approaches and software implementations

As alternative approaches to semiparametric regression we only cover the most versatile frameworks for performing function-on-scalar regression. The capabilities of the respective software implementations are also outlined, a comprehensive comparison of available software implementations is given in Table 3 of Greven and Scheipl (2017b). Many specialized function-on-scalar regression methods have been proposed in the literature, oftentimes with corresponding small software implementations, which we do not cover here. See Morris (2015) for an in-depth review of this field.

The semiparametric approach of Greven and Scheipl (2017a) was already outlined extensively. The methodology is ready-to-use in the `refund` package in `R` (Goldsmith et al., 2016), the most versatile function therein being `pffr`.

One class of alternative approaches includes a pre-smoothing step prior to model estimation, meaning that each functional observation is smoothed and the resulting smooth curve is then treated as the functional observation (see, e.g., Ramsay and Silverman, 2005). The disadvantage is that the measurement error removed by the smoothing step is not taken into account in subsequent inference. On the plus side, this can allow for more efficient estimation as the smooth curve can then be represented compactly by the vector of spline coefficients yielding the smoothed curve. The `R` package `fda` is publicly available (Ramsay et al., 2014) and implements simple linear models for functional responses.

An overview of nonparametric methods and their applications is provided in Ferraty and Vieu (2006). Their regression approaches are usually based on kernel methods and are able to model highly nonlinear associations. However, the methods mostly cover only univariate models with a single covariate. Febrero-Bande et al. (2012) introduce the `R` package `fda.usc` which implements a subset of these methods and related extensions.

The componentwise gradient boosting framework of Brockhaus et al. (2016b) is spline based and extremely versatile. With boosting being a popular, very efficient yet very powerful estimation technique, it represents a neat alternative to the standard regression approach. The advantages are most noticeable when working with very high-dimensional data requiring an efficient estimation technique or when dealing with data situations with more parameters than observations, as such settings remain computationally feasible using a boosting approach. Also, the boosting approach automatically performs variable selection. However, uncertainty quantification for boosting is currently only possible using computationally expensive resampling techniques like bootstrapping (Hastie et al., 2009). The method is implemented in the `R` package `FDboost` (Brockhaus, 2016). Recently, this approach has also been extended to model the variance of functional responses conditional on covariates (Brockhaus et al., 2016a), using techniques developed in the literature on generalized additive models for location, scale and shape (GAMLSS; Mayr et al., 2012). More general details on boosting and GAMLSS can be found in the tutorials by Mayr and Hofner (2018) and Stasinopoulos et al. (2018), respectively, which are also part of this special issue.

As another alternative, fully Bayesian functional regression can be used. The most comprehensive framework we are aware of is the one of Morris (2017) and collaborators, who also provide a comprehensive comparison to the approach of Greven and Scheipl (2017a). Generally speaking, fully Bayesian approaches have the advantage that diverse between- and within-function correlation structures can be incorporated into the model in a very flexible way. Also, handling inference is much easier as approximate posterior distributions of all parameters are available in the form of MCMC samples. Readers interested in a general introduction to Bayesian distributional regression are pointed to the tutorial paper by Umlauf and Kneib (2018). Unfortunately, the Morris (2017) framework lacks a comprehensive and well-documented publicly available software implementation at the time of writing. A C++ and Matlab implementation called `WFMM` (Herrick, 2015) for conditionally Gaussian functional responses with a limited feature set is publicly available.

5 Discussion and outlook

This work provides an introduction into the general concepts of function-on-scalar regression. Important practical considerations and best practices are outlined for the most important modelling tasks. We hope that researchers can use this work as a starting point for applying functional regression models to their own data. Comprehensive `R` code for our running example is available in the online supplement.

We concentrated on the semiparametric approach of Greven and Scheipl (2017a) as this framework is rather flexible in terms of incorporating different types of covariate effects, is applicable for both regular and irregular data with possible missing values, and is accompanied by a flexible implementation of function-on-scalar regression in the `refund` package. However, important differences regarding practical aspects of the application of the existing function-on-scalar regression frameworks are also outlined. Furthermore, current limitations like the problem of accounting for phase variation and intra-functional correlation are made clear.

As this work is mainly aimed at introducing the approach to those not familiar with functional response regression and to offer advice on the correct application of such methods, it should be clear that not all methodological aspects of functional regression are covered. One crucial point we have not discussed is the use of functional principal components (fPCs) as a popular alternative to using spline basis functions. fPCs often lead to a very compact basis and nicely interpretable results. An overview on fPC-based approaches is given in Wang et al. (2016). Note that functional residuals and other functional random effects can be represented using fPCs as well in the approach described here (cf. Greven and Scheipl, 2017a).

Finally, we look forward to the ongoing development of ready-to-use and robust methodology for functional regression. Being both an important method for working with complex data structures and a field where research is still needed for some important aspects, functional regression stays one of the currently most exciting fields of modern statistics.

Acknowledgements

Financial support by the German Research Foundation for the development of `refund::pffr` through the Emmy Noether Programme, grant GR 3793/1-1 is gratefully acknowledged. We also thank the anonymous reviewers, whose insightful and constructive comments were highly appreciated and helped making the article more accessible.

References

Bauer A (2016) *Auswirkungen der Erdbebenquelldynamik auf den zeitlichen Verlauf der Bodenbewegung* [Impact of earthquake fault dynamics on the temporal development of ground motion]. Master's thesis, Ludwig-Maximilians-Universität, Munich, Germany. URL https://epub.ub.uni-muenchen.de/31976/

——— (2017) bauer-alex/FoSIntro: v1.0 of the FoSIntro package. doi:10.5281/zenodo. 1012730

Breuer A, Heinecke A, Rettenberger S, Bader M, Gabriel A-A and Pelties C (2014) Sustained petascale performance of seismic simulations with seissol on supermuc. In *International Supercomputing Conference*, pages 1–18. Cham: Springer.

Brockhaus S (2016) *FDboost: Boosting functional regression models*. R package version 0.2-0. URL https://CRAN.R-project.org/package=FDboost

Brockhaus S, Fuest A, Mayr A and Greven S (2016a) *Signal regression models for location, scale and shape with an application to stock returns*. arXiv preprint arXiv:1605. 04281.

Brockhaus S, Melcher M, Leisch F and Greven S (2016b) Boosting flexible functional regression models with a high number of functional historical effects. *Statistics and Computing*, **27**, 913–26.

Carr D, Lewin-Koh N, Maechler M and Sarkar D (2016) *hexbin: Hexagonal Binning Routines*. R package version 1.27.1. URL https://CRAN.R-project.org/package=hexbin

Diggle P, Heagerty P, Liang K-Y and Zeger S (2002) *Analysis of longitudinal data*. Oxford: Oxford University Press.

Duchon J (1977) Splines minimizing rotation-invariant semi-norms in sobolev spaces. In Schempp W and Zeller K, eds. *Constructive theory of functions of several variables*, Lecture Notes in Math. Vol. 571, pages 85–100. Berlin and New York: Springer-Verlag.

Efron B (1979) Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, **7**, 1–26.

Efron B and Tibshirani RJ (1994) *An introduction to the bootstrap*. CRC Press.

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

Fahrmeir L, Kneib T, Lang S and Marx B (2013) *Regression: Models, methods and applications*. Berlin Heidelberg: Springer.

Febrero-Bande M and de la Fuente MO (2012) Statistical computing in functional data analysis: The r package fda. usc. *Journal of Statistical Software*, **51**, 1–28.

Ferraty F and Vieu P (2006) Nonparametric functional data analysis: Methods, theory, applications and implementations.

Gelman A, Carlin JB, Stern HS and Rubin DB (2014) *Bayesian data analysis*, (Vol. 2). Boca Raton, FL: Chapman & Hall/CRC.

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*. R package version 0.1-15. URL https://CRAN.R-project.org/package=refund

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

———(2017b) Rejoinder. *Statistical Modelling*, **17**, 100–15.

Hastie T, Tibshirani R and Friedman J (2009) *The elements of statistical learning: Data mining, inference and prediction*, 2nd edition. New York: Springer.

Herrick R (2015) *WFMM*, version 3.0 edition. The University of Texas MD Anderson Cancer Center. URL https://biostatistics.mdanderson.org/SoftwareDownload/SingleSoftware.aspx?Software_Id=70

Krivobokova T, Kneib T and Claeskens G (2010) Simultaneous confidence bands for penalized spline estimators. *Journal of the American Statistical Association*, **105**, 852–63.

Leeb H and Pötscher BM (2005) Model selection and inference: Facts and fiction. *Econometric Theory*, **21**, 21–59.

Marra G and Wood SN (2011) Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis*, **55**, 2372–87.

———(2012) Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics*, **39**, 53–74.

Marron JS, Ramsay JO, Sangalli LM, and Srivastava A (2015) Functional data analysis of amplitude and phase variation. *Statistical Science*, **30**, 468–84.

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.

Mayr A and Hofner B (2018) Boosting for statistical modelling. A non-technical introduction. *Statistical Modelling*, **18**, 365–84.

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

———(2017) Comparison and contrast of two general functional regression modelling frameworks. *Statistical Modelling*, **17**, 59–85.

Nychka D (1988) Bayesian confidence intervals for smoothing splines. *Journal of the American Statistical Association*, **83**, 1134–43.

Pelties C, Gabriel AA and Ampuero JP (2014) Verification of an ader-dg method for complex dynamic rupture problems. *Geoscientific Model Development*, **7**, 847–66.

Plummer M (2016) *rjags: Bayesian graphical models using MCMC*. R package version 4-6. URL https://CRAN.R-project.org/package=rjags.

Pya N and Wood SN (2016) A note on basis dimension selection in generalized additive modelling. arXiv preprint arXiv:1602.06696.

R Core Team (2016) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Ramsay JO and Silverman BW (2005) *Functional data analysis*. New York: Springer.

Ramsay JO, Wickham H, Graves S and Hooker G (2014) *fda: Functional Data Analysis*. R package version 2.4.4. URL https://CRAN.R-project.org/package=fda

Ramsay JO, Hooker G and Graves S (2009) *Functional data analysis with R and MATLAB*. New York: Springer.

Reiss PT, Huang L and Mennes M (2010) Fast function-on-scalar regression with penalized basis expansions. *International Journal of Biostatistics*, **6**.

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

Stasinopoulos M, Rigby RA and de Bastiani F (2018) A distributional regression approach using GAMLSS. *Statistical Modelling*, **18**, 248–73.

Tucker JD (2016) *fdasrvf: Elastic functional data analysis*. R package version 1.7.1.

Umlauf N and Kneib T (2018) A primer on bayesian distributional regression. *Statistical Modelling*, **18**, 219–47.

Wang J-L, Chiou J-M and Mueller H-G (2015) *Review of functional data analysis*. arXiv preprint arXiv:1507.05135.

——— (2016) Functional data analysis. *Annual Review of Statistics and Its Application*, **3**, 257–95.

Wood SN (2003) Thin plate regression splines. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **65**, 95–114.

———(2006) *Generalized additive models: An introduction with R*. CRC press.

———(2013) On p-values for smooth components of an extended generalized additive model. *Biometrika*, **100**, 221–28.

———(2016) *Just another gibbs additive modeller: Interfacing jags and mgcv*. arXiv preprint arXiv:1602.02539.

Wood SN, Li Z, Shaddick G and Augustin NH (2016a) Generalized additive models for gigadata: Modelling the uk black smoke network daily data. *Journal of the American Statistical Association*, **112**, 1199–1210.

Wood SN, Pya N and Säfken B (2016b) Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, **111**, 1548–63.

Part IV.

APC Analysis

8. Semiparametric APC Analysis of Destination Choice Patterns: Using Generalized Additive Models to Quantify the Impact of Age, Period, and Cohort on Travel Distances

Contributing article

Weigert, M., Bauer, A., Gernert, J., Karl, M., Nalmpatian, A., Küchenhoff, H., and Schmude, J. (2021). Semiparametric APC analysis of destination choice patterns: Using generalized additive models to quantify the impact of age, period, and cohort on travel distances. *Tourism Economics.*

Code repository

https://github.com/MaxWeigert/TravelDistAPC

Copyright information

Author contributions

The concept for the publication was jointly created by all authors excluding Asmik Nalmpatian. Maximilian Weigert performed the statistical literature review and proposed the use of generalized additive models with a two-dimensional tensor product spline. Alexander Bauer proposed ridgeline matrices as novel visualization concept. All other parts of the statistical methodology were jointly developed by Maximilian Weigert and Alexander Bauer. Helmut Küchenhoff supervised the statistical analysis. Maximilian Weigert and Alexander Bauer jointly wrote the R codebase for all analyses, with some contribution of Asmik Nalmpatian. Maximilian Weigert wrote the initial draft of the literature review and the statistical methodology section. Johanna Gernert and Marion Karl jointly developed the tourism science parts under guidance of Jürgen Schmude. Johanna Gernert wrote initial drafts for the literature review and the respective parts of effect interpretation and the conclusion. All other parts were mainly written by Maximilian Weigert and Alexander Bauer. Alexander Bauer performed the main part in revising the scientific and linguistic style of all parts of the publication and in intertwining them to ensure a consistent style.

All authors took part in extensive idea discussions and proofreading. This publication was produced as part of the research project TourIST (Tourism in Space and Time) conceived by Marion Karl, Jürgen Schmude, Alexander Bauer and Helmut Küchenhoff and which was funded by the German Research Foundation (DFG) under Grant Nos KU 1359/4-1, SCHM 850/22-1, and KA 4976/2-1.

Supplementary material

- Online supplementary material: https://journals.sagepub.com/doi/suppl/10.1177/1354816620987198
- Accompanying conference poster: https://www.researchgate.net/publication/353852226_Visualization_techniques_for_semiparametric_APC_analysis_Using_Generalized_Additive_Models_to_examine_touristic_travel_distances

Special Issue: Panel Data Models in Tourism Research

TOURISM ECONOMICS
The Business and Finance of Tourism and Recreation

Tourism Economics
1–24
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1354816620987198
journals.sagepub.com/home/teu
SAGE

Semiparametric APC analysis of destination choice patterns: Using generalized additive models to quantify the impact of age, period, and cohort on travel distances

Maximilian Weigert
LMU Munich, Germany

Alexander Bauer
LMU Munich, Germany

Johanna Gernert
LMU Munich, Germany

Marion Karl
University of Queensland, Australia

Asmik Nalmpatian
LMU Munich, Germany

Helmut Küchenhoff
LMU Munich, Germany

Jürgen Schmude
LMU Munich, Germany

Abstract

This study investigates how age, period, and birth cohorts are related to altering travel distances. We analyze a repeated cross-sectional survey of German pleasure travels for the period 1971–2018 using a holistic age–period–cohort (APC) analysis framework. Changes in travel distances are attributed to the life cycle (age effect), macro-level developments (period effect), and generational membership

Corresponding author:
Maximilian Weigert, Department of Statistics, Statistical Consulting Unit StaBLab, LMU Munich, Ludwigstraße 33, 80539 Munich, Germany.
Email: maximilian.weigert@stat.uni-muenchen.de

(cohort effect). We introduce ridgeline matrices and partial APC plots as innovative visualization techniques facilitating the intuitive interpretation of complex temporal structures. Generalized additive models are used to circumvent the identification problem by fitting a bivariate tensor product spline between age and period. The results indicate that participation in short-haul trips is mainly associated with age, while participation in long-distance travel predominantly changed over the period. Generational membership shows less association with destination choice concerning travel distance. The presented APC approach is promising to address further questions of interest in tourism research.

Keywords
APC analysis, cohort analysis, destination choice, generalized additive models, ridgeline matrices, travel distance

Introduction

Overcoming geographic distances is, by definition, one of the constitutive elements of tourism as people need to temporarily travel to places outside their everyday environment to be defined as tourists (Cooper and Hall, 2016). Alterations of tourist flow can be attributed to time-related factors, including developments across the life cycle, over time periods or between successive generations (Oppermann, 1995). Technological developments in transportation (Castro et al., 2020), economic conditions of the source market (Sun and Lin, 2019) and the greater availability of information through modern communication technology (Yang et al., 2018) have facilitated long-distance travel over time. Apart from these external influences, destination choice, and hence travel distance, depends on sociodemographic and psychological characteristics of the traveler (Wong et al., 2016). Travel behavior changes over the course of someone's life cycle due to changing personal circumstances and increasing age (Bernini and Cracolici, 2015) and between generations (Lohmann and Danielsson, 2001). Furthermore, travel-related factors such as time availability also affect destination choice (McKercher and Mak, 2019).

Thorough analyses of such temporal patterns particularly rely on quantifying and comprehensively communicating the developments over age, period, and cohort. The separation of these factors is performed with statistical age–period–cohort (APC) analysis methods. Therein, each temporal dimension describes characteristic developments regarding the individual traveler or external circumstances of holiday trips. Following Yang and Land (2008), age effects represent the ageing process of an individual, period effects refer to external events and environmental changes, and cohort effects relate to specific groups of individuals who experience the same events in a specific span of time. APC analyses require long-term panel or repeated cross-sectional data (Yang and Land, 2013). Due to the sparse availability of adequate data and the complexity of existing methods, only few studies in tourism research have examined alterations in travel behavior based on all three temporal dimensions so far (e.g. Oppermann, 1995). In this work, we introduce an established APC approach into tourism research by analyzing the temporal development of travel distances of German tourists, investigating the research question:

> How are age, period, and cohort related to altering travel distances?

The key challenge in APC analyses is to separate these temporal effects by overcoming the identification problem that each component is a linear combination of the others (Clayton and

Schifflers, 1987), for example, age = period − cohort. In other words, the three components can never be fully separated and interpretation requires a thorough understanding of their interrelations. Statistical models based on linear effect structures only yield a unique solution if further assumptions about these interrelations are made. The quality of the solution highly depends on the adequacy of the substantial assumptions and the underlying data.

Novel to tourism research and based on a repeated cross-sectional German survey which covers travel information from almost 50 years, our study highlights the potential of holistic APC modeling to generate a more comprehensive understanding of the factors that drive changes in travel behavior. We provide three major contributions. First, we introduce a state-of-the-art approach for APC analysis into the field of tourism research. It is based on a generalized additive modeling framework, where cohorts are represented as an interaction between age and period. The approach is applicable for panel and repeated cross-sectional data as well as individual and aggregated data. Secondly, we introduce ridgeline matrices and partial APC plots as novel graphical tools for analyzing APC structures to facilitate the communication of complex temporal patterns. The former build on the visualization concept of Lexis diagrams (Carstensen, 2007). Finally, we contribute new insights about the key factors that trigger destination choice by analyzing the associations of age, period, and cohort with altering travel distances using a comprehensive statistical approach which has not yet been applied in tourism science. The latter comprises both the pure analysis of APC structures and the inclusion of further variables potentially associated with altering travel distances. In terms of practical implications, understanding the spatiotemporal movements of tourists and their influencing factors can support practitioners and policy-makers in the planning and management of destinations, including future travel behavior predictions.

Literature review

Geographic distance and destination choice

Tourism literature emphasizes the importance of geographic distance in destination choices (Lee et al., 2012; Yang et al., 2018). Adopted from Tobler's (1970) first law of geography, the negative impact of distance between origin and destination on destination choice can be explained by distance decay theory: tourism demand declines with increasing geographic distance (McKercher et al., 2008). This spatial decline of outbound tourism demand is associated with rising physical, temporal, and monetary costs (Taylor and Knudson, 1973).

Advances in transportation and communication technologies coupled with reduced travel time and costs have facilitated long-haul travel in the last decades (McKercher and Mak, 2019). These developments lead to the assumption that the negative impact of distance (i.e. the friction effect of distance) on tourism demand diminished over time (Yang et al., 2018), indicating a strong period effect. Even so, studies investigating temporal changes of international tourism flows show that geographic distance still has a substantial impact on demand patterns, supporting the robustness of distance decay theory (Lee et al., 2012; McKercher and Mak, 2019).

How far one is willing to travel depends on sociodemographic and psychographic traits of the traveler such as age, household size, or income (Eugenio-Martin and Campos-Soria, 2011), tripographic characteristics such as trip duration (Guillet et al., 2011), and socioeconomic factors of the place of residence (Sun and Lin, 2019; Wong et al., 2016). The distribution of traveled distances can be visualized by demand curves. Their shape varies depending on the respective source market, type of tourist, and type of travel (McKercher and Mak, 2019; Wong et al., 2017), indicating the influence of such factors on destination choices.

Distance decay studies usually take advantage of available macro data on tourist flows. They use aggregated international tourist arrival or departure data to examine the association between travel distance and international travel patterns (McKercher and Mak, 2019; McKercher et al., 2008; Sun and Lin, 2019). The data sources limit these studies to the analysis of period effects (i.e. the influence of societal and economic factors on temporal changes). For example, Sun and Lin (2019) found that economic welfare and transport capacity are key factors for longer travel distances. The influence of sociodemographic or travel-related characteristics on tourist flows is often neglected in such studies (Yang et al., 2018). Identifying the key factors for changing destination choices considering both external and internal factors remains a challenge in tourism research. It requires both long-term data on individual level and complex statistical approaches such as APC analysis.

Age–period–cohort analysis

Association of age, period, and cohort with travel behavior.

Travel behavior is changing over time due to various reasons. To explain temporal developments, research suggests considering three dimensions: age, period, and cohort (e.g. Oppermann, 1995). The effect of an individual's age on the propensity to travel or the type of holiday experience is generally explained by life cycle theory (Bernini and Cracolici, 2015). According to this concept, age-related shifts in travel behavior are mainly associated with the varying life stages an individual or family passes (Chen and Shoemaker, 2014), ranging from childhood, young adulthood, newly married couple, parenting, empty nest to retirement. Most notably, the different stages are characterized by the change of marital status as well as altering income levels over the life cycle (Bowen and Clarke, 2009). This indicates that age serves as a proxy for these and other personal changes such as physical health (Scheiner and Holz-Rau, 2013). Studies indicate a nonlinear age effect on tourism demand including a small mid-30s to 40s dip in the overall decreasing curve (Collins and Tisdell, 2002). In terms of distance traveled, Oppermann (1995) found a bimodal pattern showing that the propensity for long-distance travel has its maximum among young people in their 20s and a second peak around age 50 (i.e. when children have moved out). The overall negative correlation of age with long-haul travel is associated with a gradual decline in health and mobility (You and O'leary, 2000). However, due to social change and modern life cycles, modifications and extensions of the traditional family life cycle need to be considered. For example, while single-parent families are less likely to choose long-distance destinations, this is not the case for couples of the same age without children (Collins and Tisdell, 2002).

Alterations in travel behavior are also caused by external factors of the macro environment simultaneously affecting people of all ages (Pennington-Gray and Spreng, 2002). These period effects comprise various factors including single events such as terror attacks, which have short- and long-term impacts on travel behavior (Karl et al., 2017), pandemic crises (Romagosa, 2020), as well as long-term trends such as economic developments in the source market (Wong et al., 2016), technological advances in transportation (Castro et al., 2020), and mobile technology (Cohen et al., 2014) or climate change (Gössling et al., 2012). While modern developments in transportation permanently encourage long-distance travel, events such as economic downturns can deter people from traveling overseas (Sun and Lin, 2019). New climate change protection policies and changes in consumers' perception of air travel (e.g. flight shame) may reduce travel distance in the future

(Becken and Carmignani, 2020). Potential long-term changes are also expected due to the COVID-19 pandemic which leads to a revival of short-distance travel (Romagosa, 2020). The impact of such external factors on (future) travel behavior is assessed in tourism demand modeling. Methodological approaches in this field comprise predictive methods purely based on historic tourist data (i.e. time series studies) or based on known causal relationships between demand and explanatory variables (i.e. econometric studies) (Song and Li, 2008). More recent advances in tourism demand modeling include artificial intelligence-based models, where different data sources can be combined to estimate models that are mainly focused on deriving sound predictions (Song et al., 2019).

Besides age and period effects, researchers indicate that travel behavior is also shaped by generational membership (McKercher et al., 2020). According to generational theory, members of a birth cohort share collective values and experiences through epochal events (Pendergast, 2010), reflected in similar consumer behavior patterns (Glover and Prideaux, 2009). Reviewing consumer behavior research in tourism, Cohen et al. (2014) identified generational membership as one of the core influential factors of tourism behavior. The specific beliefs and attitudes of each generation remain consistent over the ageing and life cycle process (Schewe and Noble, 2000). Targeted cohort analyses were applied in various tourism research settings, investigating generational differences regarding travel motivation and preferences (Chen and Shoemaker, 2014; Huang and Lu, 2017; Pennington-Gray et al., 2003), tourism expenditure (Bernini and Cracolici, 2015), online travel information search (Beldona, 2005), tourism experiences sought (Lehto et al., 2008), activity participation (You and O'leary, 2000), and destination choice (Huang and Lu, 2017; Li et al., 2013). Regarding destination selection, studies found that younger generations are more inclined to travel abroad to visit off-the-beaten path destinations (Li et al., 2013). This can be related to the effect of generation-specific socialization experiences on travel behavior (Oppermann, 1995).

Evidently, changes in travel behavior are triggered simultaneously and interactively by the effects of age, period, and cohort. In tourism research, the above stated cohort analyses aim to separate these effects with the main goal of identifying generation-specific consumption patterns. These insights are used for market segmentation (Schewe and Noble, 2000) and as a tool for tourism forecasting (Pennington-Gray et al., 2002). However, due to the sparse availability of long-term data on tourist behavior, studies investigating the temporal variations of travel behavior with a comprehensive (APC) approach are rare (Bernini and Cracolici, 2015). Instead, cohort analyses in tourism are often based on single cross-sectional surveys (e.g. Huang and Lu, 2017) or include only a few time points (e.g. Beldona, 2005). Such data settings exacerbate a reliable separation of age and cohort effects in these studies. Although research confirmed an association between these factors and destination choice (e.g. Bernini and Cracolici, 2015; Oppermann, 1995), to our knowledge no empirical studies have yet analyzed travel distance alterations based on all three temporal dimensions.

Statistical APC approaches. Examining to which extent observed developments can be attributed to each temporal component requires a joint analysis framework. APC analyses have primarily originated in epidemiological science to analyze mortality rates for specific population groups defined by age, period, and cohort (Kupper et al., 1985). The last decades showed an increasing use of APC methods across various research fields (Yang and Land, 2013), also due to the availability of more sophisticated statistical approaches. Descriptive analysis is usually performed with Lexis diagrams (Figure 1), that is, a two-dimensional table or graph depicting age groups and periods in

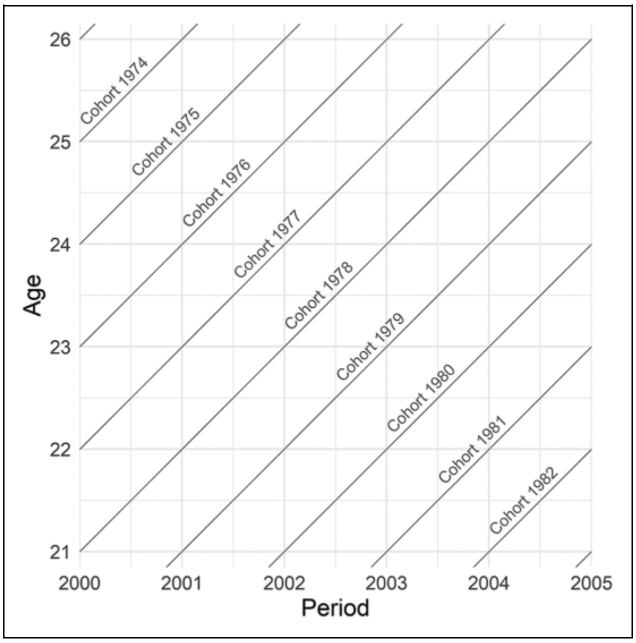

Figure 1. Sketch of a Lexis diagram. Period and age are displayed on the *x*-axis and *y*-axis, respectively. Cohorts are represented as diagonals.

horizontal and vertical direction, respectively (Carstensen, 2007). Accordingly, unique cohorts are displayed along the diagonals.

The most popular version of an APC model for repeated cross-sectional data is defined as a multiplicative three-factor regression model (see, e.g. Holford, 1983). More generally, it is a generalized linear model (GLM, Nelder and Wedderburn, 1972) of the form

$$g\left(\mu_{\text{apc}}\right) = \beta_0 + \beta_a \times \text{age}_a + \beta_p \times \text{period}_p + \beta_c \times \text{cohort}_c \tag{1}$$

where μ_{apc} denotes the expected value of an exponential family response for age group $a = 1, \ldots, A$, period $p = 1, \ldots, P$, and cohort $c = 1, \ldots, C$, $g(\cdot)$ denotes the link function, β_0 denotes the intercept, and β_j $(j \in \{a, p, c\})$ denotes the linear coefficients. The parameterization of this classical APC model is based on cross-sectional data aggregated over age groups, periods, and cohorts. However, the structure is easily adaptable to individual data (e.g. Fannon et al., 2018). Panel data can be analyzed by introducing random effects into the model (Diggle et al., 2002).

Given its linear predictor and the linear dependency of age, period, and cohort, model (1) cannot be estimated without setting constraints on the effect structures. Numerous strategies have been developed to overcome the identification problem. While early methods often used strict linear constraints such as the equality of two of the three effects (e.g. Fienberg and Mason, 1979), modern approaches rely on less restrictive assumptions. For instance, Bayesian hierarchical models restrict first- and second-order differences of the effects (e.g. Schmid and Held, 2007); the intrinsic estimator applies a form of principal components regression (Fu, 2000). Clayton and Schifflers (1987) give a detailed consideration to the identification problem and common issues in estimation

Table 1. Overview of the generations in the data and the respective periods and age groups in which they were observed.

Generation	Birth years	Relative frequency (%)	Observed periods	Observed ages
Generations born before 1939	1874–1938	25.0	1971–2018	33–99
Silent Generation	1939–1946	13.8	1971–2018	25–79
Baby Boomer	1947–1966	37.5	1971–2018	14–71
Generation X	1967–1982	17.7	1981–2018	14–51
Generation Y	1983–1994	5.2	1997–2018	14–35
Generation Z	1995–2010	0.8	2009–2018	14–23

Note: Generations before 1939 are not of special interest and are summarized.

and interpretation. A thorough overview of existing methodology is given by Yang and Land (2013).

A model class which has gained popularity in APC analysis since the late 1990s (e.g. Carstensen, 2007; Heuer, 1997) is spline-based regression. This approach overcomes the identification problem by estimating potentially nonlinear age, period, and cohort effects. Building on this approach, Clements et al. (2005) propose an APC model using a bivariate spline function depending on age and period in a generalized additive regression model (GAM). The resulting two-dimensional interaction surface implicitly contains information about cohorts on the diagonals. Given data on aggregated level, the model structure is given by

$$g\left(\mu_{ajc}\right) = \beta_0 + f_{ap}\left(\text{age}_a, \text{period}_p\right) \tag{2}$$

where $f(\cdot,\cdot)$ is a nonlinear interaction surface represented by a two-dimensional spline basis. As a direct extension of GLMs, generalized additive regression is a robust and flexible approach for modeling nonlinear effect structures (Wood, 2017). GAMs are applicable to all exponential family responses and to research settings with additional explanatory variables. Penalized splines (e.g. Eilers and Marx, 1996) enable the estimation of nonlinear relationships in the data but avoid overfitting by imposing a penalty on the effect's roughness. The application of GAMs is facilitated by sophisticated and freely available software (e.g. Wood, 2017).

Data and methods

Database

The data used in this study were collected in the *Reiseanalyse*, an annual cross-sectional survey on pleasure travel among approximately 7500 German residents (FUR, 2020b). Survey data are available from 1971 to 2018 and comprise around 227,000 holiday trips. Travelers from former East Germany have been included since 1990. The target population comprises (West) German citizens until 2009, and all German-speaking residents thereafter. Data are representative respective to federal state, city size, age, sex, household size and income, education level and citizenship (FUR, 2020a). We focus on five generations following the classification of Herhoffer and Meurer (2018). A description of these generations and their observed data is given in Table 1.

We analyze travel distances of the main holiday trips, each one defined as the personally most important trip within a year, lasting at least 5 days. This comprises both domestic and outbound

Table 2. Overview of the travel distance categorization used in the analyses.

Travel distance	Exemplary destinations	Relative frequency
<500 km	Germany and neighboring countries	1971: 57.9% 2018: 31.9%
500–1000 km	Neighboring and close European countries	1971: 24.5% 2018: 18.7%
1000–2000 km	European countries (e.g. Spain, Portugal, Malta, and Finland)	1971: 15.3% 2018: 28.4%
2000–6000 km	North Africa, Middle East, Russia, and Mongolia	1971: 1.5% 2018: 11.5%
>6000 km	America, Africa (excluding North Africa), Asia, and Australia	1971: 0.8% 2018: 9.5%

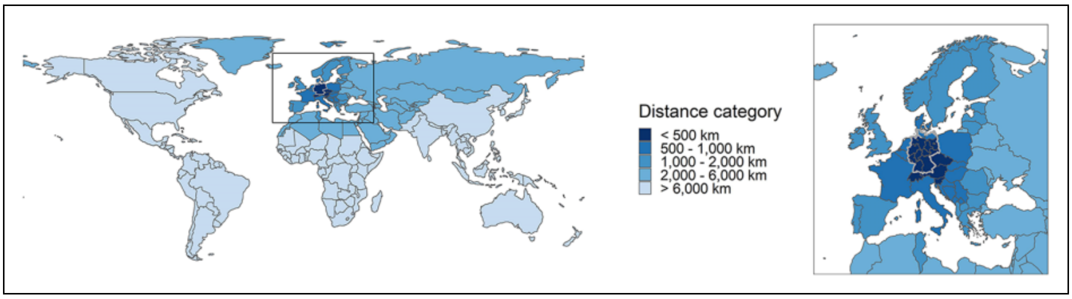

Figure 2. Travel distance categorization for travelers from Bavaria, Germany. Germany is framed in gray.

travel. Direct distances were calculated in kilometers between the region of origin (i.e. the centroid of the federal state) and the destination. For the latter, we typically use the centroid of the stated country. Information on farther destinations was often not available explicitly but only as part of its greater region (e.g. "Southeast Asia"). To distinguish short-, medium-, and long-haul travel, distances were analyzed in five categories similar to Frick et al. (2014). Distance categories and exemplary destination countries are displayed in Table 2 and Figure 2. Since travel distances are approximated from the respective federal state of origin, some European countries are not assigned to a consistent group.

Methods

Our methodological framework consists of descriptive and model-based analysis of APC structures. For descriptive visualization, we introduce *ridgeline matrices* as a novel technique. These are a two-dimensional extension of ridgeline plots (Wilke, 2018), an established tool to display densities against a secondary grouping variable. In accordance with Lexis diagrams, we display age groups along the horizontal axis and periods along the vertical axis, so that diagonals represent specific cohorts. The resulting plot layout enables a direct comparison of distributions over several age groups, periods, and cohorts. For our individual-level data setting, we exploit the survey structure by calculating travel distances based on the weighted observations, leading to representative results for German travelers within each group. Following Clements et al. (2005), our modeling approach builds on model (2), addressing the identification problem by implicitly regarding the cohort effect as a statistical interaction between age and period, represented by the diagonal of the estimated nonlinear surface. We apply semiparametric additive logistic regression to model the individual travel distance categories as binary outcomes. In accordance with Fannon et al. (2018), we use the following model structure for our repeated cross-sectional data setting with observations on individual level

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + f_{\text{ap}}(\text{age}_i, \text{period}_i), \quad i = 1, \ldots, n \tag{3}$$

with π_i the probability to travel in the respective distance category, β_0 the intercept, $f(\cdot,\cdot)$ a two-dimensional nonlinear function, and n the number of individuals. In the remainders of this work, we call model (3) the *pure APC model*. All quantitative interpretations and visualizations of effects in this study are based on odds ratios (OR $= \frac{\pi_i}{1-\pi_i}$).

We represent the two-dimensional function $f(\cdot,\cdot)$ by a tensor product basis, defined as the Kronecker product of two one-dimensional marginal spline bases over age and period (Wood, 2017). More specifically, each marginal basis function of age is multiplied with each marginal basis function of period to obtain the two-dimensional spline basis. We use penalized B-splines (Eilers and Marx, 1996) to define marginal spline bases, each of them made up of 10 basis functions and the penalization of second-order differences. The estimated tensor product surface is visualized by a heatmap, giving a compact overview of the interrelations of age, period, and cohort. To facilitate the comprehension of effect structures, we use a lower resolution of the heatmap by computing mean effects over APC groups of 5 years. Uncertainty is displayed using 95% pointwise confidence intervals (Marra and Wood, 2012). As an additional graphical tool, individual marginal effects for age, period, and cohort f_a(age), f_p(period), and f_c(cohort) offer a more accessible visualization of temporal developments by focusing on a specific dimension only. This also facilitates effect strength comparisons within and between different models. The mean marginal effects are extracted from the tensor product estimate

$$\begin{aligned} f_a(\text{age}_a) &= \frac{1}{P}\sum_{p=1}^{P} f_{ap}\left(\text{period}_p | \text{age}_a\right) \\ f_p\left(\text{period}_p\right) &= \frac{1}{A}\sum_{a=1}^{A} f_{ap}\left(\text{age}_a | \text{period}_p\right) \\ f_c(\text{cohort}_c) &= \frac{1}{A \times P}\sum_{a=1}^{A}\sum_{p=1}^{P} f_{ap}\left(\text{age}_a, \text{period}_p | \text{cohort}_c\right) \end{aligned} \tag{4}$$

To visualize the interplay of APC effects, we propose partial APC plots as an extension of marginal effect plots with specific focus on the interrelation of two selected temporal factors. In our experience, this substantially facilitates communication of the model complexity to practitioners. In addition to one marginal effect of interest, partial APC plots display appropriate slices of the tensor product where, for example, the nonlinear variation over cohorts is shown for one specific age group only.

The pure APC model aims at a descriptive interpretation of the estimated temporal structures. Causal conclusions should not be drawn as age, period, and cohort each represent underlying internal and external factors that are not directly incorporated in the model. To estimate the attribution of observed temporal developments to such factors, we integrate additional covariates into the model structure

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + f_{ap}(\text{age}_i, \text{period}_i) + \eta_i \tag{5}$$

where η_i denotes a linear predictor containing further (non)linear effects. While we primarily focus on the pure APC model in this study, we also estimate an extended covariate model by evaluating a selected set of sociodemographic and travel-related variables on individual level to exemplarily demonstrate the application of APC models in tourism research. Effects obtained through this covariate model have to be interpreted with caution since we do not account for all factors potentially associated with travel behavior.

Based on both models, we apply additive logistic regression for all five distance categories, similar to a multinomial modeling approach. Model performance is evaluated by area under the curve (AUC) values (Japkowicz and Shah, 2011), calculated on a hold-out test set comprising 20% of the data when re-estimating each model on the remaining training set. AUC values vary between 0.55 and 0.67 for the pure APC models and 0.58 to 0.83 for the models including covariates with best model performances for the highest (">6000 km": 0.66 pure model, 0.83 covariate model) and lowest ("<500 km": 0.63, 0.72) distance categories.

The changes in the underlying population in 1990 and 2010 were accounted for by sensitivity analyses. For this purpose, all models were re-calculated based only on the population of Western German travelers and German citizens, respectively. No substantial deviations from the presented results were found. Results of these analyses are listed in the Online Appendix C.

All statistical analyses were performed with the open source software R (R Core Team, 2019). Models are estimated with the function gam from the package mgcv (Wood, 2017), and all visualizations are based on the package ggplot2 (Wickham, 2016). Code for the statistical analysis is freely available in an open source GitHub repository (Weigert et al., 2020).

Results

Descriptive analysis

Over the last 50 years, Germans travel considerably longer distances for their main holiday. As illustrated in Figure 3, trips of less than 500 km have decreased, while the share of trips with a distance of 2000 km and more has steadily increased. Since 2000, the distribution of trips across all distance categories is stable with one-third of trips being conducted within 500 km.

The demand curve for pleasure trips in the most recent year 2018 (Figure 4) follows the European tourism demand pattern observed by McKercher and Mak (2019). Proceeding from the majority of holidays spent in rather close destinations, demand declines with increasing distance. The popularity of package holiday destinations in the Mediterranean regions is reflected by a secondary peak around 1800 km distance.

The two-dimensional ridgeline matrix (Figure 5) represents an extension of the demand curve and gives a first impression of the extent to which travel distances simultaneously change over age, period, and cohort. Density functions of distances are displayed on a log10 scale to focus on changes in lower distance categories. The figure reveals increasing travel distances since the 1970s as the highest peak of the densities moves towards the right for all age groups. An association

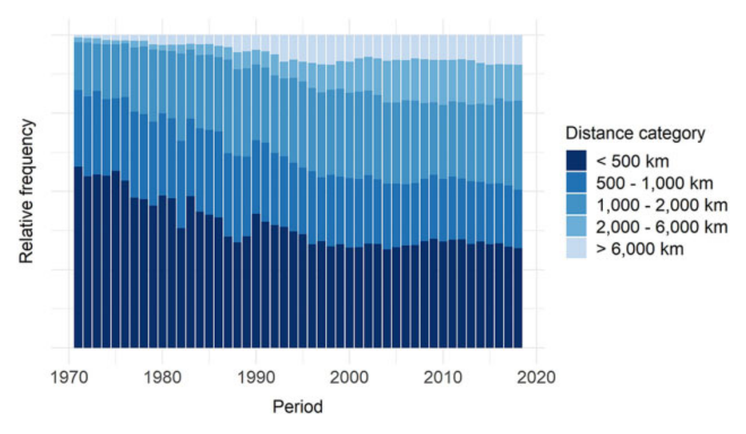

Figure 3. Relative frequency of travel distance of main holiday trips between 1971 and 2018. Some abrupt developments are caused by minor changes in the survey design.

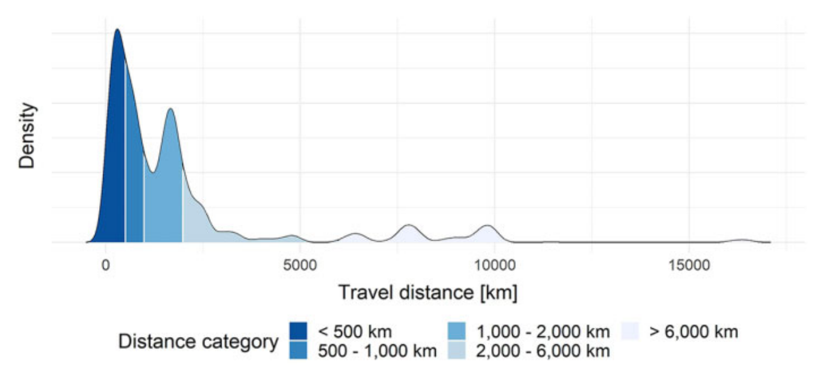

Figure 4. Distribution of travel distances of German travelers in 2018. Travel distance values 0 mark travels inside the traveler's federal state. Higher density values encode higher frequency.

between age and travel distances is visible as travel distance is consistently highest for 20- and 30-year-olds. Only minor differences are visible comparing different generations.

APC models

Pure APC model. Our attribution of the observed developments to the three temporal dimensions is based on separate logistic regression models purely conditioning on age, period, and cohort. The main model results are visualized by a heatmap of the estimated tensor product surface allowing the comparison of areas with a higher chance to travel in the examined distance category and areas with a respective lower chance. The according heatmap for distance category ">6000 km" is displayed in the left panel of Figure 6. Overall, younger age groups, recent periods, and younger cohorts are all attributed the highest chance to travel long distance. Substantial uncertainty is only

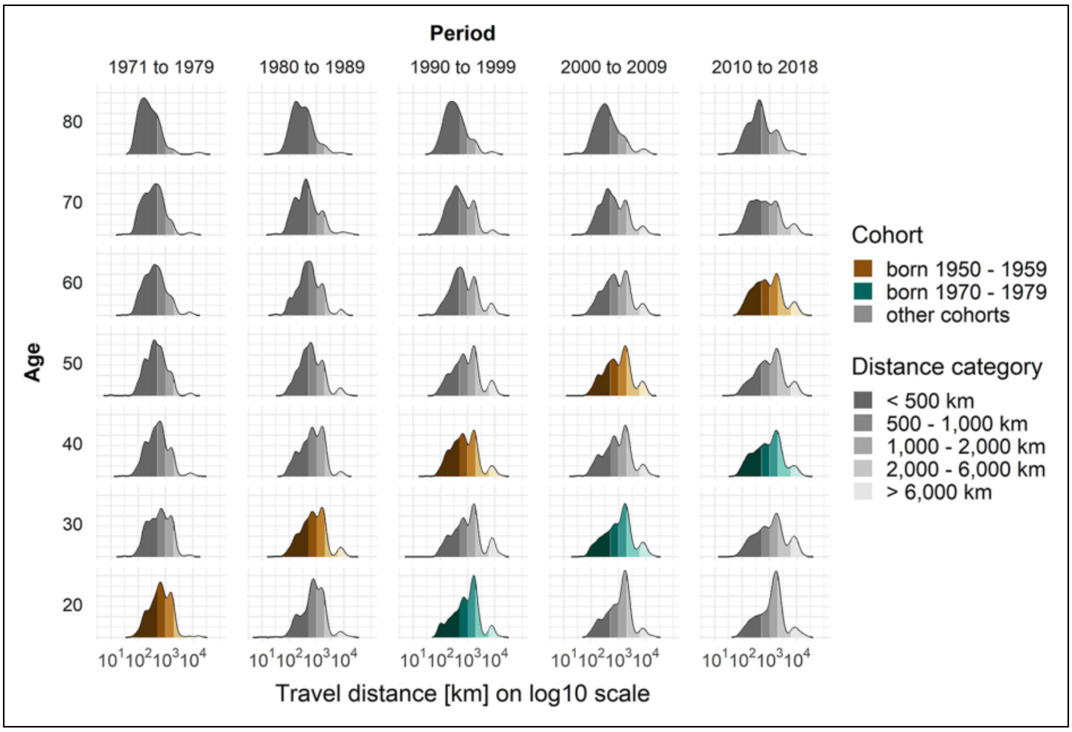

Figure 5. Ridgeline matrix depicting the development of travel distances for different age, period, and cohort groups. Cohorts born between 1950 and 1959 and between 1970 and 1979 as representatives of Baby Boomers and Generation X are exemplarily highlighted brown and green, respectively. Higher densities encode higher frequency. Distances are displayed on a log10 scale.

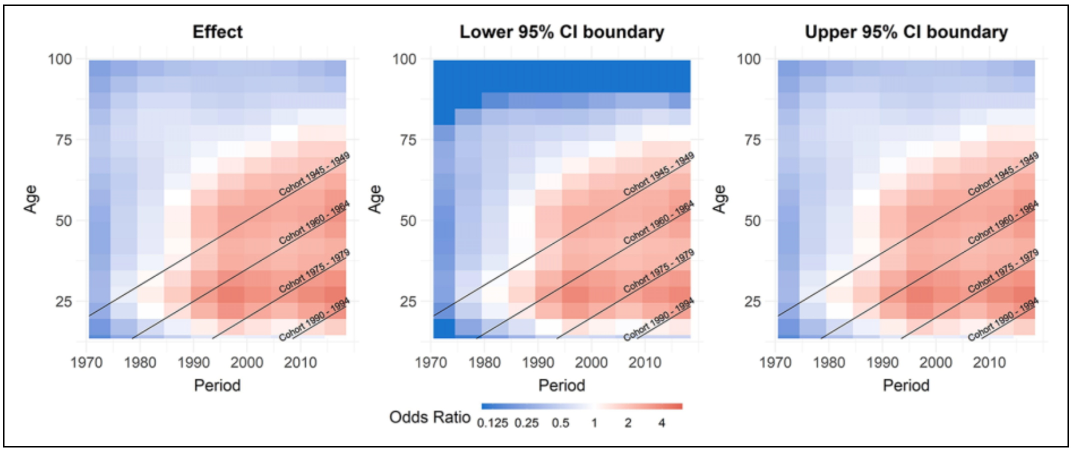

Figure 6. Heatmaps of the estimated tensor product surface (left panel) and the respective lower (center) and upper (right) 95% CI boundary for distance category ">6000 km". Effects are averaged over 5-year blocks. Exponentiated values smaller than 0.1 are trimmed to 0.1. CI: confidence interval.

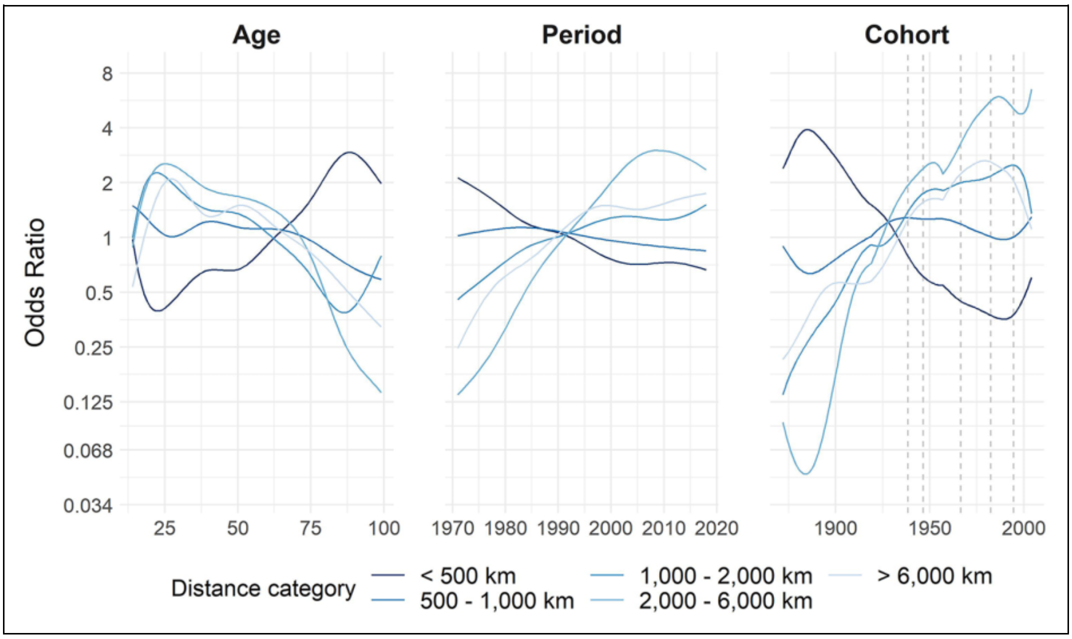

Figure 7. Estimated marginal odds ratios of age, period, and cohort for each distance category on a log2 scale. The dashed vertical lines in the cohort plot mark the boundaries between the generations defined in "Database" section.

present for age groups 90 or higher since very few such travelers were observed. The uncertainty of the effect estimates is similar for all distance category models (see the Online Appendix A).

Figure 7 shows the respective marginal effects for the temporal domains. The displayed ORs have a multiplicative interpretation, dependent on the currently focused distance category: for example, the age effect for distance category "<500 km" shows that the chance to make one's holiday trip within 500 km is about twice as high for persons aged 62 (estimated OR $\approx$ 0.98) compared to persons aged 30 (OR $\approx$ 0.49) since the effects show a difference in chance of around $+100\% \, (= \frac{0.98}{0.49} - 1)$.

Age. The age effects show pronounced differences between short- and long-haul travel. The tendency for short-haul trips within 500 km increases with age (age 23: OR $\approx$ 0.39) and reaches its peak at age 88 (OR $\approx$ 2.94). Teenagers also are more likely to travel to closer destinations. Reversed and partly bimodal age effects are obtained for travel distances over 1000 km with the highest chances around age 25 (">6000 km": OR $\approx$ 2.01) and 50 (">6000 km": OR $\approx$ 1.50). The dip between 35 years and 45 years becomes more pronounced with increasing distances and is most visible in the distance category ">6000 km" (age 41: OR $\approx$ 1.30). From the mid-50s onwards, the chance for long-distance holidays decreases continuously.

The results are in accordance with life cycle theory (e.g. Collins and Tisdell, 2002; Oppermann, 1995). The increase in choosing longer distance travel between the ages of 14 and 30 might be explained by increasing travel experience. Teenagers most commonly are in the early stages of their travel careers (independent from their parents) and prefer more familiar, low-risk destinations closer to home. In contrast, self-sufficient young people in their 20s seem to become more adventurous and therefore are more likely to choose distant destinations (Karl, 2018). The changing marital status between mid-30 and mid-40 associated with parenting reduces preference for long-haul trips as families with dependent children prefer easily accessible and safe destinations (Collins and Tisdell, 2002; Karl, 2018). With the transition to the empty nest stage, the demand for distant destinations is growing again due to the reduction of travel constraints (Bernini and Cracolici, 2015). The decline in travel distance with age can be explained by decreasing physical health, limited mobility, and reduced disposable income of the elderly (You and O'leary, 2000).

Period. Regarding the period effect, long-distance travels have strongly increased over the last decades. Particularly, the effect structures of short- and long-distance travels reversed around the year 1992, suggesting that Germans were inclined to visit rather close destinations beforehand, while the tendency to choose distant destinations has been increasing ever since. The strongest period effect is observed for destinations within "2000–6000 km" (1971: OR $\approx$ 0.14; 2018: OR $\approx$ 2.35) as well as ">6000 km" (1971: OR $\approx$ 0.25; 2018: OR $\approx$ 1.75). Since 2000, the effects are comparably stable.

The growth in travel distance over time can be mainly explained by technological developments in transportation which have led to distant destinations being faster and more cheaply accessible, especially for low-income populations (Castro et al., 2020). Additionally, the regular confrontation with information about foreign destinations increased with the advent of digital media (Beldona, 2005; Kim et al., 2015) leading to a decrease in perceived distance (Yang et al., 2018). According to positive correlations between the economic situation of the source market and outbound travel participation (Sun and Lin, 2019; Wong et al., 2016), it can be assumed that the growing economy of Germany is reflected in the period effect. The increase in travel within 2000–6000 km distance can be attributed to the growing popularity of specific destinations, such as Turkey or Egypt. The emerging saturation since 2000 can be linked to declining population dynamics and flattening economic growth rates (Frick et al., 2014).

Cohort. The cohort effect shows a clear association between generational affiliation and distance traveled. Overall, younger generations show a greater chance for overseas travels and lower tendencies for short-haul trips than older generations. For instance, members of Generation Y (mean OR $\approx$ 5.69) have more than twice the chance to travel within 2000–6000 km compared to Baby Boomers (mean OR $\approx$ 2.59), while travels to destinations within 500 km are about 47% more likely for Baby Boomers (OR $\approx$ 0.53) than for Generation Y (mean OR $\approx$ 0.36). In parts, the cohort effect reflects the observed age differences since, for example, our data on the youngest cohorts only comprise teenagers. Further detail is given in the discussion of Figure 8.

The cohort effect is in line with other research showing that younger generations travel to more distant (Oppermann, 1995) and international destinations (Pennington-Gray et al., 2002). The higher probability of long-distance travel among young cohorts can be attributed to the socialization processes, also shaped by advances in transport and communication technologies in formative years (Oppermann, 1995), increasing the potential to gain greater travel experiences in childhood. This is closely linked to younger cohorts showing higher tendencies to be novelty seekers that prefer nonmainstream destinations (Li et al., 2013).

Comparison of effects. While substantially varying travel distances are observed over all temporal dimensions, the marginal association structures show differences in their effect strengths. As given in Table 3, the chance for short-haul trips, especially those under 500 km, is mainly associated with

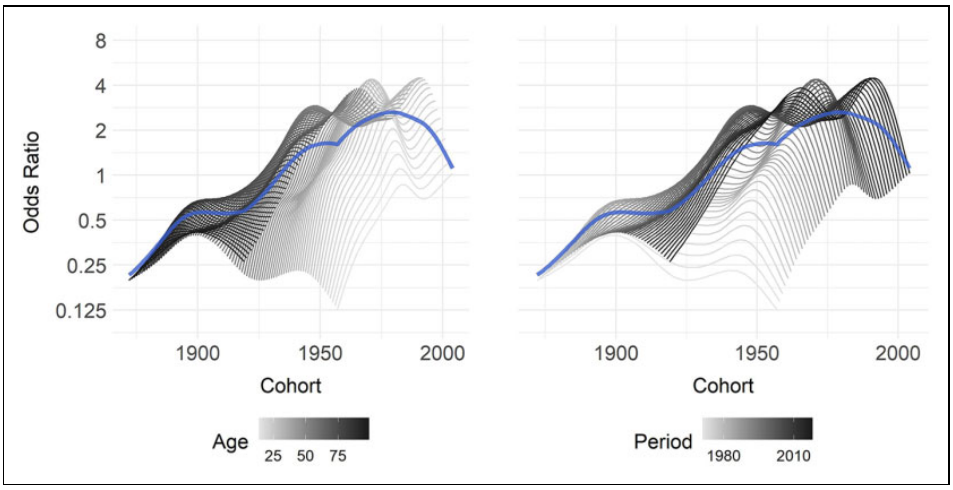

Figure 8. Partial APC plot of estimated odds ratios for the cohort effect dependent on age group (left panel) and period (right) for the model ">6000 km." The mean marginal effect is marked as bold blue line. APC: age–period–cohort.

Table 3. Overview of marginal effects of the pure APC model (see Figure 7).

Model	Effect	Value with maximum OR	Value with minimum OR	Maximum OR	Minimum OR	Ratio
<500 km	Age	88	23	2.94	0.39	**7.5**
	Period	1971	2018	2.13	0.66	3.2
	Cohort	1939	1989	0.77	0.36	2.1
500–1000 km	Age	14	99	1.49	0.59	**2.5**
	Period	1983	2018	1.14	0.85	1.3
	Cohort	2004	1989	1.30	0.97	1.3
1000–2000 km	Age	22	86	2.27	0.39	**5.8**
	Period	2018	1971	1.51	0.46	3.3
	Cohort	1994	2004	2.50	1.35	1.9
2000–6000 km	Age	25	99	2.54	0.14	18.1
	Period	2009	1971	3.01	0.14	**21.5**
	Cohort	2004	1939	6.54	1.97	3.3
>6000 km	Age	27	99	2.10	0.32	6.6
	Period	2018	1971	1.75	0.25	**7.0**
	Cohort	1979	2004	2.64	1.11	2.4

Note: APC: age–period–cohort; OR: odds ratio. For each model and effect, the following information is listed, from left to right. Variable value where the OR reaches its maximum/minimum; maximum/minimum of the OR; ratio between the respective maximum OR and minimum OR. The maximum ratios per model are highlighted in bold. According to the generations defined in "Database" section, cohort effects are considered for birth years from 1939 onwards only.

age differences. Long-distance travel predominantly varies over the period. Particularly within destinations in 2000–6000 km distance, a noteworthy period effect is shown, underlining the findings of the marginal effects. Differences between generations born from 1939 onwards are less pronounced regarding travel distances. Overall, the effect strengths underpin the developments visualized in the ridgeline matrix (Figure 5).

The observed tendencies may imply that choosing short-haul destinations depends on personal characteristics and age-related travel constraints such as physical or family restrictions (You and O'leary, 2000). Contrarily, long-distance travel might be more constrained by macro-level factors such as developments in transport technology attributed to reduced costs for long-haul travel or economic growth in the source market leading to an increase in disposable income which can be used for more expensive long-distance travel (Sun and Lin, 2019). In Germany, the positive economic development in the time period investigated in this study has made long-term travels considerably more accessible. Moreover, technological advances in transportation (Castro et al., 2020) have considerably reduced flight costs for German travelers in the past five decades. The access to a wider range of information sources through modern communication technologies further reduces perceived (cultural) distances to previously inaccessible destinations (Yang et al., 2018), which might also explain the strong increase in long-distance travel over time. In comparison with age and period, generational membership seems to be less important for alterations in travel distances. This broadens previous insights of cohort studies that focused on generational differences without accounting for all three temporal dimensions (e.g. Bernini and Cracolici, 2015).

Interrelations of age, period, and cohort. Figure 8 shows an exemplary partial APC plot that highlights the interrelations between the temporal dimensions. In addition to the marginal cohort effect, it includes one line for each *partial cohort effect*, that is, for the estimated differences between cohorts when just focusing on travelers with a specific age (left panel) or travels in a specific period (right panel). The displayed effects originate from the model for long-distance travels ">6000 km".

The figure displays different kinds of information: First, it shows which cohort entails observations in which age group or period, as already listed in Table 1. For instance, the age-dependent plot highlights that the youngest cohorts solely comprise observations of teenagers. Secondly, it is easily deducible how substantial this partial observation structure affects the marginal cohort effect. The drop in the marginal effect for the youngest cohorts can be fully traced back to the observed age groups, since teenagers travel distinctively lower distances than travelers in their 20s or 30s. Thirdly, the partial cohort effects separately display the cohort differences in each age group and period. For example, 14-year-olds show less variation over the observed cohorts than 20-year-olds. Overall, very young (light gray lines) and very old (dark gray lines) people show more consistent travel distances than middle-aged, generally less constrained travelers. Regarding the cohort differences for a given period, more consistent travel distances are observed in the older periods (light gray) compared to the most current periods (dark gray). The partial age and period effects and the effects on other distance categories are given in the Online Appendix B.

Covariate model. Travel behavior is shaped by several internal and external factors (Moutinho, 1987). To showcase the integration of additional factors associated with travel behavior, particularly destination choice, we extend the pure APC model for distance category ">6000 km" by the following internal covariates: (i) inflation-adjusted household net income (as a nonlinear effect)—the integration of this covariate is motivated by the income elasticity of tourism demand which assumes an overall positive correlation, especially for outbound travel participation (Eugenio-Martin and Campos-Soria, 2011); (ii) household size—this reflects the marital, financial, and

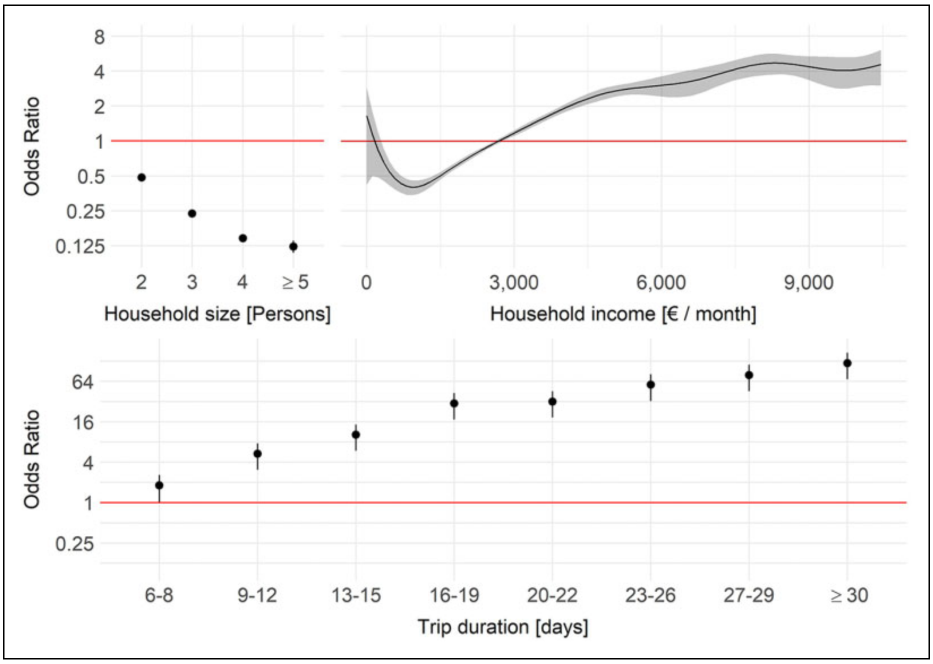

Figure 9. Estimated odds ratios of the variables household income, household size (reference category: one-person household) and trip duration (reference category: trips of 5 days) in the model for the distance category ">6000 km" on a log2 scale. Uncertainty is displayed by 95% confidence intervals.

social situation of an individual and the related travel constraints (Alegre and Pou, 2004). It is well-established that the larger the household, the less likely it is to travel abroad (Guillet et al., 2011); (iii) trip duration is included because of its positive association to travel distances since higher costs owing to longer distances are compensated by longer trip lengths (Jackman et al., 2020). External, macro-level factors like the economic climate in the source market and technological developments are not explicitly accounted for. As outlined, these are jointly reflected by the period effect.

Overall, Figure 9 illustrates substantial associations between the chance to travel farther than 6000 km and all included covariates. While trip duration shows a very strong positive association, the chance for long-distance travel decreases continuously with increasing household size. More specifically, someone from a two-person household has on average a 293% ($\approx \frac{0.48}{0.12} - 1$) higher chance to travel farther than 6000 km than a member of a $\geq$5-person household. Regarding the positive income effect, the chance for long-distance travel increases almost linearly within income levels of 1000–5000 € before flattening for high-income values (1000 €: OR $\approx$ 0.40; 8000 €: OR $\approx$ 4.63).

The inclusion of additional covariates alters the strengths of the estimated age, period, and cohort effects compared to the pure APC model (Figure 10). Accounting for household size attenuates the age-related dip around 40. This can be related to the impact of travel constraints caused by specific personal circumstances of this age group. The variations in the period and cohort effect are mainly triggered by conditioning on the trip duration. More specifically, assuming trips of equal length, the chance for holiday trips over 6000 km increases more steeply both over

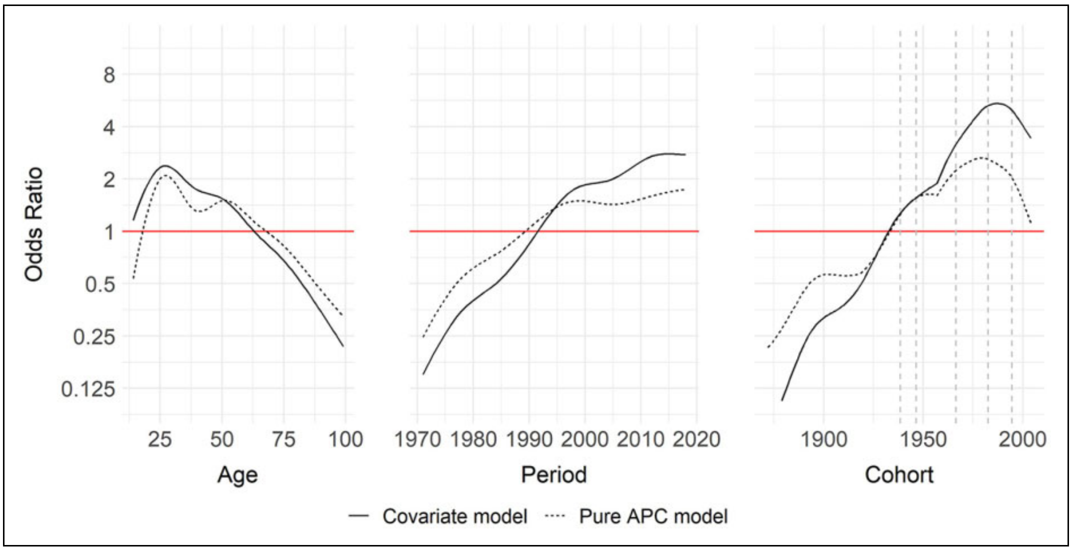

Figure 10. Estimated odds ratios of the full model for distance category ">6000 km versus <6000 km" including covariates (covariate model) compared to the APC effects in the model without covariates (pure APC model) on a log2 scale. The dashed vertical lines in the cohort plot mark the boundaries of the analyzed generations. APC: age–period–cohort.

time and across generations underlining the higher affordability and easier accessibility of long-haul trips in recent years and for younger cohorts.

Conclusion

Comprehension of changing destination choice patterns requires a thorough understanding of all temporal interrelations and their respective drivers. APC analysis is an established tool for the investigation of such time-related changes, separating cohort effects from age and period (Yang and Land, 2013). Nevertheless, this method and its possibilities for the examination of behavioral changes in tourism science are not yet fully exploited. The purpose of this study was threefold: (1) establish a flexible, state-of-the-art statistical modeling approach for APC analysis in tourism science, (2) introduce ridgeline matrices and partial APC plots as novel graphical tools for visualizing complex APC structures in a comprehensive way, and (3) generate new insights into temporal developments of travel behavior by applying the comprehensive APC approach, based on a rich secondary data set.

From a methodological perspective, our contributions focus on refining and showcasing a widely applicable and well accessible APC approach. We build our modeling framework on semiparametric GAMs to overcome the identification problem without using restrictive assumptions on the temporal effects. The approach can be formulated for aggregated and individual data settings (i.e. travel information for groups of travelers or for individual travelers) as well as repeated cross-sectional or panel data, making it adaptable to a variety of research settings. The GAM framework is highly flexible and offers robust and efficient concepts for estimation and inference, accompanied by sophisticated and freely available software. Since APC analyses rely on a thorough understanding of the temporal interrelations, we offer innovative visualizations to facilitate their comprehension. Ridgeline matrices present an easily accessible tool for displaying three-dimensional temporal changes. On a model-based level, partial APC plots are a novel way to visualize bivariate interrelations of the individual effects. Their application is vital in APC analyses since birth cohorts are most commonly not observed across all available age groups and periods. In such settings, partial APC plots can help understand to which extent the estimated association structures are due to the specific data structure.

From a tourism perspective, the holistic APC framework contributes to new insights about temporal changes of destination choice. It offers a more comprehensive alternative to previously applied approaches in tourism research. In combination with our long-term data set comprising travel behavior on individual level, it allows a deeper understanding of the temporal structures. Our study confirms that alterations in travel behavior occur in accordance with life cycle theory (age), macro-level developments in economy and society (period), and generational theory (cohort). Moreover, we produced new insights on the main temporal drivers that alter destination choices. In contrast to cohort studies focusing on generational differences (e.g. Huang and Lu, 2017), our findings especially suggest that cohort differences seem less pronounced when all three temporal dimensions are considered. Contrary to common approaches in studies on changes in travel behavior (e.g. Bernini and Cracolici, 2015), our simultaneous analysis of developments over age, period, and cohort does not neglect any specific interrelations in the temporal structures. Since destination choice is determined by various internal and external factors (Wong et al., 2017), another benefit of our framework is the possibility to incorporate additional explanatory variables. For example, our approach allows to investigate to what extent the observed developments can be attributed to specific characteristics of the travel decision process. If data are modeled on individual level, this most notably comprises the simultaneous incorporation of explaining variables on individual (e.g. income of the traveler) and macro level (e.g. general economic indices). Especially for studies observing individual travelers, accounting for further variables offers great potential for future tourism research by identifying the influences and interactions of socioeconomic and travel-related factors (e.g. travel motivation or transportation mode). Often it is the interplay between such internal and external factors, related to the tourist and the destination, that shapes travel decision-making (Karl, 2018) and consequently tourism demand. For instance, the individual motivation to travel and the price level at and transport costs to a destination commonly influence tourists' destination choices (Nicolau and Más, 2006). Finally, both the academic sector and the industry can make use of these newly generated insights. Understanding which and how different factors cause changes in travel behavior (e.g. domestic and outbound tourist flows) may lead to better predictions of future tourism demand, supporting touristic stakeholders in tourism planning and management.

Using repeated cross-sectional data from an established long-term survey has its advantages (valid and representative sample, cost-efficiency) but is not without limitations. A critical aspect which had to be considered in the analysis were changes of the underlying population in 1990 after German reunification, and in 2010 from German citizens to the German-speaking population. However, our comparative sensitivity analyses show that these modifications have no substantial impact on the overall results. Regarding our main findings, transferability is limited to source markets like Western European countries, with conditions similar to the German market (economic situation, transportation system and freedom of travel). Future studies should apply our proposed approach to other source markets to investigate how APC structures change when different macroeconomic and sociocultural conditions are considered. Finally, as generally is the case in observational studies, findings should be interpreted with caution regarding causal relationships.

The overall APC effects can be differentiated reasonably well. Causal conclusions, however, should not be drawn as all observed temporal interdependencies can be traced back to specific socioeconomic factors, societal changes, or shared socialization processes that affect each individual tourist. Accordingly, a clear distinction between these underlying factors is not possible without explicitly accounting for them as further covariates in the model. To make accurate predictions of tourism demand, future research needs to focus on integrating relevant factors from tourism demand modeling (e.g. economic development and technological advancement) and the travel decision-making literature (see review by Smallman and Moore, 2010) into the outlined analyses of individual travel data. Among others, this includes the impact of economic changes in the source market, technological advancement leading to reduced transport costs or political events (e.g. German reunification and establishment of the Schengen area) on destination choices regarding short- or long-haul destinations.

Due to the identification problem, separating the effects of age, period, and cohort remains the crucial challenge in APC analysis. Generally speaking, the GAM framework is an adequate basis for estimating mean APC structures in widespread research settings. Since perfect separation of the three temporal effects is not possible, future research should specifically focus on tools to make association structures more accessible. Especially model-related visualization techniques such as partial APC plots are promising to be further refined for this purpose. Ridgeline matrices are extendable by further building on additional concepts of ridgeline plots (Wilke, 2018) to display the distribution of each matrix cell conditional on, for example, socioeconomic groups. While we designed both novel visualization techniques specifically for the application on cross-sectional data, their adaptation to panel data remains to be evaluated. Regarding our application, alternative statistical approaches to model travel distances should be taken into consideration. In the regression context, this comprises modeling the raw distances as a response as well as the evaluation of more complex techniques like functional data analysis (Bauer et al., 2018) to compare the demand curves between different groups.

In conclusion, the outlined modeling approach proved its worth in the application on travel distances and contributed to deeper knowledge in destination choice. Combining the flexibility of semiparametric regression with modern visualization tools offers great potential for future studies analyzing temporal changes in diverse fields of (tourism) research.

Acknowledgments

The authors would like to thank Forschungsgemeinschaft Urlaub und Reisen e.V. for providing the underlying data and help regarding questions of any type to the data set. They also thank the anonymous reviewer whose insightful comments were highly appreciated and helped making the article more accessible.

Declaration of conflicting interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work has been funded by the German Research Foundation (DFG) under Grant Nos KU 1359/4-1, SCHM 850/22-1, and KA 4976/2-1 and by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A. The authors of this work take full responsibility for its content.

ORCID iD

Maximilian Weigert https://orcid.org/0000-0003-4400-134X
Marion Karl https://orcid.org/0000-0002-0666-4997

Supplemental material

Supplemental material for this article is available online.

References

Alegre J and Pou L (2004) Micro-economic determinants of the probability of tourism consumption. *Tourism Economics* 10(2): 125–144.

Bauer A, Scheipl F, Küchenhoff H, et al. (2018) An introduction to semiparametric function-on-scalar regression. *Statistical Modelling* 18(3-4): 346–364.

Becken S and Carmignani F (2020) Are the current expectations for growing air travel demand realistic? *Annals of Tourism Research* 80: 102840.

Beldona S (2005) Cohort analysis of online travel information search behavior: 1995-2000. *Journal of Travel Research* 44(2): 135–142.

Bernini C and Cracolici MF (2015) Demographic change, tourism expenditure and life cycle behaviour. *Tourism Management* 47: 191–205.

Bowen D and Clarke J (2009) *Contemporary Tourist Behaviour: Yourself and Others as Tourists*. Wallingford: CABI Tourism Texts. CABI.

Carstensen B (2007) Age–period–cohort models for the lexis diagram. *Statistics in Medicine* 26(15): 3018–3045.

Castro R, Lohmann G, Spasojevic B, et al. (2020) The future past of aircraft technology and its impact on stopover destinations. In: Yeoman I and McMahon-Beattie U (eds) *The Future Past of Tourism, The Future of Tourism*. Bristol: Channel View Publications, pp. 93–104.

Chen SC and Shoemaker S (2014) Age and cohort effects: the American senior tourism market. *Annals of Tourism Research* 48: 58–75.

Clayton D and Schifflers E (1987) Models for temporal variation in cancer rates. II: age–period–cohort models. *Statistics in Medicine* 6(4): 469–481.

Clements MS, Armstrong BK and Moolgavkar SH (2005) Lung cancer rate predictions using generalized additive models. *Biostatistics* 6(4): 576–589.

Cohen SA, Prayag G and Moital M (2014) Consumer behaviour in tourism: concepts, influences and opportunities. *Current Issues in Tourism* 17(10): 872–909.

Collins D and Tisdell C (2002) Age-related lifecycles: purpose variations. *Annals of Tourism Research* 29(3): 801–818.

Cooper C and Hall CM (2016) *Contemporary Tourism: An International Approach*. 3rd edn. Oxford: Goodfellow Publishers Ltd.

Diggle P, Diggle PJ, Heagerty P, et al. (2002) *Analysis of Longitudinal Data*. Oxford: Oxford University Press.

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

Eugenio-Martin JL and Campos-Soria JA (2011) Income and the substitution pattern between domestic and international tourism demand. *Applied Economics* 43(20): 2519–2531.

Fannon Z, Monden C and Nielsen B (2018) *Age-period-cohort modelling and covariates, with an application to obesity in England 2001-2014 (No. 2018-W05). Economics Group, Nuffield College: University of Oxford.*

Fienberg SE and Mason WM (1979) Identification and estimation of age-period-cohort models in the analysis of discrete archival data. *Sociological Methodology* 10: 1–67.

Forschungsgemeinschaft Urlaub und Reisen eV (FUR) (2020a) Reiseanalyse. How Germany travels. Available at: https://reiseanalyse.de/reiseanalyse-2018/ (accessed 14 May 2020).

Forschungsgemeinschaft Urlaub und Reisen eV (FUR) (2020b) Survey of tourist demand in Germany for holiday travel and short breaks. Available at: https://reiseanalyse.de/wp-content/uploads/2019/08/RA2020_Infoflyer_EN.pdf (accessed 14 May 2020).

Frick R, Belart B, Schmied M, et al. (2014) Langstreckenmobilität: Aktuelle Trends und Perspektiven: Grundlagenstudie. Grundlagenstudie.

Fu WJ (2000) Ridge estimator in singulah oesiun with application to age-period-cohort analysis of disease rates. *Communications in Statistics-Theory and Methods* 29(2): 263–278.

Glover P and Prideaux B (2009) Implications of population ageing for the development of tourism products and destinations. *Journal of Vacation Marketing* 15(1): 25–37.

Gössling S, Scott D, Hall CM, et al. (2012) Consumer behaviour and demand response of tourists to climate change. *Annals of Tourism Research* 39(1): 36–58.

Guillet BD, Lee A, Law R, et al. (2011) Factors affecting outbound tourists' destination choice: the case of Hong Kong. *Journal of Travel & Tourism Marketing* 28(5): 556–566.

Herhoffer PA and Meurer J (2018) Konsumentengenerationen 2018: Premium- und Luxus-Studie. Available at: http://www.keylens.com/wp-content/uploads/2018/05/Konsumgenerationen-2018_Branchenreport-Touristik.pdf (accessed 18 February 2020).

Heuer C (1997) Modeling of time trends and interactions in vital rates using restricted regression splines. *Biometrics* 53: 161–177.

Holford TR (1983) The estimation of age, period and cohort effects for vital rates. *Biometrics* 39: 311–324.

Huang Q and Lu Y (2017) Generational perspective on consumer behavior: China's potential outbound tourist market. *Tourism Management Perspectives* 24: 7–15.

Jackman M, Lorde T, Naitram S, et al. (2020) Distance matters: the impact of physical and relative distance on pleasure tourists' length of stay in Barbados. *Annals of Tourism Research* 80: 102794.

Japkowicz N and Shah M (2011) *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge, MA: Cambridge University Press.

Karl M (2018) Risk and uncertainty in travel decision-making: tourist and destination perspective. *Journal of Travel Research* 57(1): 129–146.

Karl M, Winder G and Bauer A (2017) Terrorism and tourism in Israel: analysis of the temporal scale. *Tourism Economics* 23(6): 1343–1352.

Kim H, Xiang Z and Fesenmaier DR (2015) Use of the internet for trip planning: a generational analysis. *Journal of Travel & Tourism Marketing* 32(3): 276–289.

Kupper LL, Janis JM, Karmous A, et al. (1985) Statistical age-period-cohort analysis: a review and critique. *Journal of Chronic Diseases* 38(10): 811–830.

Lee HA, Guillet BD, Law R, et al. (2012) Robustness of distance decay for international pleasure travelers: a longitudinal approach. *International Journal of Tourism Research* 14(5): 409–420.

Lehto XY, Jang S, Achana FT, et al. (2008) Exploring tourism experience sought: a cohort comparison of Baby Boomers and the Silent Generation. *Journal of Vacation Marketing* 14(3): 237–252.

Li X, Li X and Hudson S (2013) The application of generational theory to tourism consumer behavior: an American perspective. *Tourism Management* 37: 147–164.

Lohmann M and Danielsson J (2001) Predicting travel patterns of senior citizens: how the past may provide a key to the future. *Journal of Vacation Marketing* 7(4): 357–366.

Marra G and Wood SN (2012) Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics* 39(1): 53–74.

McKercher B and Mak B (2019) The impact of distance on international tourism demand. *Tourism Management Perspectives* 31: 340–347.

McKercher B, Chan A and Lam C (2008) The impact of distance on international tourist movements. *Journal of Travel Research* 47(2): 208–224.

McKercher B, Lai B, Yang L, et al. (2020) Travel by Chinese: a generational cohort perspective. *Asia Pacific Journal of Tourism Research* 25(4): 341–354.

Moutinho L (1987) Consumer behaviour in tourism. *European Journal of Marketing* 17: 872–909.

Nelder JA and Wedderburn RW (1972) Generalized linear models. *Journal of the Royal Statistical Society: Series A (General* 135(3): 370–384.

Nicolau JL and Más FJ (2006) The influence of distance and prices on the choice of tourist destinations: the moderating role of motivations. *Tourism Management* 27(5): 982–996.

Oppermann M (1995) Travel life cycle. *Annals of Tourism Research* 22(3): 535–552.

Pendergast D (2010) Getting to know the Y Generation. In: Benckendorff P, Moscardo G and Pendergast D (eds) *Tourism and Generation Y*. Wallingford: CAB International, pp. 1–15.

Pennington-Gray L and Spreng RA (2002) Analyzing changing preferences for pleasure travel with cohort analysis. *Tourism Analysis* 6: 109–121.

Pennington-Gray L, Fridgen JD and Stynes D (2003) Cohort segmentation: an application to tourism. *Leisure Sciences* 25(4): 341–361.

Pennington-Gray L, Kerstetter DL and Warnick R (2002) Forecasting travel patterns using Palmore's cohort analysis. *Journal of Travel & Tourism Marketing* 13(1-2): 125–143.

R Core Team (2019) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/ (accessed 4 January 2021).

Romagosa F (2020) The COVID-19 crisis: opportunities for sustainable and proximity tourism. *Tourism Geographies* 22: 690–694.

Scheiner J and Holz-Rau C (2013) A comprehensive study of life course, cohort, and period effects on changes in travel mode use. *Transportation Research Part A: Policy and Practice* 47: 167–181.

Schewe CD and Noble SM (2000) Market segmentation by cohorts: the value and validity of cohorts in America and abroad. *Journal of Marketing Management* 16(1-3): 129–142.

Schmid VJ and Held L (2007) Bayesian age-period-cohort modeling and prediction-BAMP. *Journal of Statistical Software* 21(8): 1–15.

Smallman C and Moore K (2010) Process studies of tourists' decision-making. *Annals of Tourism Research* 37(2): 397–422.

Song H and Li G (2008) Tourism demand modelling and forecasting—a review of recent research. *Tourism Management* 29(2): 203–220.

Song H, Qiu RT and Park J (2019) A review of research on tourism demand forecasting: launching the annals of tourism research curated collection on tourism demand forecasting. *Annals of Tourism Research* 75: 338–362.

Sun YY and Lin PC (2019) How far will we travel? A global distance pattern of international travel from both demand and supply perspectives. *Tourism Economics* 25(8): 1200–1223.

Taylor CE and Knudson DM (1973) Area preferences of midwestern campers. *Journal of Leisure Research* 5(2): 39–48.

Tobler WR (1970) A computer movie simulating urban growth in the Detroit region. *Economic Geography* 46(sup1): 234–240.

Weigert M, Bauer A and Nalmpatian A (2020) TravelDistAPC: Supplementary code. Available at: https://github.com/MaxWeigert/TravelDistAPC (accessed 4 January 2021).

Wickham H (2016) *ggplot2: Elegant Graphics for Data Analysis*. Berlin: Springer.

Wilke C (2018) ggridges: ridgeline plots in "ggplot2." R package version 0.5. 0. R foundation for statistical computing, Vienna, Austria.

Wong IA, Fong LHN and Law R (2016) A longitudinal multilevel model of tourist outbound travel behavior and the dual-cycle model. *Journal of Travel Research* 55(7): 957–970.

Wong IA, Law R and Zhao X (2017) When and where to travel? A longitudinal multilevel investigation on destination choice and demand. *Journal of Travel Research* 56(7): 868–880.

Wood SN (2017) *Generalized Additive Models: An Introduction with R*. Boca Raton, FL: CRC Press.

Yang Y and Land KC (2008) Age–period–cohort analysis of repeated cross-section surveys: fixed or random effects? *Sociological Methods & Research* 36(3): 297–326.

Yang Y and Land KC (2013) *Age-Period-Cohort Analysis: New Models, Methods, and Empirical Applications*. Boca Raton, FL: CRC Press.

Yang Y, Liu H, Li X, et al. (2018) A shrinking world for tourists? Examining the changing role of distance factors in understanding destination choices. *Journal of Business Research* 92: 350–359.

You X and O'leary JT (2000) Age and cohort effects: an examination of older Japanese travelers. *Journal of Travel & Tourism Marketing* 9(1-2): 21–42.

Author biographies

Maximilian Weigert is a PhD candidate at the Statistical Consulting Unit StaBLab at the Department of Statistics at LMU Munich. His research focuses on the application of advanced statistical and machine learning methods in spatiotemporal data settings in interdisciplinary projects. Besides his research, he works as a statistical consultant for researchers from diverse disciplines at LMU Munich.

Alexander Bauer is a PhD candidate at the Statistical Consulting Unit StaBLab at the Department of Statistics at LMU Munich. His research focuses on the application of advanced statistical methods in interdisciplinary research settings, primarily driven by spatiotemporal research questions, as for example, arising in tourism research. In addition to his research, he works as a statistical consultant for researchers from diverse disciplines at LMU Munich.

Johanna Gernert is a PhD candidate at the Department of Geography at LMU Munich. Her research focuses on touristic travel behavior in a spatiotemporal context. Her research mainly focuses on changes in destination choice, with special emphasis on the investigation of internal and external influencing factors.

Marion Karl is a postdoctoral research fellow at The University of Queensland and lecturer at LMU Munich. Her research focuses on internal and external influencing factors of travel decision-making and travel behavior. She applies a geographic perspective to explain travel behavior from both a tourist and a destination perspective.

Asmik Nalmpatian is a master's student in statistics with specialization in theory at the Department of Statistics at LMU Munich. She holds a bachelor's degree (BSc) in statistics. She is currently working as a research assistant at the Statistical Consulting Unit StaBLab to support researchers from diverse disciplines at LMU Munich.

Helmut Küchenhoff is Professor at the Department of Statistics at LMU Munich. He is Head of the Statistical Consulting Unit StaBLab. He is involved in many interdisciplinary projects in applied statistics.

Jürgen Schmude holds the Chair for Economic Geography and Tourism Research at the Department for Geography at LMU Munich (since 2008). He is Head of the Department and leader of the teaching and research unit, Economic Geography. He has also been president of the German Society of Tourism Research since November 2015. His research is focused on travel behavior, safety and security in tourism and on climate change in tourism.

9. APCtools: Descriptive and Model-based Age-Period-Cohort Analysis

Contributing article

Bauer, A., Weigert, M., and Jalal, H. (2022b). APCtools: Descriptive and Model-based Age-Period-Cohort Analysis. *Manuscript submitted for publication.*

Code repository

https://github.com/bauer-alex/APCtools

Author contributions

Alexander Bauer and Maximilian Weigert jointly created the concept for the R package and the publication as well as the R codebase for Weigert et al. (2021) which was used as basis for most parts of the R package. Alexander Bauer performed the major part of extensively polishing, partly re-writing and documenting all existing functions and in creating new convenience functions and unit tests. Maximilian Weigert substantially contributed to this process. All parts of the paper were jointly written by Alexander Bauer and Maximilian Weigert. Hawre Jalal wrote the code for the hexamaps visualization, contributed to the literature review of established R packages that deal with APC analysis and proofread the paper. This publication was produced as part of the research project TourIST (Tourism in Space and Time) conceived by Marion Karl, Jürgen Schmude, Alexander Bauer and Helmut Küchenhoff and which was funded by the German Research Foundation (DFG) under Grant Nos KU 1359/4-1, SCHM 850/22-1, and KA 4976/2-1.

APCtools: Descriptive and Model-based Age-Period-Cohort Analysis

Alexander Bauer
Statistical Consulting Unit StaBLab, Department of Statistics, LMU Munich, Germany

Maximilian Weigert
Statistical Consulting Unit StaBLab, Department of Statistics, LMU Munich, Germany

Hawre Jalal
Department of Health Policy and Management, Graduate School of Public Health, University of Pittsburgh, United States

Summary

Age-Period-Cohort (APC) analysis aims to determine relevant drivers for long-term developments and is used in many fields of science (Yang and Land 2013). The R package `APCtools` offers modern visualization techniques and general routines to facilitate the interpretability of the interdependent temporal structures and to simplify the workflow of an APC analysis. Separation of the temporal effects is performed utilizing a semiparametric regression approach. We shortly discuss the challenges of APC analysis, give an overview of existing statistical software packages and outline the main functionalities of the package.

Keywords: R; Statistical analysis; APC analysis; Age-period-cohort analysis; Hexamaps

Statement of Need

The main focus in APC analysis is on disentangling the interconnected effects of age, period, and cohort. Long-term developments of some characteristic can either be associated with changes in a person's life cycle (age), macro-level developments over the years that simultaneously affect all age groups (period), or the generational membership of an individual, shaped by similar socialization processes and historical experiences (cohort).

The critical challenge in APC analysis is the linear dependency of the components age, period, and cohort (cohort = period - age). Flexible methods and visualization techniques are needed to circumvent this *identification problem*. Several packages for APC analysis exist for the statistical software R. Package `apc` (Fannon and Nielsen 2020) implements methods based on the canonical parametrization of Kuang, Nielsen, and Nielsen (2008), which however lack flexibility and robustness when compared to nonlinear regression approaches. Package `bamp` (Schmid and Held 2007) offers routines for the analysis of incidence and mortality data based on a Bayesian APC model with a nonlinear prior. R package `Epi` (Carstensen et al. 2021) implements the methods introduced in Carstensen (2007) to analyze disease and mortality rates, including the estimation of separate smooth effects for age, period and cohort. Rosenberg, Check, and Anderson (2014) developed an R-based web tool for the analysis of cancer rates, including different estimates for marginal effect curves.

In contrast to the above software packages, `APCtools` builds on a flexible and robust semiparametric regression approach. The package includes modern visualization techniques and general routines to facilitate the interpretability of the estimated temporal structures and to simplify the workflow of an APC analysis. As is outlined below in further detail, sophisticated functions are available both for descriptive and regression model-based analyses. For the former, we use density (or ridgeline) matrices, classical heatmaps and *hexamaps* (hexagonally binned heatmaps) as innovative visualization techniques building on the concept of Lexis diagrams. Model-based analyses build on the separation of the temporal dimensions based on generalized additive models, where a tensor product interaction surface (usually between age and period) is utilized to represent the third dimension (usually cohort) on its diagonal. Such tensor product surfaces can also be estimated while accounting for further covariates in the regression model.

Descriptive Analysis

In the following, we showcase the main functionalities of the `APCtools` package on the included `travel` dataset, containing data from the German *Reiseanalyse* survey – a repeated cross-sectional study comprising information on German travelers between 1971 and 2018. Focus is on travelers between 14 and 89 years and the distance of each traveler's *main trip* – i.e. each traveler's most important trip in the respective year – and how these distances change over the temporal dimensions.

Several descriptive visualization techniques are implemented that are all based on the classical concept of Lexis diagrams where two temporal dimensions (of age, period, and cohort) are depicted on the x- and y-axis, and the remaining dimension along the diagonals. Additional to heatmaps and *hexamaps* (see below) this includes density matrices (called *ridgeline matrices* in Weigert et al. (2021)) which can be used to flexibly visualize observed distributions along the temporal dimensions. Such visualizations can for example be used to illustrate changes in travel distances. As can be seen in Figure 1 and Figure 3, longer-distance travels are mainly undertaken by young age groups and in more recent years.

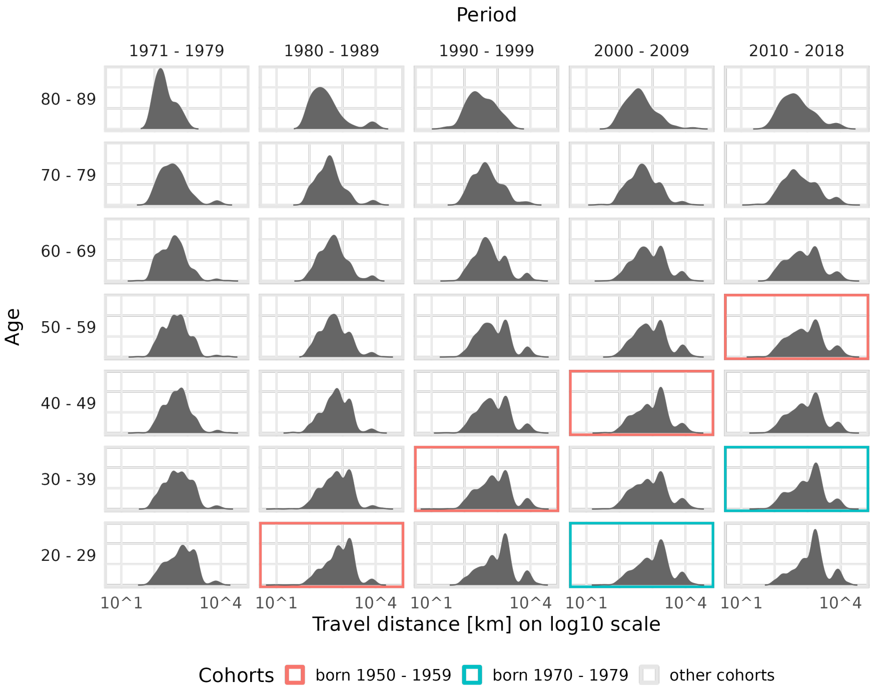

Figure 1: Density matrix of the main trips' travel distance in different age and period groups. Two cohort groups are exemplarily highlighted.

Model-based Analysis

To properly estimate the association of a process with the individual dimensions age, period, and cohort, we utilize the approach introduced by Clements, Armstrong, and Moolgavkar (2005) who circumvent the identification problem by representing the effect of one temporal dimension (e.g. cohort) based on a nonlinear interaction surface between the other two dimensions (age and period). This leads to a generalized additive regression model (GAM, Wood (2017)) of the following form:

$$g(\mu_i) = \beta_0 + f_{ap}(age_i, period_i) + \eta_i, \qquad i = 1, \ldots, n,$$

with observation index i, μ_i the expected value of an exponential family response, link function $g(\cdot)$ and the intercept β_0. The interaction surface is included as a tensor product surface $f_{ap}(age_i, period_i)$, represented by a two-dimensional spline basis. η_i represents an optional linear predictor that contains further covariates. Model estimation can be performed with functions `gam` or `bam` from R package `mgcv` (Wood 2017). As outlined in Weigert et al. (2021) this modeling approach can both be applied to repeated cross-sectional data and panel data.

Based on an estimated GAM, a heatmap of the smooth tensor product surface can be plotted (see Figure 2). Additionally, marginal effects of the individual temporal dimensions can be extracted by averaging over each dimension.

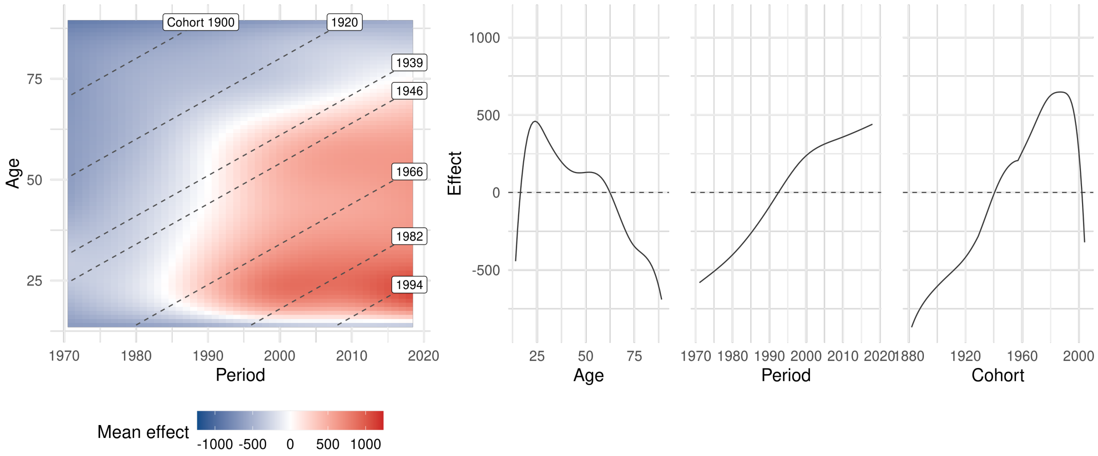

Figure 2: Heatmap of the estimated tensor product surface (left pane) and marginal APC effects based on an additive model with the travel distance as response and no further control variables (right pane).

As an alternative to classical heatmaps the raw observed APC structures or the subsequently estimated model-based tensor product surface can also be visualized using *hexamaps*, i.e. hexagonally binned heatmaps where developments over age, period, and cohort are given equal visual weight by distorting the coordinate system (Jalal and Burke 2020). This resolves the central problem of classical heatmaps where developments over the diagonal dimension are visually underrepresented compared to developments over the dimensions depicted on the x- and y-axis.

`APCtools` further provides partial APC plots, which can be used to visualize interdependencies between the different temporal dimensions (see Weigert et al. (2021) for details). Also, several utility functions are available to plot covariate effects as well as functions to create publication-ready summary tables of the central model results.

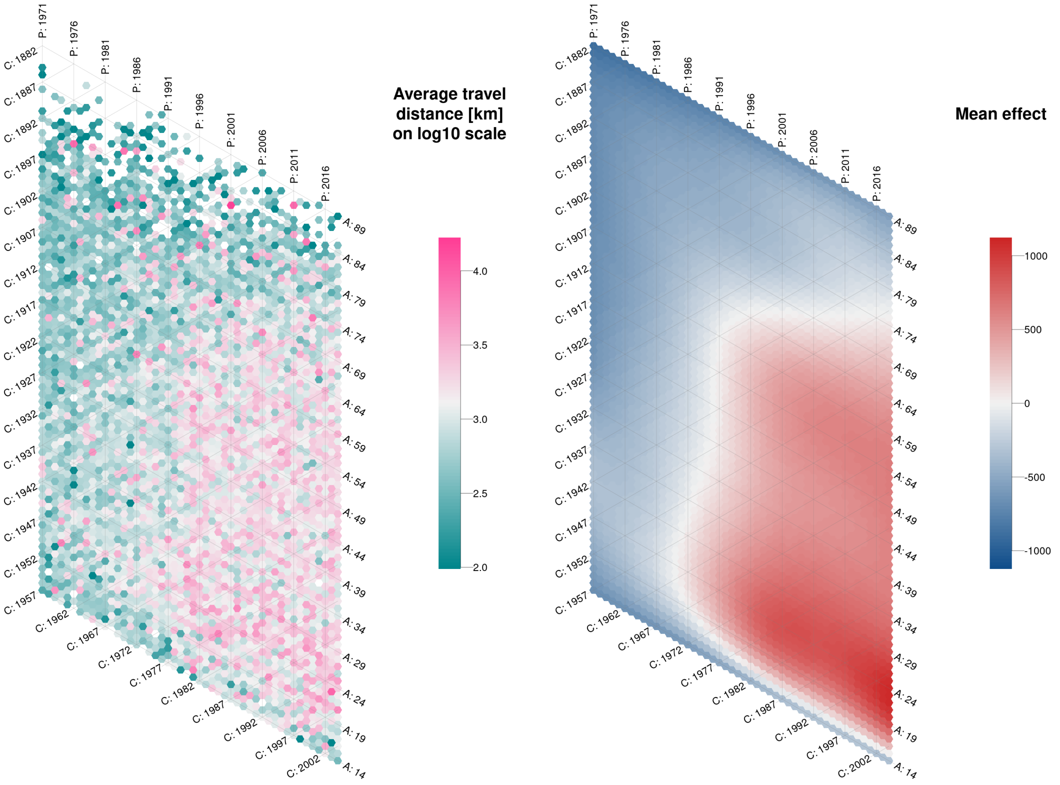

Figure 3: Hexamaps of the observed travel distances (left pane) and the estimated tensor product surface based on an additive model with the travel distance as response and no further control variables (right pane).

Acknowledgments

We thank Helmut Küchenhoff for valuable methodological contributions. This work has been partially funded by the German Research Foundation (DFG) under Grant No. KU 1359/4-1 and by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A.

References

Carstensen, Bendix. 2007. "Age–Period–Cohort Models for the Lexis Diagram." *Statistics in Medicine* 26 (15): 3018–45. https://doi.org/10.1002/sim.2764.

Carstensen, Bendix, Martyn Plummer, Esa Laara, and Michael Hills. 2021. *Epi: A Package for Statistical Analysis in Epidemiology.* https://CRAN.R-project.org/package=Epi.

Clements, Mark S, Bruce K Armstrong, and Suresh H Moolgavkar. 2005. "Lung Cancer Rate Predictions Using Generalized Additive Models." *Biostatistics* 6 (4): 576–89. https://doi.org/10.1093/biostatistics/kxi028.

Fannon, Zoe, and Bent Nielsen. 2020. *Apc: Age-Period-Cohort Analysis.* https://CRAN.R-project.org/package=apc.

Jalal, Hawre, and Donald S Burke. 2020. "Hexamaps for Age-Period-Cohort Data Visualization and Implementation in r." *Epidemiology (Cambridge, Mass.)* 31 (6): e47. https://doi.org/10.1097/EDE.0000000000001236.

Kuang, Di, Bent Nielsen, and Jens P Nielsen. 2008. "Identification of the Age-Period-Cohort Model and the Extended Chain-Ladder Model." *Biometrika* 95 (4): 979–86. https://doi.org/10.1093/biomet/asn026.

Rosenberg, Philip S, David P. Check, and William F. Anderson. 2014. "A Web Tool for Age–Period–Cohort Analysis of Cancer Incidence and Mortality Rates." *Cancer Epidemiology, Biomarkers & Prevention* 23: 2296–2302. https://doi.org/10.1158/1055-9965.EPI-14-0300.

Schmid, Volker J., and Leonhard Held. 2007. "BAMP – Bayesian Age-Period-Cohort Modeling and Prediction." *Journal of Statistical Software* 21. https://doi.org/10.18637/jss.v021.i08.

Weigert, Maximilian, Alexander Bauer, Johanna Gernert, Marion Karl, Asmik Nalmpatian, Helmut Küchenhoff, and Jürgen Schmude. 2021. "Semiparametric APC Analysis of Destination Choice Patterns: Using Generalized Additive Models to Quantify the Impact of Age, Period, and Cohort on Travel Distances." *Tourism Economics.* https://doi.org/10.1177/1354816620987198.

Wood, Simon N. 2017. *Generalized Additive Models: An Introduction with R.* CRC press. https://doi.org/10.1201/9781315370279.

Yang, Yang, and Kenneth C Land. 2013. *Age-Period-Cohort Analysis: New Models, Methods, and Empirical Applications.* Taylor & Francis. https://doi.org/10.1201/b13902.

Contributing Publications

Bauer, A., Scheipl, F., Küchenhoff, H., and Gabriel, A.-A. (2022a). Disentangling the Variation Structure of Seismic Ground Velocities – Registration for Incomplete Non-Gaussian Functional Data. *Manuscript submitted for publication.*

Wrobel, J. and Bauer, A. (2021). registr 2.0: Incomplete Curve Registration for Exponential Family Functional Data. *Journal of Open Source Software*, 6(61): 2964.

Bauer, A., Scheipl, F., Küchenhoff, H., and Gabriel, A.-A. (2018). An introduction to semiparametric function-on-scalar regression. *Statistical Modelling*, 18(3–4): 346–364.

Weigert, M., Bauer, A., Gernert, J., Karl, M., Nalmpatian, A., Küchenhoff, H., and Schmude, J. (2021). Semiparametric APC analysis of destination choice patterns: Using generalized additive models to quantify the impact of age, period, and cohort on travel distances. *Tourism Economics.*

Bauer, A., Weigert, M., and Jalal, H. (2022b). APCtools: Descriptive and Model-based Age-Period-Cohort Analysis. *Manuscript submitted for publication.*

References

Ahmad, O. B., Boschi-Pinto, C., Lopez, A. D., Murray, C. J., Lozano, R., Inoue, M., et al. (2001). Age standardization of rates: a new WHO standard. *Geneva: World Health Organization*, 9(10): 1–14.

Bauer, A. (2016). Auswirkungen der Erdbebenquelldynamik auf den zeitlichen Verlauf der Bodenbewegung. Master's thesis, Ludwig-Maximilians-Universität, Munich, Germany.

Bauer, A. (2017). *FoSIntro: Convenience Functions for Semiparametric Function-on-Scalar Regression*. R package version 1.0.

Bauer, A., Scheipl, F., Küchenhoff, H., and Gabriel, A.-A. (2017). Modeling spatio-temporal earthquake dynamics using generalized functional additive regression. In *International Workshop on Statistical Modelling 2017*.

Bauer, A., Scheipl, F., Küchenhoff, H., and Gabriel, A.-A. (2018). An introduction to semiparametric function-on-scalar regression. *Statistical Modelling*, 18(3-4): 346–364.

Bauer, A., Scheipl, F., Küchenhoff, H., and Gabriel, A.-A. (2022a). Disentangling the Variation Structure of Seismic Ground Velocities - Registration for Incomplete Non-Gaussian Functional Data. *Manuscript submitted for publication.*

Bauer, A., Weigert, M., and Jalal, H. (2022b). APCtools: Descriptive and Model-based Age-Period-Cohort Analysis. *Manuscript submitted for publication.*

Beyaztas, U., Shang, H. L., and Alin, A. (2021). Function-on-Function Partial Quantile Regression. *Journal of Agricultural, Biological and Environmental Statistics*, pages 1–26.

Brockhaus, S., Fuest, A., Mayr, A., and Greven, S. (2018). Signal regression models for location, scale and shape with an application to stock returns. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(3): 665–686.

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(4): 913–926.

Bryner, D. and Srivastava, A. (2021). Shape Analysis of Functional Data with Elastic Partial Matching. *arXiv preprint arXiv:2105.08604*.

Carstensen, B. (2007). Age-period-cohort models for the Lexis diagram. *Statistics in medicine*, 26(15): 3018–3045.

Chakraborty, A. and Panaretos, V. M. (2021). Functional registration and local variations: Identifiability, rank, and tuning. *Bernoulli*, 27(2): 1103–1130.

Chen, C. and Srivastava, A. (2021). SrvfRegNet: Elastic Function Registration Using Deep Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4462–4471.

Cheng, W., Dryden, I. L., Huang, X., et al. (2016). Bayesian Registration of Functions and Curves. *Bayesian Analysis*, 11(2): 447–475.

Claeskens, G., Silverman, B. W., and Slaets, L. (2010). A multiresolution approach to time warping achieved by a Bayesian prior-posterior transfer fitting strategy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(5): 673–694.

Clements, M. S., Armstrong, B. K., and Moolgavkar, S. H. (2005). Lung cancer rate predictions using generalized additive models. *Biostatistics*, 6(4): 576–589.

Diggle, P. J., Heagerty, P., Liang, K.-Y., Zeger, S., et al. (2002). *Analysis of Longitudinal Data*. Oxford university press.

Ebert, A., Mengersen, K., Ruggeri, F., and Wu, P. (2021). Curve Registration of Functional Data for Approximate Bayesian Computation. *Stats*, 4(3): 762–775.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical science*, 11(2): 89–121.

Eilers, P. H. and Marx, B. D. (2021). *Practical smoothing: The joys of P-splines*. Cambridge University Press.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). *Regression: Models, Methods and Applications*. Springer Science & Business Media.

Fannon, Z., Monden, C., Nielsen, B., et al. (2018). Age-period-cohort modelling and covariates, with an application to obesity in england 2001-2014. Technical report, Economics Group, Nuffield College, University of Oxford.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*, volume 76. Springer.

Fienberg, S. E. and Mason, W. M. (1979). Identification and Estimation of Age-Period-Cohort Models in the Analysis of Discrete Archival Data. *Sociological methodology*, 10: 1–67.

Fu, W. J. (2000). Ridge estimator in singulah oesiun with application to age-period-cohort analysis of disease rates. *Communications in statistics-Theory and Methods*, 29(2): 263–278.

FUR Forschungsgemeinschaft Urlaub und Reisen e.V. (2020). Survey of tourist demand in Germany for holiday travel and short breaks. (accessed 13 Feb 2022).

Gertheiss, J., Goldsmith, J., and Staicu, A.-M. (2017). A note on modeling sparse exponential-family functional response curves. *Computational statistics & data analysis*, 105: 46–52.

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(2): 215–236.

Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J., Di, C., Gellar, J., Harezlak, J., McLean, M. W., Swihart, B., Xiao, L., Crainiceanu, C., and Reiss, P. T. (2021). *refund: Regression with Functional Data*. R package version 0.1-24.

Greven, S. and Scheipl, F. (2017). A general framework for functional regression modelling. *Statistical Modelling*, 17(1-2): 1–35.

Guo, X., Kurtek, S., and Bharath, K. (2020). Variograms for spatial functional data with phase variation. *arXiv preprint arXiv:2010.09578*.

Hadjipantelis, P. Z., Aston, J. A., Müller, H.-G., and Evans, J. P. (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.

Hall, P., Müller, H.-G., and Yao, F. (2008). Modelling sparse generalized longitudinal observations with latent Gaussian processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4): 703–723.

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.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2 edition. Springer.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*, volume 43. CRC press.

Holford, T. R. (1983). The Estimation of Age, Period and Cohort Effects for Vital Rates. *Biometrics*, pages 311–324.

Horton, W. Z., Page, G. L., Reese, C. S., Lepley, L. K., and White, M. (2021). Template Priors in Bayesian Curve Registration. *Technometrics*, 63(4): 487–499.

Jalal, H. and Burke, D. S. (2020). Hexamaps for Age-Period-Cohort Data Visualization and Implementation in R. *Epidemiology (Cambridge, Mass.)*, 31(6): e47.

Knapp, G. F. (1868). *Über die Ermittlung der Sterblichkeit aus den Aufzeichnungen der Bevölkerungs-statistik*. Hinrichs.

Kneip, A. and Gasser, T. (1992). Statistical Tools to Analyze Data Representing a Sample of Curves. *The Annals of Statistics*, pages 1266–1305.

Kneip, A. and Ramsay, J. O. (2008). Combining Registration and Fitting for Functional Models. *Journal of the American Statistical Association*, 103(483): 1155–1165.

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

Learned-Miller, E. G. (2005). Data driven image models through continuous joint alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2): 236–250.

Lee, S. and Jung, S. (2016). Combined Analysis of Amplitude and Phase Variations in Functional Data. *arXiv preprint arXiv:1603.01775*.

Ling, N. and Vieu, P. (2018). Nonparametric modelling for functional data: selected survey and tracks for future. *Statistics*, 52(4): 934–949.

Liu, Y., Li, M., and Morris, J. S. (2020). Function-on-scalar quantile regression with application to mass spectrometry proteomics data. *The Annals of Applied Statistics*, 14(2): 521–541.

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.

Maier, E.-M., Stöcker, A., Fitzenberger, B., and Greven, S. (2021). Additive Density-on-Scalar Regression in Bayes Hilbert Spaces with an Application to Gender Economics. *arXiv preprint arXiv:2110.11771*.

Mattar, M. A., Ross, M. G., and Learned-Miller, E. G. (2009). Nonparametric curve alignment. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3457–3460. IEEE.

Matuk, J., Bharath, K., Chkrebtii, O., and Kurtek, S. (2019). Bayesian Framework for Simultaneous Registration and Estimation of Noisy, Sparse and Fragmented Functional Data. *arXiv preprint arXiv:1912.05125*.

McDonnell, E. I., Zipunnikov, V., Schrack, J. A., Goldsmith, J., and Wrobel, J. (2021). Registration of 24-hour accelerometric rest-activity profiles and its application to human chronotypes. *Biological Rhythm Research*, pages 1–21.

Morris, J. S. (2017). Comparison and contrast of two general functional regression modelling frameworks. *Statistical Modelling*, 17(1-2): 59–85.

Müller, M. (2015). *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. Springer.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3): 370–384.

Nunez, E., Lizarraga, A., and Joshi, S. H. (2021). SrvfNet: A Generative Network for Unsupervised Multiple Diffeomorphic Functional Alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4481–4489.

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

Pelties, C., Gabriel, A.-A., and Ampuero, J.-P. (2014). Verification of an ADER-DG method for complex dynamic rupture problems. *Geoscientific Model Development*, 7(3): 847–866.

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer.

Rao, A. R. and Reimherr, M. (2021). Modern Non-Linear Function-on-Function Regression. *arXiv preprint arXiv:2107.14151*.

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.

Rügamer, D., Kolb, C., Fritz, C., Pfisterer, F., Bischl, B., Shen, R., Bukas, C., Thalmeier, D., Baumann, P., Klein, N., et al. (2021). deepregression: a Flexible Neural Network Framework for Semi-Structured Deep Distributional Regression. *arXiv preprint arXiv:2104.02705*.

Sangalli, L. M., Secchi, P., Vantini, S., and Vitelli, V. (2010). k-mean alignment for curve clustering. *Computational Statistics & Data Analysis*, 54(5): 1219–1233.

Sarkar, S. and Panaretos, V. M. (2021). CovNet: Covariance Networks for Functional Data on Multidimensional Domains. *arXiv preprint arXiv:2104.05021*.

Scheipl, F., Gertheiss, J., Greven, S., et al. (2016). Generalized functional additive mixed models. *Electronic Journal of Statistics*, 10(1): 1455–1492.

Schmid, V. J. and Held, L. (2007). Bayesian Age-Period-Cohort Modeling and Prediction – BAMP. *Journal of Statistical Software*, 21(8): 1–15.

Srivastava, A., Wu, W., Kurtek, S., Klassen, E., and Marron, J. S. (2011). Registration of Functional Data Using Fisher-Rao Metric. *arXiv preprint arXiv:1103.3817*.

Stöcker, A., Brockhaus, S., Schaffer, S. A., Bronk, B. v., Opitz, M., and Greven, S. (2021). Boosting functional response models for location, scale and shape with an application to bacterial competition. *Statistical Modelling*, 21(5): 385–404.

Telesca, D. and Inoue, L. Y. T. (2008). Bayesian Hierarchical Curve Registration. *Journal of the American Statistical Association*, 103(481): 328–339.

Tormene, P., Giorgino, T., Quaglini, S., and Stefanelli, M. (2009). Matching incomplete time series with dynamic time warping: an algorithm and an application to post-stroke rehabilitation. *Artificial intelligence in medicine*, 45(1): 11–34.

Tucker, J. D. (2014). *Functional Component Analysis and Regression Using Elastic Methods*. Ph.D. thesis, The Florida State University.

Tucker, J. D. (2020). *fdasrvf: Elastic Functional Data Analysis*. R package version 1.9.4.

Tucker, J. D., Shand, L., and Chowdhary, K. (2021). Multimodal Bayesian registration of noisy functions using Hamiltonian Monte Carlo. *Computational Statistics & Data Analysis*, page 107298.

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

Uphoff, C., Rettenberger, S., Bader, M., Madden, E. H., Ulrich, T., Wollherr, S., and Gabriel, A.-A. (2017). Extreme scale multi-physics simulations of the tsunamigenic 2004 sumatra megathrust earthquake. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16.

Vitelli, V. (2019). A novel framework for joint sparse clustering and alignment of functional data. *arXiv preprint arXiv:1912.00687*.

Volkmann, A. (2021). *multifamm: Multivariate Functional Additive Mixed Models*. R package version 0.1.1.

Volkmann, A., Stöcker, A., Scheipl, F., and Greven, S. (2021). Multivariate functional additive mixed models. *Statistical Modelling*, page 1471082X211056158.

Wagner, H. and Kneip, A. (2019). Nonparametric registration to low-dimensional function spaces. *Computational Statistics & Data Analysis*, 138: 49–63.

Weigert, M., Bauer, A., Gernert, J., Karl, M., Nalmpatian, A., Küchenhoff, H., and Schmude, J. (2021). Semiparametric APC analysis of destination choice patterns: Using generalized additive models to quantify the impact of age, period, and cohort on travel distances. *Tourism Economics*.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*. CRC press.

Wood, S. N., Li, Z., Shaddick, G., and Augustin, N. H. (2017). Generalized Additive Models for Gigadata: Modeling the U.K. Black Smoke Network Daily Data. *Journal of the American Statistical Association*, 112(519): 1199–1210.

Wrobel, J. and Bauer, A. (2021). registr 2.0: Incomplete curve registration for exponential family functional data. *Journal of Open Source Software*, 6(61): 2964.

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

Wu, W. and Srivastava, A. (2014). Analysis of spike train data: Alignment and comparisons using the extended Fisher-Rao metric. *Electronic Journal of Statistics*, 8(2): 1776–1785.

Yang, Y. and Land, K. C. (2013). *Age-Period-Cohort Analysis: New Models, Methods, and Empirical Applications*. Taylor & Francis.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005). Functional Data Analysis for Sparse Longitudinal Data. *Journal of the American statistical association*, 100(470): 577–590.

Zhu, H., Morris, J. S., Wei, F., and Cox, D. D. (2017). Multivariate functional response regression, with application to fluorescence spectroscopy in a cervical pre-cancer study. *Computational statistics & data analysis*, 111: 88–101.

Eidesstattliche Versicherung (Affidavit)

(Siehe Promotionsordnung vom 12. Juli 2011, § 8 Abs. 2 Pkt. 5)

Hiermit erkläre ich an Eides statt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, den 04.03.2022 ________________ Alexander Bauer