Statistical Analysis of Functional Data: Multivariate Responses, Misaligned Data and Local Inference

PhD thesis

Niels Aske Lundtorp Olsen

Department of Mathematical Sciences
University of Copenhagen

This thesis has been submitted to the PhD School of the Faculty of Science,
University of Copenhagen.

Niels Aske Lundtorp Olsen
Email: niels.olsen@math.ku.dk

Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
2100 København Ø
Denmark

Academic advisor:	Assoc. Professor Bo Markussen, *University of Copenhagen.*
Assessment committee:	Professor Helle Sørensen, *University of Copenhagen.* Professor R. Todd Odgen, *Columbia University.* Assoc. Professor Ana-Maria Staicu, *North Carolina State University.*
Submission date:	July 31, 2018.
ISBN:	978-87-7078-915-8.

Abstract

Functional data analysis is characterised by relatively small sample sizes, many observations per curve, and an issue about misaligned data. In this thesis we develop new methods and models for functional data analysis (FDA). The focus is on multivariate responses, misalignment and local inference, three challenging fields within functional data analysis.

In the first paper of the thesis we consider a new model for multivariate, misaligned functional data. We develop low-parametric warp and cross-correlation models, and we apply the model to three different data sets. We also use of the last data set in a classification study, where we compare our model to a number of state-of-the-art methods.

The second paper of the thesis is about the same topic, but with a very different approach. By a clever parametrisation using the Cholesky decomposition, we develop a model framework that potentially allows for very fast computations.

The third paper of the thesis is about local inference for functional data. We develop a functional analogue to the *Benjamini-Hochberg* method as a way to deal with the multiple comparisons problem. The paper contains theoretical results about control of false discovery rates, two simulation studies and an application to satellite measurements of Earth temperatures.

The last paper of the thesis contains a statistical study of conidial discharge, where we extend the model from the first article in the context of generalised linear models. In the application we study the intensity of conidial discharge as function of time, for mycelia stored at three different temperatures.

Resumé

Funktional dataanlyse er karakteriseret ved relativt små stikprøvestørrelser, mange observationer for hver kurve og en problemstilling omkring misalignede data. I denne afhandling udvikles nye metoder og modeller for funktionel dataanalyse (FDA). Fokus er på flerdimensionale responser, misalignment og lokal inferens, tre udfordrende områder indenfor funktionel dataanalyse.

I afhandlingens første artikel betragter vi en ny model for multivariate, misalignede funktionelle data. Vi udvikler lavparametriske warp- og krydskorrelationsmodeller, og modellen anvendes på tre forskellige datasæt. Det sidste af datasættene benyttes til et klassifikationsstudie, hvor vi sammenligner vores model med en række state-of-the-art metoder.

Afhandlingens anden artikel omhandler samme emne, men med en meget anderledes tilgang. Ved på smart vis at parametrisere med cholesky-dekompostionen udvikler vi en modelramme der potentielt tillader meget hurtige beregninger.

Afhandlingens tredje artikel omhandler lokal inferens for funktionelle data. Vi udvikler en funktionel analog til *Benjamini-Hochberg*-metoden til at håndtere problemstillingen omkring multiple tests. Artiklen indeholder teoretiske resultater om kontrol af False Discovery Rates, to simulationsstudier og en anvendelse på satellitmålinger af globale temperaturer.

Afhandlingens sidste artikel omhandler et statistisk studie af afskydning af konidiesporer, hvor vi videreudvikler en af de førnævnte modeller i konteksten af generaliserede lineære modeller. I anvendelsen undersøger vi hvordan sporeafskydningsintensiteten udvikler sig som funktion over tid, for mycelier opbevaret ved tre forskellige temperaturer.

Preface

This thesis has been submitted as a formal requirement for obtaining the PhD degree at University of Copenhagen. The work has been carried out from August 2015 to July 2018, with the research mostly being done at Department of Mathemtical Sciences (MATH).

My first practical encounter with the topic of this thesis, namely functional data analysis, was a course taught by Helle Sørensen and Anders Tolver at University of Copenhagen in 2014, where I was introduced to standard tools for functional data analysis such as smoothing, PCA, and registration of functional data. I wrote my master thesis about multivariate longitudinal models, and from that I had a good starting point for a PhD about multivariate functional data analysis. Since then, the project has expanded in various directions to cover multivariate responses, discrete responses, misaligned data as well as local inference. As a part of my PhD, I was at long-term stay at Politecnico di Milano in Italy. This was a very fruitful and inspiring stay, and the resulting collaboration led to the third paper of this thesis.

The thesis starts with an introduction and discussion of certain aspects of functional data analysis followed by some supplementary material to the papers. The first two papers are work carried out in MATH, on the initiatives of Lars Lau Raket and Bo Markussen. The fourth paper in the thesis is the result of collaboration initiated by teaching in the course *Applied Statistics*. One of the students, Pascal Herren, had done experiments on spore discharge of fungi, and the data was used in the mandatory course project. Later we met and initiated a collaboration with both a statistical and a biological scope.

First of all, I wish to thank Bo Markussen, my supervisor, for guidance, help and patience, and for always being available for a discussion. My next thanks goes to Lars Lau Raket, who co-supervised me for a period, also for guidance, help and patience. I would also like to thank Anders Tolver for many discussions about functional data, and a special thanks goes Simone Vantini for arranging my stay to Milan.

A warm thanks goes to my office mates, both in Copenhagen and in Milan, and to my fellow colleagues at MATH for companionship and creating a great atmosphere. Finally I want to thank my family for all your support over the years.

Contents

1 Introduction

1.1 Introduction to functional data

Although functional data analysis today is a well-established field within statistics, few authors on functional data analysis (FDA) textbooks present any rigorous definition of (what actually is) functional data. Popular textbooks such as Horváth & Kokoszka (2012) and Ramsay & Silverman (2005) begin by introducing examples of functional data and avoid rigorous definitions.

This is an approach to which the author agrees; however a very general definition of functional data is statistical data associated with smooth functions $M \to M'$, where M and M' are manifolds, typically subsets of $\mathbb{R}^n$ and $\mathbb{R}^{n'}$, respectively. For the majority of applications, M is a sub-interval of $\mathbb{R}$, in that case, the smooth functions are smooth curves.

These smooth curves are not observed themselves, but are observed at discrete time points with noise and/or shifted in domain (misalignment), and the noise is often correlated across the domain M; this is referred to as *serial correlation*. Furthermore, one typically have few samples but many observations, and the locations and numbers of observations may be different from sample to sample. It might even be that the observed data are actually discrete but randomly generated from an underlying smooth process of interest.

Some of these features (although rarely all of them) are usually present in functional data, and the methodology used to analyse and perform inference in this kind of data is, in the author's opinion, what constitutes functional data analysis. It should be noted that there are no clear distinctions between functional data analysis and related fields of statistics such as spatial, longitudinal and multivariate data analysis, and we will not elaborate further on this in this thesis. Popular references on multivariate and longitudinal data analysis are Koch (2013) and Diggle et al. (2002), respectively.

The term 'functional data analysis' was coined by Jim Ramsay in 1982 (Ramsay 1982) and since then, it has become a large and diverse field within statistics. The aim of this thesis is to extend and contribute to some of the methodology of functional data, with a particular focus on misalignment and multivariate functional data.

This section will look into some typical features of functional data analysis with a focus on the content of the papers along with the author's opinion on some of these issues.

Two examples of functional data sets A popular example of functional data is the *Canadian Weather data set* (Ramsay & Silverman 2005, Ramsay et al. 2018), which has a

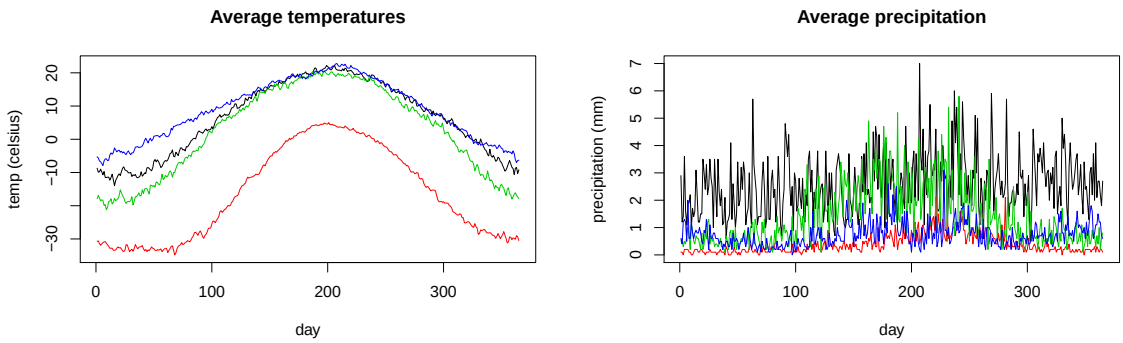


FIGURE 1.1: Temperature and precipitation curves from four Canadian weather stations (BLACK: Montreal, RED: Resolute, GREEN: Winnipeg, BLUE: Kamloops)

simple and interpretable setting. The data set consists of daily temperature and precipitation averages of 35 Canadian weather stations across the country, see Figure 1.1 for an example. There might not be an obvious research question to be asked from the data, but understanding the variation across weather stations, and comparisons across regions and between temperatures and precipitation would be of interest. One would need to deal with noise in data in a clever ways; in particular, the precipitation is very noisy. Identifying and describing the variation is not a straightforward task either. Tools such as principal component analysis would be recommended, and with a sound application of functional data methodology, one gets an understanding of how average temperatures and precipitation varies across Canada.

Another example of functional data is from a study by Thomsen et al. (2010) on data-driven detection of *horse lameness*. Data consist of a large number of acceleration profiles of trotting horses. Figure 1.2 shows acceleration profiles in three directions of a trotting horse with an artificially introduced (non-permanent) lameness on its left foreleg. This is multivariate functional data with an obvious periodic structure of data, which can and should be exploited in a data analysis. Alignment is an obvious issue in these data; timing can vary substantially between different repetitions/horses and also varies between cycles of the individual repetitions. By comparing acceleration profiles for different horses under various settings, the study aims at classifying lameness of the horses, both in terms of severity and where the lameness is located (ie. which leg is lame).

The nature of functional data While we tend to think of functional data as continuous objects from a suitable function space, functional data are always observed as discrete data, due to obvious constraints on measurement and data storage devices. The data is thus always multivariate in nature, but the special structure of functional data and the high ratio of "data dimension" to number of samples (in a more general setting known as the '*curse of dimensionality*') means that standard tools for multivariate analysis are inadequate for functional data. Furthermore, the number of data points per curve typically varies between curves and one would often have some missing data points; a simple consequence of this is

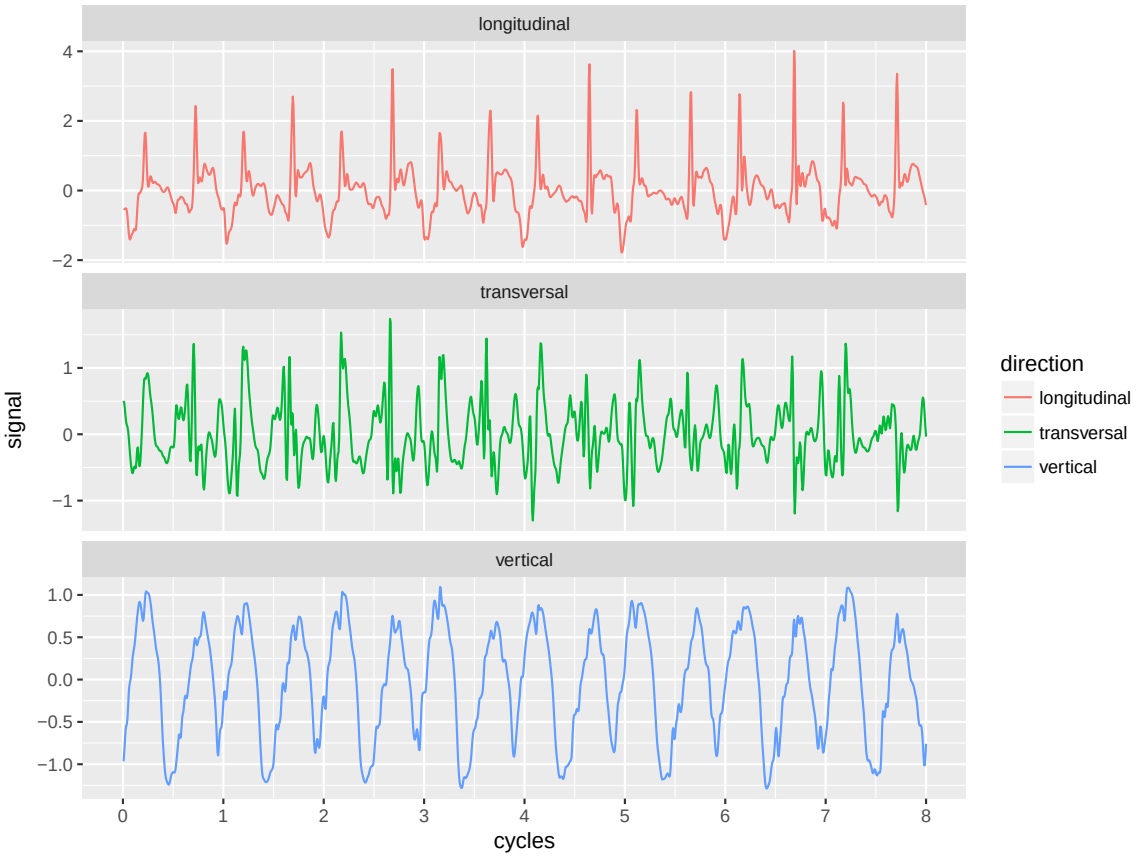


FIGURE 1.2: acceleration profiles of a trotting horse with an artificially introduced (non-permanent) lameness on its left foreleg.

that models for functional data should be able to handle different data sizes.

Functional data sets and the scopes of corresponding statistical analyses are diverse as the examples of the previous paragraph illustrate, but some of the topics that often arise in functional data analysis are stated below.

1. *Smoothing of data*: Going from discrete observations to continuous curves is always part of functional data analysis; it can be done on the level of individual curves or population means.

2. *Alignment*: Observed data are often misaligned in the sense that they have the same shape but individual curves have been deformed in one dimension (e.g. time).

3. *Regression*: There are likely to be some covariates present in data, how do we estimate and quantify the effects of these in a good way? Here one should take into account that pre-processing of data may play a large role.

4. *Inference*: Is there an effect and where is the effect?: Given a set of covariates C for some parameter space $\mathcal{C}$, we would often like to test a hypothesis $C \in \mathcal{U}$ for a given subset $\mathcal{U} \subset \mathcal{C}$. Secondly we might want to test the hypothesis $C(t) \in \mathcal{U}(t)$ across the domain of the functional data allowing us to select "areas of effect", also known as *domain selection*.

This ordering loosely reflects the frequency in functional data, and also the order which these steps often are carried out.

There are various other topics often treated in articles and monographs on functional data analysis such as classification, clustering, prediction, depth analysis. Paper I contains a classification study, but apart from that, none of these topics will be considered in this thesis.

Multivariate functional data versus univariate functional data The majority of literature on functional data deals with univariate functional data, that is functional data $[a, b] \to \mathbb{R}$, whereas multivariate functional data are functions $[a, b] \to \mathbb{R}^k$ for $k \geq 2$.

Often multivariate functional methods are considered trivial extensions of univariate methodology, yet this is rarely done in practice. Two examples of studies, where multivariate functional data are analysed as were they univariate functional data, are Sørensen et al. (2012) and Raket et al. (2016).

The multivariate response adds an extra dimension af variation in comparison to univariate functional data, having three layers of variability: across dimension, across time and across subjects/repetitions. This extra layer of variability adds some challenge compared to univariate responses:

- *Alignment*: Alignment of multivariate data must compromise between alignment of individual coordinates. See Section 1.2 for more details.

- *Correlation*: Repeated measurements of multivariate data can exhibit two forms of correlation: *serial correlation* and *cross-correlation*. Serial correlation is correlation over time and also present in univariate functional data. Cross-correlation on the other hand is correlation *between* coordinates $(x_1(t), \ldots, x_k(t))$. Much power and information may be gained by incorporating cross-correlation into the data analysis; a statistical analysis that does not use any inner products between coordinates is making an implicit assumption of no cross-correlation. Papers I and II introduce new models for modelling cross-correlation that varies over time.

- *Visualization and implementation* This is a minor issue, but a relevant one. Plotting univariate functional data is easy: time on one axis and response on the other axis. Visualisation is a basic descriptive tool and allows us to detect and understand variation in data.

  Furthermore, not all software implementations of functional data can handle multivariate responses which adds a practical challenge.

In this thesis we will look into some new dedicated methods for multivariate functional data in Papers I and II.

1.1.1 The basi(c)s: a very brief introduction to modelling functional data

A model for functional data If we don't consider misalignment in data, covariates and complicated stuff, the "basic" model for functional data is

$$y_i(t) = \theta(t) + \epsilon_i(t), \quad t \in [a, b], \quad i = 1, \ldots N \tag{1.1}$$

where $\theta : [a, b] \to \mathbb{R}^k$ is the *mean function*, and ϵ is residual/amplitude variation.

The observed data consist of $m = m_1 + \cdots + m_N$ pairs of discrete observations $\{(t_{1j}, y_{1j})\}_{j=1}^{m_1}, \ldots, \{(t_{Nj}, y_{Nj})\}_{N=1}^{m_N}$ and the primary aim is to infer θ from data.

As the set of smooth functions on $[a, b]$ is an infinite-dimensional vector space, one cannot hope for a complete identification of θ, and one has to rely on a finite representation for θ. For that we use a pre-specified set of *basis functions* $\{\phi_k\}_{k=1}^M$, $\phi_k : [a, b] \to \mathbb{R}$, and define $\theta = \sum \phi_k c_k$, where $\{c_k\}_{k=1}^M$ are coefficients for the basis functions. The basis functions should be chosen such that $span\{\phi_1, \ldots, \phi_K\}$ is 'flexible enough to capture the characteristics of θ'. Basis functions are usually chosen for their mathematical properties; b-spline bases and Fourier bases are typical choices. Ramsay & Silverman (2005) lists a number of other basis systems, including the popular wavelet transforms (Daubechies 1992).

Fitting curves to data The estimation problem now consists of finding the spline coefficients c. If $N = 1$ or one models one set of coefficients per curve ie. $c = \{c_n\}$, we speak of the *smoothing problem*. The coefficients of c may be found using *least squares*:

$$\hat{c} = \arg\min_c \sum_{i=1}^{N} (\boldsymbol{y}_i - \phi_i c)^\top \Sigma^{-1} (\boldsymbol{y}_i - \phi_i c) \tag{1.2}$$

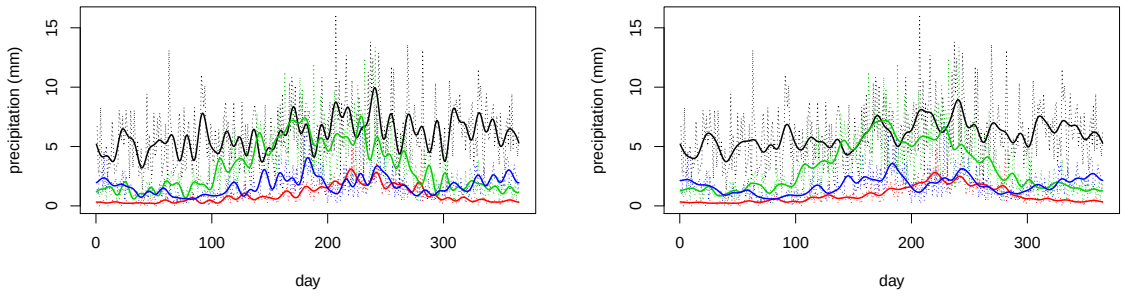


FIGURE 1.3: Canadian weather data: Left: no penalisation. Right: penalisation

where Σ is the residual covariance; $\Sigma = \sigma^2 I$ corresponds to iid. noise; in which (1.2) reduces to ordinary least squares.

Penalisation Often, in order to control the roughness of the estimated θ, a penalisation term is introduced. This may be done in form of a differential operator L s.t. $||L\theta||^2$ is penalised, where $||L\theta||$ is the L^2-norm of $L\theta$. e.g. $L\theta = \theta''$. If we introduce a roughness penalty, the smoothing problem becomes:

$$\hat{c} = \arg\min_c \sum_{i=1}^{N} (\boldsymbol{y}_i - \phi_i c)^\top \Sigma^{-1} (\boldsymbol{y}_i - \phi_i c) + \lambda L(\theta) \tag{1.3}$$

where λ is a parameter controlling the amount of regularisation. Without a penalisation on the roughness, there is a risk of overfitting data when using a large number of basis functions, and the higher-order derivatives of θ might behave wildly. As θ is a linear function of c, it generally holds true that $||L\theta||^2 = c^\top R c$ for some positive semi-definite matrix R. In that case, we can rewrite (1.3):

$$\hat{c} = \arg\min_c \sum_{i=1}^{N} (\boldsymbol{y}_i - \phi_i c)^\top \Sigma^{-1} (\boldsymbol{y}_i - \phi_i c) + \lambda c^\top R c \tag{1.4}$$

which has the closed-form solution:

$$\hat{c} = (\sum_i \phi_i^\top \Sigma^{-1} \phi_i + \lambda R)^{-1} \sum_i \phi_i^\top \Sigma^{-1} \boldsymbol{y}_i \tag{1.5}$$

The optimal value of λ is usually found by *generalised cross-validation* (GCV) or similar methods.

As an example of the effect of using a roughness penalty, we will use the precipitation curves of Fig 1.

We use a Fourier basis with 65 basis functions per curve, and following Ramsay & Silverman (2005) we use a penalty of $L\theta = \theta' - \theta'''$, such that L maps the first harmonics to zero. Fourier

bases have the property that all basis functions are orthogonal to each other. This implies that R is diagonal. Choosing the roughness penalty using the GCV criterion, we get $\lambda = 0.0193$. The smoothed curves, with and without penalisation are displayed in Figure 1.3. We see a clear effect of penalisation – the penalised curves are much less wiggly and "more reasonable", in particular the green and black curves(Kamloops and Montreal, respectively). One should not blindly trust the use of penalisation when smoothing functional data. The results when using GCV depends on number of bases, choice of penalisation operator, covariance structure and other things.

Smoothing as pre-processing Many functional data analysts would say that the smoothing procedure described in this section constitutes the *pre-processing* of the raw data. With the discrete observations $\{(t_{1j}, y_{1j})\}_{j=1}^{m_1}$ replaced by the functions/curves $\theta_1, \dots \theta_N$, we are now ready to continue the analysis and look into issues such as alignment, principal components etc. This approach is used in e.g. Ramsay & Silverman (2005) and Srivastava & Klassen (2016).

The author is sceptical about such approaches. While it is very desirable to convert the original data into functions, it is important that the pre-processing of data at most has a negligible influence on the subsequent data analysis. Papers on functional data analysis tend rarely to reflect on these issues, despite that in many cases it is not evident if making (reasonable) changes to the pre-processing could have made a difference in the subsequent data analysis.

Other smoothing techniques There are various alternatives to smoothing than using basis functions, such as linear interpolation and kernel methods, none of which will be considered here. It appears that using basis functions for pre-processing is by far the most common approach in FDA literature. A likely reason for this is the smoothness properties (in terms of derivatives) and data reduction characteristics resulting from this approach.

Smoothing and derivatives Derivatives of functions often play an important role in FDA; derivatives are associated with change, and the quantity of change (e.g. growth) is important in many studies (Ramsay & Silverman 2005).

A fundamental challenge is that derivatives are not observed, but must somehow be processed from data. This may be done by models that mimics derivatives in smart ways, or by appropriate smoothing of the observed data. Particular care should be taken when using derivatives from pre-processed data. Derivatives are local features of the smoothed data, so they will be even more sensitive to pre-processing of data than the smoothed signals themselves. Despite that many studies and examples within FDA use derivatives of pre-processed curves, the author is not aware of any robustness studies on this subject.

Derivatives indirectly or directly play a role in the papers; this be in the form of hypotheses (Papers III and IV), warping and estimation (Papers I, II and IV) or covariance structures (Papers I and IV).

1.1.2 Inference for functional data

Functional data analysis is sometimes done in context of scientific questions that relate to statistical hypotheses. An example of this is the *two population test*: Let $x_1, \ldots, y_n$ and $y_1, \ldots, y_m$ be independent realisations of random variables X and Y, respectively, can we based on data reasonably believe that $X \stackrel{D}{=} Y$, or is there evidence in data to conclude that $X \stackrel{D}{\neq} Y$? Here $\stackrel{D}{=}$ denotes equality in distribution. When realisations of X and Y are curves, this question becomes hard to answer.

Inference is less often encountered in functional data analysis than in other fields of statistics. Functional data analysis more often deals with "summary statistics" such as identifying mean curves/trajectories and principal modes of variation than inferential problems, and for reasons stated below inference in functional data carries several issues.

In principle, hypothesis testing in functional data analysis is simple: Set up a correct model for data, estimate parameters under null and alternative hypotheses, calculate *likelihood-ratio test* or some other suitable test statistic and compare this to the distribution under the null hypothesis. If the distribution is infeasible, use asymptotical theory for approximation. This is hypothesis testing as it would likely be presented in a book on theoretical statistics. Unfortunately, functional data rarely follows that theory.

Using the two-population framework as reference/example, some issues regarding inference for functional data are stated below. Assume that instances of X and Y are functions on the domain M:

1. *Small sample sizes*: Although each curve (should) consist of many observations, sample sizes in functional data are generally small. Many tests used in practice in statistical inference are based on asymptotical properties of the tests, which one cannot rely on for small sample sizes.

2. *Parametric and semi-parametric tests*: Many tests rely on Gaussianity, and though Gaussian models are convenient for functional data, there is no evidence that this assumption is generally true. For modelling purposes, non-Gaussianity is not a serious issue, but for statistical testing, it is important.

3. *Identifying differences*: Suppose that we are able to conclude $X \stackrel{D}{\neq} Y$. This does not answer how the distributions of X and Y differ, nor which differences are relevant, something which depends on the application. One possible direction is to explore for which $t \in M$ that $X_t \stackrel{D}{\neq} Y_t$. This is what we call *local inference* or *domain selection*.

Some inferential questions are considered in Paper IV. The null hypotheses in that paper are highly non-linear and formulated in terms of latent variables, which make estimation under null hypotheses unfeasible in practice.

Local inference Local inference is a small but interesting topic within functional data analysis. The main issue in local inference is the *multiple comparisons problem* since local inference involves a continuum of hypotheses. Local inference is the main topic of Paper III. For recent reviews of local inference we refer to Paper III and Abramowicz et al. (2018).

1.2 Functional data with temporal variation

Overview Although time is inherent in functional data, temporal variation between objects is a curious thing, often considered a "nuisance" to be filtered away. Temporal variation is also known as *phase variation*, and functional data, where temporal variation is present, is known as *misaligned* or *unregistrered* (functional) data. The concept is variously known as *warping*, *alignment* or *registration*. This section is intended as a reflection on temporal variation including novel ideas presented in Section 1.2.3. For recent reviews of this subject, we refer to Wang et al. (2016) and Marron et al. (2015).

Why study temporal variation? When modelling functional or longitudinal data, we tend to think of a set of idealised, but unobserved, systems $\mathbf{X} = \{x_i\}$ where x_i is a function $[a, b] \to \mathbb{R}^k$. Depending on the setup, it could be population means, covariate effects etc.

Generally, t is physical time in this setting, and the aim of the analysis is to understand and characterise $\mathbf{X}$, and possibly relate it to relevant scientific questions. Although laws of nature dictate that the effects of time are the same for all entities, it is rarely the case that biological systems (and many other types of systems) have the exact same temporal evolution; ie. if an experiment is repeated twice under the same conditions, it is unlikely that timings observed in the two outcomes are identical.

As an example of this, consider the following setup (Figure 1.4) from (Grimme 2014, chapter 3.6), where a person is to lift a cylinder across another cylinder ten times in total (for simplicity only the z-coordinate is shown).

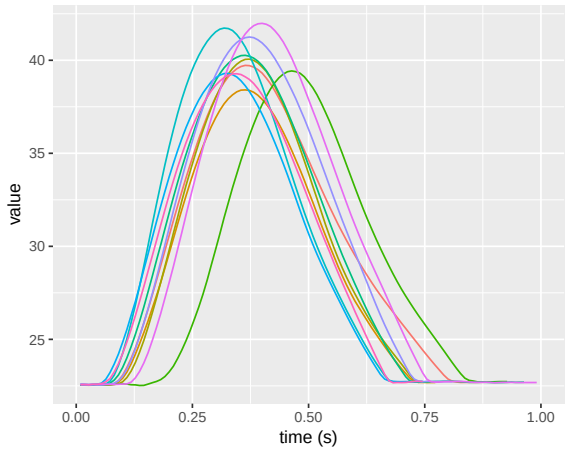


FIGURE 1.4: Ten repetitions of a cylinder experiment

There are some clear constraints in this experiment: the start and end points are well-defined, and so is the obstacle in the middle. But it is implausible to assume that timing within the movement is the same across repetitions; as is also evident in the figure.

So somehow we must incorporate this temporal variation into our analysis. There are many approaches to this, some better than others. A key concept is *warping functions*, that model the temporal deviation from the idealised system; warping functions map the idealised time into observed time. Some authors prefer the opposite formulation: warping functions map observed time into idealised time or system time. These formulations are equally good, but they represent (slightly) different ways of thinking: is it the underlying, unobserved signal, that should be warped, or is it the observed trajectories that should be back-transformed into aligned trajectories?

An illustration of the two formulations is shown below:

$$\boxed{\theta} \xrightarrow{\text{warping}} \theta \circ v \longrightarrow \boxed{Y} \qquad \boxed{\theta} \longrightarrow Y \circ v \xrightarrow{\text{warping}} \boxed{Y} \tag{1.6}$$
$$\uparrow \epsilon \qquad\qquad\qquad\qquad \uparrow \epsilon$$

Here Y denotes the observed curve, θ population mean, ϵ residual variance, and v is a warping function.

1.2.1 Some approaches to misaligned functional data

Basically, I would say that there are three approaches of handling misaligned data:

1. *registration of the observed curves*: Using some suitably chosen method for estimating individual warping functions v_i, the observed data (possibly smoothed versions of this) are mapped into the registered data $\tilde{y}_i = v_i^{-1} \circ y$, which are used in the subsequent analysis. If there were no variation in data besides misalignment, the registered functions would all be identical.

2. *registration as part of the modelling*: One constructs a model on the form $y_i(t) = \theta(v_i(t)) + z_i(t)$, where z_i is amplitude variation and (typically) assumed to be uncorrelated from the warp v_i. Maximum likelihood or similar is then used to predict/estimate v_i and estimate parameters for the model. Such models are highly non-linear and rarely have closed-form estimators, so estimation methods mostly alternate between estimating/predicting warps and estimating parameters describing z_i.

3. *landmark registration*: Sometimes, functional data feature "landmarks"; certain peaks or other characteristic features (typically of scientific interest) that are present throughout the data. These landmarks are identified by the data analyst, and the data are mapped into registered data $\tilde{y}_i$ such that landmarks have the same temporal position for all registered curves.

There are many variations on these approaches and no clear distinction between the methods. One may consider landmark registration as a methodology in the *registration of the observed curves* family, but generally the warping functions are not of interest in landmark registration; only the warped curves are.

Landmark registration is a simple and very intuitive approach: certain features or "landmarks" in data are identified, typically by the researcher, and aligned such that the landmarks are placed at the same location in time. These landmarks often have importance in the experimental context of the data; typical landmarks are local maxima or minima of certain magnitude after initial smoothing of data.

As a simple example, consider the following small example (Figure 1.5), which shows the vertical position of the left foot of a walking person:

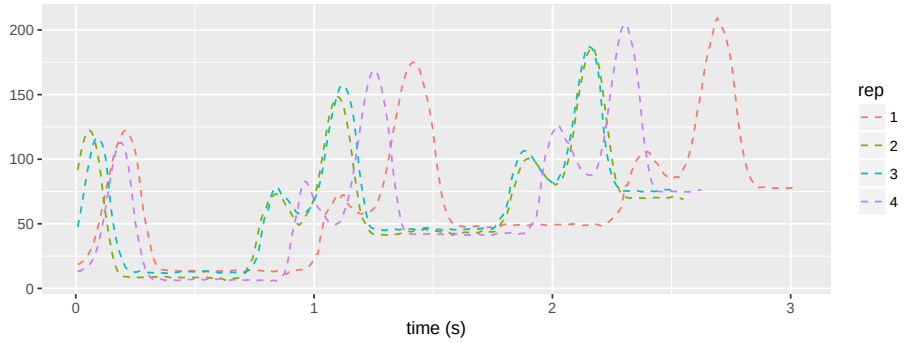


FIGURE 1.5: Four repetitions of a walking sequence

There are four repetitions in total and approximately three gait cycles per repetition. We can clearly identify some landmarks: a small peak shortly after the foot raises from the surface, and a large peak before it lands again. These peaks differ in timing, but since they represent the same underlying movement, one would like these landmarks to be aligned. The times for take-off and landing can also be used as landmarks.

The author believes that landmark registration is a strong tool, at least for a preliminary analysis, and that it to a reasonable extent can used as benchmark/reference for other methods: a method for registration of functional data is adequate if it aligns data such that landmarks are given the same location. An obvious drawback of landmark registration is the fact that data must have identifiable landmarks that are present in all curves for a registration to work.

A popular family of methods, belonging to the first class of approach listed above and widely considered among state-of-the-art, uses equivalence classes of functional representations: Let f_1 and f_2 be appropriate functions with common domain D, and let W be a family of warping functions $D \to D$. We say that f_1 and f_2 are equivalent if there exists $h \in W$, such that $f_2 = f_1 \circ h$ (Marron et al. 2015). The key idea is then to have a metric that is invariant under simultaneous warpings, $d(f, g) = d(f \circ h, g \circ h)$. Two popular methods are the square-root velocity transform (Srivastava & Klassen 2016), which has been used in many applications,

and the correlation criterion (Sangalli et al. 2009). For a discussion of approaches, we refer to Vantini (2012) and Marron et al. (2015).

The author prefers the second approach (registration as part of the modelling) combined with random warping functions. One of the main philosophical reasonings for treating warps as random variables is that individual warps are associated with individual curves, ie. the warp for curve i is associated with curve i and only that curve, and if we were to repeat the experiment again, the warps would presumably be different. It thus makes good sense to treat warp functions as *random* and *latent* quantities with an inherent uncertainty.

Model-based approaches for registration are faithful to data: a mean curve is a mean curve, and although it may be hard to estimate, there is no need for defining it in terms of some complicated average of equivalence classes. Furthermore, there is no need for smoothing of data: the noise can be accounted for using a suitable model for this.

Most models for phase variation assume that phase and amplitude variation are uncorrelated. This includes models such as landmark registration that do not specify any distance measure related to warping.

Identifiability Another fundamental issue in misaligned FDA is that of *identifiability* – can warping curves be identified? If misalignment were the only kind of variation in data, the answer should be yes. But with amplitude variation also present in data, this is not so easy; one has to identify which parts of the variation is due to warp variation and which is due to amplitude variation. If one has a large population of curves and a correctly specified model, this should be doable. But sample sizes in FDA are usually small, and checking model assumptions may be difficult, so identifying/quantifying amplitude and phase variation is generally hard.

Returning to our previous example of cylinder movement, we have now included the x-coordinate in Figure 1.6. For the z-coordinate, it is clear that there is some amplitude variation in data – if the curves were placed on top of each other, there would still be quite some variation left. If we instead focus on the x-coordinate, we see that the curves can approximately be warped into each other, but this is not a proof that phase variation explains almost all of the variation. Indeed, if we combine all three coordinates (y-coordinate not shown here) as done in Paper I, we reach the conclusion that much variation is also due to amplitude variation.

1.2.2 Modelling and prediction of warps

If the temporal variation of functional data could easily be inferred from data, there would be no reason for the many papers on this subject. This is however not the case, and quite some effort has to be done in terms of modelling and prediction of warping functions.

No matter which approach is applied, some effort has to be done in terms of warp prediction and there is no default method for penalizing/estimating warps. One has to specify a family of possible warps and an optimization criterion for warp prediction. The constraints, flexibility and regularization put on warping functions will naturally affect the results.

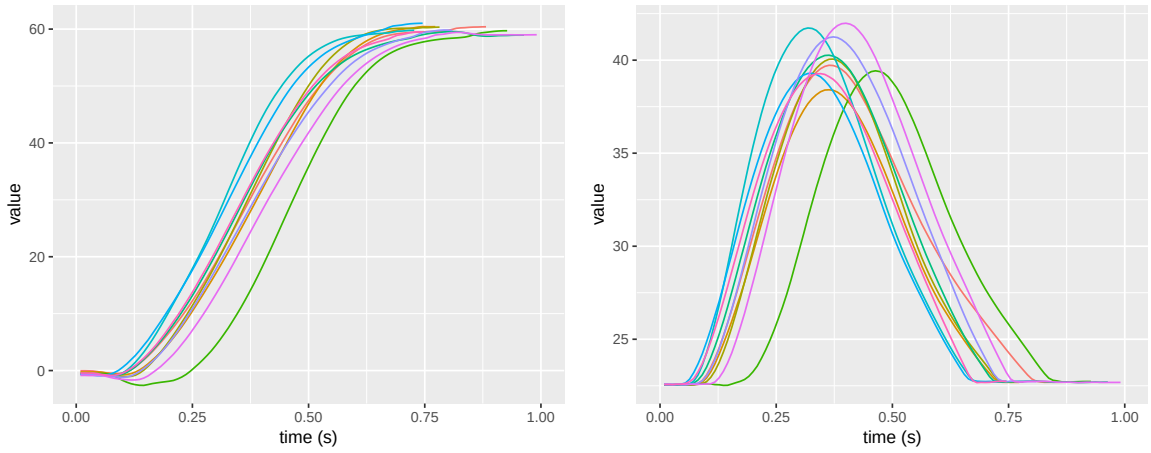


FIGURE 1.6: Ten repetitions of a cylinder experiment. Left panel: x-coordinate. Right panel: z-coordinate.

Most models of warping functions define a class of warping functions in terms of a suitable vector space V, where the modeling takes place. Along with this, there is an injective mapping $\phi : V \to L^2(D)$, where D is the domain of the warping functions, identifying $w \in V$ with a warping function. An example is *shift registration*, the most basic non-trivial alignment procedure: here $V = \mathbb{R}$ and $(\phi(w))(t) = t - w$.

Many warping methods follow the generic framework defined below:

Definition 1 (Generic framework for modelling and prediction of warping functions)**.** Let y be a functional observation and $\theta : D \to \mathbb{R}^k$ a target curve. Let V be real vector space, and ϕ an injective mapping $\phi : V \to L^2(D)$.

If we warp the observation y (compare the left panel in Equation (1.6)), warp prediction amounts to minimising

$$L_{warp}(w) + L_{amp}(y \circ \phi(w) - \theta). \tag{1.7}$$

If we instead warp the target curve (compare the right panel in Equation (1.6)), the minimisation criterion is

$$L_{warp}(w) + L_{amp}(y - \theta \circ \phi(w)) \tag{1.8}$$

Here L_{warp} and L_{amp} are parameter-dependent loss functions, and $(y-\theta\circ\phi(w))$ or $(y\circ\phi(w)-\theta)$, respectively, is the amplitude signal. Parameters of L_{warp} and L_{amp} needs to estimated as part of the process, and (1.7) and (1.8) can be viewed as posterior likelihoods for the full model.

Often the loss functions will be quadratic forms. This implicity amounts to some Gaussianity assumption on w and the amplitude signal.

Most models assume that phase and amplitude variation are uncorrelated in sense that L_{warp} and L_{amp} are additive as in Definition 1. This assumption is convenient and plausible in

many applications. One possible explanation that authors rarely incorporate phase-amplitude correlation is simply the small sample sizes of FDA: There is too little data to firmly establish a correlation, taking into account the fact that predicting warping functions is not a trivial task. A notable paper on correlated phase and amplitude variation is Hadjipantelis et al. (2015); they argue that phase and amplitude is in fact correlated; their study contains 50000 profiles.

Besides that the modeling of warping could be made easier and natural when defined on a vector space, vector spaces allow good representations and implementations for warping functions.

Literature has many examples on registration methods following the generic framework presented above, and using vector spaces is a good tool for representations and software implementations of warping functions. As the transformation ϕ is usually non-linear (and has to be if the warping functions are required to be monotonous, see below), optimization is inherently non-linear. Raket et al. (2014) introduced a model where $V = \mathbb{R}^k$ for k small, and $w \in V$ is modelled as a latent Gaussian random variable. This model gives a large and flexible class of warping functions and has been used with success in a number of applications. We use this model in Papers I and IV.

Properties of warping functions There are natural properties of warping functions, although some of these are too restrictive for some applications. In the following we let v refer to a warping function:

- *Monotonicity*: Timing may vary between repetitions, but things rarely go backwards. Thus it is commonly required that warping functions are strictly increasing functions, ie. $t_2 > t_1 \Rightarrow v(t_2) > v(t_1)$.

- *Group structure*: The set of feasible warps for a given phase variation model should be a group under composition. Examples are the group of increasing affine transformations $\{x \mapsto \alpha x + \beta | \alpha > 0, \beta \in \mathbb{R}\}$, and the group of increasing diffeomorphisms on [0,1].

- *Smoothness*: Warping functions are generally assumed to obey some degree of smoothness, often enforced by L_{warp}. If $f : A \to B$ and $g : B \to C$ are C^{k_1} and C^{k_2} functions, respectively, then $g \circ f$ is a $C^{min(k_1,k_2)}$ function.

- *Fixed end points*: If we let $[a, b]$ be the domain of v, we require $v(a) = a$ and $v(b) = b$.

- *The identity as the mean*: The identity function should be the "most central" in the distribution of warps. In the framework of definition 1, this can defined as $\mathrm{E}[w] = 0$ with $v = \phi(w)$.

  This is slightly related to the identifiability issue – if we require $\mathrm{E}[w] = 0$, then we cannot apply a common, non-trivial warping function to the population of warps and still have $\mathrm{E}[w] = 0$.

So which of these mathematical properties are important? Well it certainly depends on the application. In the author's opinion, monotonicity and the identity as the mean are key properties, whereas fixed end points depend much on the application, and the group structure property is too strong and rules out many spline-based approaches. Generally speaking, we should choose a class of warping functions sufficiently flexible to capture the temporal variation in data, but there is little advantage of being more flexible than that, and a low-dimensional class of warping functions may be sufficient as demonstrated in Paper I.

Alignment of multivariate functional data Registration of multivariate functional data deserves a paragraph on its own. Multivariate functional data, $[a, b] \to \mathbb{R}^k$, have the important property that the images in $\mathbb{R}^k$ of the curves are unchanged by warping. This put a limit on how much variation in data that can be explained by warping effects; differences in resulting trajectories must be explained by amplitude or residual variance.

A key issue in registration of multivariate functional data is that any alignment procedure – directly or indirectly – involves a trade-off between optimal alignment of the individual coordinates. This generally amounts to introduce some kind of weighting of coordinates, which must be estimated from data. There is a correspondence between weights and amplitude variance; a small weight on coordinate x_i corresponds to a large amplitude variation for that coordinate.

1.2.3 Dynamical modelling of functional data using warped solutions of ODEs

In the previous paragraph we assumed that variation can be separated into two additive components: amplitude or residual variation and phase variation, which I believe ought to be estimated and modeled jointly.

However, it might be case that this separation cannot be done, and that the warping does not take effect on the trajectories but some underlying quantity such as a differential equation. Using derivatives for modeling functional data has been done to some extent; a notable example is *principal differential analysis* (PDA) introduced by Ramsay (1996). In this paragraph, we will present a new approach of modeling misaligned functional data.

First presented by Olsen & Tolver (2017), assume that the idealised system is generated by a differential equation on the form

$$x^{(m)}(t) = F(x(t), x'(t), \ldots, x^{(m-1)}(t), t) \tag{1.9}$$

where t denotes physical time, we introduce warping effects by changing the time argument to warped time:

$$x_i^{(m)}(t) = F(x_i(t), x_i'(t), \ldots, x_i^{(m-1)}(t), v_i(t)) \tag{1.10}$$

That is, the dynamics of the system is like the idealised system, but with the time parameter moved to $v_i(t)$. It is a crucial part of this model that the differential equations are **inhomogenous** as functions of t.

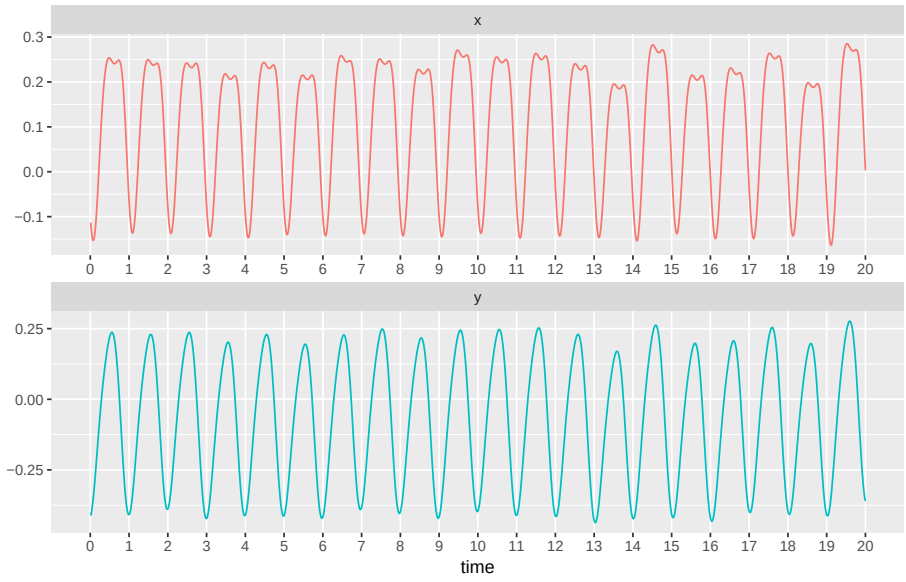


FIGURE 1.7: Twenty cycles of a two-dimensional simulation of functional data generated using a warped differential equation.

For instance, returning to our hand movement example, movement of the hand is governed by the laws of physics and inputs sent from the brain. This input (which determines the acceleration) will go through different phases depending on the state of the movement. These phases can be assumed to be (roughly) identical across repetitions, but the timing will likely vary between repetitions. Furthermore, there will be feedback from the current state to the input sent from the brain – the brain will automatically (and unconsciously) use the perceived position of the hand and cylinder to adjust the nerve signals sent to the hand.

Whereas all these interactions may be (almost) linear in nature; the result is a complex interplay between amplitude and phase variation that cannot be captured by common methods. An example of this is shown in Figure 1.7: the dynamics of this system is been generated by a linear (inhomogenous) second-order differential equation, which has been warped. Despite the fact that the derivative of the warp, v', differs by less than 6% from the constant function in this simulation, we observe much variation in data.

This discussion has so far been about locomotion, but these thoughts og ideas should not be restricted to locomotion; it may apply to other biological systems as well.

Although the best options available, current models for warping may not be the correct approach to functional data with temporal variation, if data truly are from complex, self-interacting systems where temporal variation is irremovable from other sources of variation.

1.3 Motivating introduction to the papers

1.3.1 Paper I

As mentioned in the previous section, multivariate functional data has got little attention in the FDA literature. In this paper, we look into dedicated methodology for multivariate misaligned functional data. Using a model by some of the authors as starting point, we extend this to misaligned, multivariate functional data. As part of the modelling, we develop a new low-parametric model called *dynamical correlation structure*, which allows for multivariate longitudinal data with time-varying cross-correlation.

There are three data examples: growth of Danish boys, repeated walking sequences and a hand movement experiment. We spent most time and devote more pages to the third data example, where all features in our model are exploited in full. In the end, we make a classification study on these nice data, comparing our model to state-of-the-art methods.

The paper is published as Olsen et al. (n.d.).

1.3.2 Paper II

A number of standard and often applied model for multivariate Gaussian data exists, such as *factor analysis* and *probabilistic prinipal component analysis*, but comparatively less literature exist for multivariate data measured over time. Another issue is the computational speed – inversion of matrices require $O(n^3)$ operations, so fast procedures that scale better with rank is of much interest.

In this paper we develop a full-scale framework for multivariate functional data called *Markov Component Analysis* (MCA). A strong property is that likelihood calculations etc. can be done in linear speed in terms of observations. The main idea is to use a number of underlying latent components that resembles a mixed effects model. Put into the right framework we get full-rank parametric model, where calculations require $O(n)$ operations, where n is the number of observation points per curve. There is a close connection between Markov Component Analysis and the Kalman filter, and we also develop a flexible model for warping functions within the MCA framework. We have preliminary results from a data application on trotting horses with artificially introduced lamenesses.

1.3.3 Paper III

Domain selection for functional data is a topic that has got little attention in literature.

In this article we extend *false discovery rates* – a popular and well-known quantifier for the multiple comparisons problem – to a functional data setting. Along with this, we extend the *Benjamini-Hochberg* procedure to functional data. There are many examples of correction procedures in literature, some which might be better than the BH procedure, but we focus on this procedure due to its simplicity and popularity.

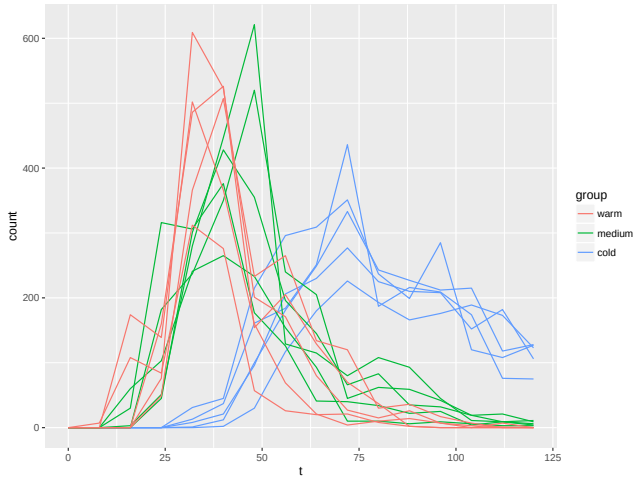


FIGURE 1.8: Data from Paper IV

Our approach can be applied to any open subset of $\mathbb{R}^k$, and weighting is allowed. We use pointwise p-values to define the the functional BH procedure. This is advantageous as we do not need to identify covariance structures or make use of fancy tests – the tests used to define p-values need only consider the pointwise values of the sample data, and so all the standard statistical tests are available. Theoretical results are shown, which also outline a simple algorithm.

We have two simulation studies and a really nice data example – satellite measurements of Earth temperatures. It's scientifically relevant, it has an unusual domain (S^2) and it has a hugely complicated covariance structure, that we prefer not to model. So we use the functional BH procedure – easy, fast and interpretable.

1.3.4 Paper IV

This manuscript is a collaboration with biologists using data from a master's project from Department of Plant and Environmental Sciences at University of Copenhagen. As the scope is more oriented towards biology, some background and motivation may be needed for the more statistically-oriented reader.

The context is a fungus pathogen that infects insects by discharging spores (conidia). It has a temporal pattern that make functional data methods interesting; previous studies have applied simple methods that does not take temporal variation into account.

Therefore, we wanted to apply the methodology of Paper I. The challenge is that data is inherently discrete with lots of zeros (see Figure 1.8), so Gaussian models would be inadequate. This calls for generalized linear models approaches. We extend the model of Paper I to use a negatively binomial response. This is successfully applied to data, and some inferential questions are also considered.

2 Supplementary material for papers

Scientific papers are usually a collaboration of several authors with individual opinions and ideas. More importantly, scientific papers must constrain themselves in terms of length and material and usually have a scope. Thus, a selection of material must take place and there might interest or relevant (background) material that is omitted. In this chapter we present supplementary material that was not included in the papers.

2.1 Continuous-time markov component analysis

This is an addendum to Paper II: *Markov component analysis for multivariate functional data*. The purpose is to extend the MCA to a continuous setting, and connect this to random variables and statistics.

Introduction In Manuscript II we are in the continuous-time domain when defining phase variation, but apart from that everything else is kept in a discrete formulation. As data is observed discretely, this is a natural approach, and relevant quantities can be defined in terms of matrices and vectors.

However, the Cholesky calculus of MCA can naturally be defined using a continuous domain, ie. instead of vectors and matrices we have functions and operators in $L^2([a,b];R^k)$, where $[a,b]$ is the domain of our data. A few things get easier for the continuous-domain MCA; notably we can easily separate multiplication operators (compare diagonal matrices) and integral operators.

An important reference is Markussen (2013), which applies differential operators in a functional data setting. In that paper, calculations can be performed in linear time, and the Green's functions related to the differential operators can be seen as covariance operators for Gaussian processes.

Setting and definitions

Setting The context is the function space $(L^2[a,b]\,;\,\mathbb{R}^d)$, i.e. the set of square integrable functions $[a,b] \to \mathbb{R}^d$ equipped with the standard inner product. We'll omit the d whenever possible. We will use $\mathcal{B}(L^2[a,b])$ to denote the space of bounded linear operators on $(L^2[a,b]\,;\,\mathbb{R}^d)$.

In this paper we deal with certain subclasses within the set of sums of bounded integral operators and multiplication operators on $(L^2[a,b];\mathbb{R}^d)$. It is well-known that this set is a linear subspace of $\mathcal{B}(L^2[a,b])$ that is closed under composition.

Definition 2 (Integral operator)**.** An *integral operator* is an operator $K \in \mathcal{B}(L^2[a,b])$ given by

$$Kf(t) = \int_a^b K(t,s)f(s)\,\mathrm{d}s$$

$K(\cdot,\cdot)$ is called the *kernel* of the operator. We will use K to refer both to the function and the kernel unless otherwise stated.

An integral operator K is *self-adjoint* iff $K(s,t) = K(t,s)$ for a.a. $s,t \in [a,b]$.

Proposition 2.1.1. *Let K and L be integral operators in $\mathcal{B}(L^2[a,b])$, and let M be a multiplication operator in $\mathcal{B}(L^2[a,b])$. Then the adjoint of K, K^*, has kernel*

$$K^*(s,u) = K(u,s)^\top, \quad s,u \in [a,b] \tag{2.1}$$

The composition $M = KL$ is also an integral operator, which has kernel:

$$M(s,t) = \int_a^b K(s,u)L(u,t)\,du, \quad s,t \in [a,b] \tag{2.2}$$

Proof: Reference.

Definition 3. A *lattice operator* is an integral operator in $\mathcal{B}(L^2[a,b])$ with kernel

$$K(s,t) = \begin{cases} \alpha(s)\beta(t)^\top \in \mathbb{R}^{d\times d} & a < t \le s < b \\ \beta(s)\alpha(t)^\top \in \mathbb{R}^{d\times d} & a < s \le t < b \end{cases} \tag{2.3}$$

It assumed that $\alpha(t), \beta(t) \in \mathbb{R}^{d\times q}$ for some common $q \ge 1$ with $\alpha(t)\beta(t)^\top$ symmetric for all t. We shall refer to q as the **order** of the lattice operator (whereas d is reserved for dimension). A lattice operator of order 1 is called a *simple* or *prime lattice operator*. Any lattice operator can viewed as a sum of prime lattice operators. If $d = 1$, there is no need take care of transpose matrices and the formula reduces to:

$$\alpha(s \vee t)\beta(s \wedge t)^\top \in \mathbb{R}^{d\times d} \quad s,t \in [a,b] \tag{2.4}$$

Notation. *We shall identify a multiplication operator M with its pointwise multiplication function and use the same symbol for both, i.e.*

$$Mf(t) = M(t)f(t) \quad \textit{for } t \in [0,1]$$

We'll always assume that M as a function is continous everywhere.

The most important multiplication operator is of course the identity operator I, which simply corresponds to a 'white noise' Gaussian process.

Definition A *factorizable operator* (of order q) is the sum of a lattice operator and a positive definite multiplication operator, where q refers to the order of the lattice operator. Note that a multiplication operator is positive definite iff its entries are positive definite almost surely.

Definition 4 (Triangular operator)**.** A *triangular operator*, or alternatively *lower triangular operator*, is an integral operator O with kernel:

$$O(s,u) = 1_{u \leq s} \alpha(s) f(u)^\top, \quad u, s \in [a,b] \tag{2.5}$$

Here it assumed that $\alpha(t), f(t) \in \mathbb{R}^{d \times q}$ for some common $q \geq 1$, which we shall refer to as the *order* of the operator.

Definition 5. An MCA operator is the sum of a triangular operator and a multiplication operator.

Assume an MCA operator is given by

$$Tg(t) = D(t) + \alpha(t) \int_a^t f(s)^\top g(s)\, d\mathrm{s} \tag{2.6}$$

Then we shall refer to the triple (D, α, f) as its *components*, which will be another way of specifying an MCA operator. The order of MCA operator is defined as the order of the lattice operator.

It is easily seen that sums and composition of MCA operators are again an MCA operator. We will consider inverses of MCA operators in Section 2.1

What's the main idea of using these operators? As we shall see, there is a correspondence between factorizable operators and MCA operators. Factorizable operators can be viewed as covariances of Gaussian stochastic processes, while MCA operators have some really nice computational properties and can be seen as the Cholesky decompositions of factorizable operators. The interesting details will be described in the following:

Decomposing factorizable operators into MCA operators

In this section we shall see how to decompose a factorizable operator $F = K + M$ into triangular operators of the same order s.t. $F = TT^*$.

First the special case $M = I$:

Proposition 2.1.2. *Let K be a lattice operator of order q. Then $I + K = (I + O)(I + O^*)$ where O is a triangular operator of order q.*

In particular, if $K(s,t) = \begin{cases} \alpha(s)\beta(t)^\top \in \mathbb{R}^{d \times d} & a < t \leq s < b \\ \beta(s)\alpha(t)^\top \in \mathbb{R}^{d \times d} & a < s \leq t < b \end{cases}$
then the kernel of O is given by:

$$O(s,u) = 1_{u \leq s} \alpha(s) f(u)^\top, \quad u, s \in [a,b] \tag{2.7}$$

where α is the same, and f satisfy the non-linear forward integral equation:

$$\beta(t) = f(t) + \alpha(t) \int_a^t f(s)^\top f(s)\, ds \tag{2.8}$$

Although difficult to solve explicitly, we emphasize that f can be calculated in linear time of t.

Proof. We have $(I+O)(I+O)^* = I+O+O^*+OO^*$ and thus O must satisfy $K = O+O^*+OO^*$. Assuming O has components $\tilde{\alpha}, f$ it is seen that we have a solution if:

$$\alpha(s)\beta(t)^\top = \tilde{\alpha}(s)f(t)^\top + \tilde{\alpha}(s) \int_a^t f(u)^\top f(u)\tilde{\alpha}(t)^\top \, d\mathrm{u} \quad a < t \leq s < b \tag{2.9}$$

$$\beta(s)\alpha(t)^\top = f(s)\tilde{\alpha}(t)^\top + \tilde{\alpha}(s) \int_a^s f(u)^\top f(u)\tilde{\alpha}(t)^\top \, du \quad a < s \leq t < b \tag{2.10}$$

Setting $\tilde{\alpha} = \alpha$ and assuming (2.8), we get the desired result. □

Now for the general proposition:

Theorem 2.1.3 (Decompostion of factorizable operators)**.** *Let $F = K + M$ be a factorizable operator of order q, where K is the integral part and M the multiplication part. Then F decomposes $F = TT^*$, where $T = O + D$ is an MCA operator of order q with triangular part O and multiplication part D.*

Assume K has kernel $K(s,t) = \begin{cases} \alpha(s)\beta(t)^\top \in \mathbb{R}^{d\times d} & a < t \leq s < b \\ \beta(s)\alpha(t)^\top \in \mathbb{R}^{d\times d} & a < s \leq t < b \end{cases}$

Then $D(t) = M(t)^{1/2}$, and the the kernel of O is given by:

$$O(s,u) = 1_{u\leq s}\alpha(s)f(u)^\top, \quad u, s \in [a,b] \tag{2.11}$$

where α is the same and f satisfy the non-linear forward integral equation:

$$\beta(t) = D(t)f(t) + \alpha(t) \int_a^t f(s)^\top f(s)\, ds \tag{2.12}$$

Proof. A slight generalization of propostion 2. □

Note that positive semidefiniteness of M is sufficient for Theorem 2.1.3 to work.

Corollary 2.1.4. *Let K be a lattice operator of order q. Then $K = OO^*$, where O is a triangular operator of order q.*

Integral and differential equations It is easily seen that under mild assumptions (2.8) and (2.11) can be re-written in terms of differential equations for f. However, these differential equations will not be linear as they contain quadratic terms of f! It is known that such differential equations may show explosion behaviour, and are thus not *a priori* guaranteed to have a valid solution.

Inversion of MCA operators

One particularly nice feature of MCA operators is the fact that finding or applying inverses correspond to solving differential equations.

Theorem 2.1.5. *Let T be an MCA operator with $T = O + D$. Let $x \in L^2[a, b]$. Assume $y = T^{-1}x \Leftrightarrow x = Ty$. y can be found using:*

$$y(t) = D(t)^{-1}[x(t) - \alpha(t) \int_a^t f(s)^\top y(s)\, ds] \tag{2.13}$$

Alternatively, the following formula using an intermediate variable v may be more suitable to implementation:

$$\begin{aligned} v(0) &= 0 \\ y(t) &= D(t)^{-1}[x(t) - \alpha(t)v(t)] \\ v'(t) &= f(t)^\top y(t) \end{aligned} \tag{2.14}$$

Proof. It is easy to verify that said formula gives both the right and left inverse of T. That the stated formula gives rise to a linear operator under the assumption that D is bounded away from 0 is easily verified. □

The explicit inversion of an MCA operator as an MCA operator itself is described in the follwing theorem:

Theorem 2.1.6. *Let $T = O + D$ be an MCA operator of order q with D positive definite. Assume O has kernel:*

$$O(s, u) = 1_{u \le s} \alpha(s) f(u)^\top, \quad u, s \in [a, b] \tag{2.15}$$

Then T has an inverse $T^{-1} = D^{-1} + \tilde{O}$ that is also an MCA operator of order q with kernel for $\tilde{O}$:

$$O(s, u) = 1_{u \le s} \tilde{\alpha}(s) \tilde{f}(u)^\top, \quad u, s \in [a, b], \tag{2.16}$$

if κ given by the forward integral:

$$\kappa(a) = I_q, \quad \kappa'(t) = -f(t)^\top D(t)^{-1} \alpha(t) \kappa(t) \quad \text{for } t \in [a, b] \tag{2.17}$$

is non-singular for all $t \in [a, b]$. The lattice components of T^{-1} are given by:

$$\tilde{\alpha}(t) = D(t)^{-1} \alpha(t) \kappa(t), \quad \tilde{f}(t) = -D(t)^{-1} f(t) \kappa(t)^{-1,\top} \tag{2.18}$$

Note that Theorem 2.1.5 can be applied in some cases in which Theorem 2.1.6 is not applicable.

Proof. If T_1 and T_2 are two MCA operators with components (D_1, α_1, f_1) and (D_2, α_2, f_2), respectively, then by changing the order of integration, we get that $T_1 T_2 g(t)$ equals:

$$\begin{aligned} &D_1(t) D_2(t) g(t) + \\ &\int_a^t \left[D_1(t) \alpha_2(t) f_2(u)^\top + \alpha_1(t) f_1(u)^\top D_2(u) + \alpha_1(t) \int_u^t f_1(s)^\top \alpha_2(s)\, ds\, f_2(u) \right] g(u)\, du \end{aligned} \tag{2.19}$$

Identifying $T = T_1$, and setting $D_2(t) = D_1(t)^{-1}$, and α_2, f_2 as in the proposed operator, the rest is an exercise to verify that the expression inside of the brackets is 0 for all choices of u and t. It is verified analogously that we also have a left inverse. □

Factorizable operators and the Wiener processes

There is a strong connection between lattice operators and stochastic integrals, where the integrand is deterministic. We refer to Adler & Taylor (2009) for a definition of deterministic stochastic integration.

Proposition 2.1.7. *Let K be a lattice operator of order q with kernel*

$$K(s,t) = \begin{cases} \alpha(s)\beta(t)^\top \in \mathbb{R}^{d\times d} & a < t \leq s < b \\ \beta(s)\alpha(t)^\top \in \mathbb{R}^{d\times d} & a < s \leq t < b \end{cases} \tag{2.20}$$

Then $K = OO^*$, where O is an triangular operator of order q, s.t. $O(s,u) = 1_{u\leq s}\alpha(s)f(u)^\top$ where f satisfies $\alpha(t)\int_a^t f(s)^\top f(s)\, d\mathrm{s} = \beta(t)$.

Furthermore, let X be a (Gaussian) stochastic process given by

$$X(t) = \alpha(t)\int_a^b f(s)\, d\mathrm{W_s} \tag{2.21}$$

where W_s is a q-dimensional Wiener process. Then $Cov(X(s), X(t)) = K(s,t)$.

One can say X is generated by O "acting" on the Wiener process.

Proof. Corollary 2.1.4. □

The last part is a simple way of describing the generating process of X, and Proposition 2.1.7 connects the theory of lattice operators with that of stochastic integrals.

Statistics

Continuous-time MCA has the potential for being used in a statistical setting, similar to the discrete-time MCA. As data is observed discretely, one has to define an embedding or smoothing of discrete observations into $L^2[a,b]$.

We will not go into details on this, but we remark that since none of the operators use 'local properties' (ie. involves derivatives), there is a certain degree of robustness in relation to the embedding of the data. One approach is to follow that of Markussen (2013).

In order to define likelihood expressions, we need a definition for the log-determinant of MCA and factorizable operators:

Definition 6 (Log-determinant)**.** Let T be an MCA operator with components (I, α, f). The log-determinant of T is defined as the "diagonal integral":

$$\log\det T = \tfrac{1}{2}\int_a^b \operatorname{tr}(\alpha(t)f(t)^\top)\,dt \tag{2.22}$$

Let K be a factorizable operator with multiplication part I. If K decomposes $K = TT^*$, we define the log-integral of K as $\log\det K = 2\log\det T$.

Why this seemingly odd definition of a log-determinant? We do not even use the logarithm at all. However, it makes sense as a limit of matrix approximations of T. If M is a function $[0,1] \to \mathbb{R}^k$, then under regularity assumptions we have,

$$\sum_{i=1}^N \log\det|I + M(\tfrac{i}{N})/N| \approx$$

$$\sum_{i=1}^N \log(1 + \operatorname{tr}[M(\tfrac{i}{N})/N]) \approx \sum_{i=1}^N \operatorname{tr}[M(\tfrac{i}{N})/N] \approx \int_0^1 \operatorname{tr} M(t)\,dt \tag{2.23}$$

which should be related to the formula for the log-determinant of a triangular matrix $\mathbf{T} \in \mathbb{R}^{n\times n}$:

$$\log\det\mathbf{T} = \sum_{i=1}^n \log\mathbf{T}_{ii} \tag{2.24}$$

The determinant of the inverse behaves in the usual way:

Proposition 2.1.8. *Let T be an MCA operator with components (I, α, f). Then* $\log\det(T^{-1}) = -\log\det T$.

Proof. Let $(I, \tilde\alpha, \tilde f)$ be the components of T^{-1}. By Theorem 2.1.6 this is valid, and we easily see that $\tilde\alpha(t)\tilde f(t)^\top = -\alpha(t)f(t)^\top$, and thus $\log\det(T^{-1}) = -\log\det T$. □

Let N functional observations $y_1, \ldots, y_N$ in $L^2[a,b]$ from the MCA model $y_i \sim N(0, TT^*)$ be given. Using the framework presented, we define the log-likelihood as:

$$l_y(T) := 2N\log\det T + \sum_{i=1}^N \int_a^b ||T^{-1}y_i(t)||^2\,d\mathrm{t} \tag{2.25}$$

If T^{-1} is of order q has components $(I, \tilde\alpha, \tilde f)$, this becomes:

$$-N\int_0^1 \operatorname{tr}\tilde\alpha(t)\tilde f(t)^\top d\mathrm{t} + \sum_{i=1}^n \int_a^b \tilde f(s)\int_a^s \tilde\alpha(s)^\top y_i(s)\,ds\,d\mathrm{t} \tag{2.26}$$

which can easily be evaluated in linear time.

As the space of possible αs or fs is infinite-dimensional, one has to put restrictions on α and f (or $\tilde\alpha$ and $\tilde f$), which could be in terms of a finite basis expansion or by adding a penalisation term. Note that the relationship between the components of T and T^{-1} is highly non-linear. It is possible to write down functional derivatives of (2.25) for $\tilde\alpha$ and $\tilde f$ which can be used for gradient descent-algorithms and solving score equations.

Discussion

We have seen that MCA can naturally be defined in a continuous-time setting, and that it has similar properties as and formulae reflecting discrete MCA. It has yet to be implemented in a statistical setting, but we devised a framework and we expect it to be robust in relation to the embedding of data into function space.

2.2 Interval-wise testing procedure for spherical domains with application to Earth temperature data

This is an addendum to Paper III: *False discovery rates for functional data*. The purpose is to extend the *interval-wise testing* procedure (IWT) (Pini & Vantini 2017) to spherical domains and apply it to the Earth temperature data.

For the ease of presentation we consider only the unit sphere, S^2. The extension to general spheres and squares is trivial. Measurability will be assumed whenever needed.

The framework is not restricted to any particular class of tests, but we propose to use permutation tests as done in Pini & Vantini (2017). The main reason for using permutation tests is that they are exact tests which are asymptotically as powerful as parametric tests; a great advantage in the setting of functional data.

2.2.1 The one-dimensional interval-wise testing prcedure

Assume that we observe M functional signals $\xi_1, \dots, \xi_M : (0,1) \to \mathbb{R}$. For each interval $I = (a,b) \subseteq (0,1)$ define the corresponding interval-wise null hypothesis H_0^I as

$$H_0^I : \quad \xi_1(t) \stackrel{D}{=} \cdots \stackrel{D}{=} \xi_M(t) \quad \forall t \in I \tag{2.27}$$

and alternative hypothesis:

$$H_A^I : \quad \xi_i(t) \stackrel{D}{\neq} \xi_j(t) \text{ **for some** } t \in I, i, j \in \{1, \dots, M\} \tag{2.28}$$

Assume we are given interval-wise p-values with the property that (H_0^I true) $\Rightarrow p^I \sim U(0,1)$ for all intervals I.

Define the unadjusted and adjusted p-value functions respectively by:

$$p, \tilde{p} : (0,1) \to [0,1], \quad p(t) = \limsup_{I \to t} p^I, \quad \tilde{p}(t) = \sup_{t \in B} p^I$$

where $I \to t$ indicates that the endpoints of the interval converge to t.

Theorem 2.2.1. *The adjusted p-value function provides control of the interval-wise error rate. That is for $\alpha \in (0,1)$:*

$$\forall I : \quad H_0^I \text{ *is true*} \Rightarrow P[\forall t \in I : \tilde{p}(t) \leq \alpha] \leq \alpha \tag{2.29}$$

Proof. Theorem A3 of Pini & Vantini (2017). □

2.2.2 Spherical IWT

Definition 7 (Interval-wise testing on spheres)**.** For each point $p \in S^2$ and radius $r > 0$, let $B_{p,r}$ be the associated ball, and define $\mathcal{B}_{S^2}$ as the set of all balls on S^2 with positive radii. Note that the following construction does not depend on whether we use geodesic or euclidean distance on the sphere.

Now assume that we observe M real-valued functional signals, $\xi_1, \dots, \xi_M : S^2 \to \mathbb{R}$. For each ball $B \in \mathcal{B}_{S^2}$, let H_0^B be the null hypothesis and alternative hypothesis

$$H_0^B : \quad \xi_1(t) \overset{D}{=} \cdots \overset{D}{=} \xi_M(t) \quad \forall t \in B \tag{2.30}$$

$$H_A^B : \quad \xi_i(t) \overset{D}{\neq} \xi_j(t) \text{ **for some** } t \in B, i, j \in \{1, \dots, M\} \tag{2.31}$$

For each B, let p_B be a p-value function that satisfies (H_0^B true) $\Rightarrow P(p_B \leq \alpha) \leq \alpha$.

Define the *unadjusted p-value function*:

$$p : S^2 \to [0,1], \quad p(t) = \limsup_{B \to t} p_B$$

where $B \to t$ is understood as balls containing t with decreasing radii.

Define the *adjusted p-value function*

$$\tilde{p} : S^2 \to [0,1], \quad \tilde{p}(t) = \sup_{B \ni t} p_B$$

The un-adjusted and adjusted p-value functions control the pointwise and ball-wise error-rates respectively.

Proposition 2.2.2. *The adjusted p-value function provides control of the ball-wise? error rate, that is for $\alpha \in (0,1)$:*

$$\forall B \in \mathcal{B}_{S^2} : \quad H_0^B \text{ is true } \Rightarrow P[\forall t \in B : \tilde{p}(t) \leq \alpha] \leq \alpha \tag{2.32}$$

Proof. Analogous to Theorem A3 of Pini & Vantini (2017). □

Remark 2.2.3. *In this definition of spherical IWT, we adjust the p-value of $t \in S^2$ by all balls containing t. This is but one choice of adjustments, and one could choose smaller as well as larger adjustment sets, e.g. all convex sets containing t. However (as for most other multiplicity correction procedures), there is a trade-off between the power of the test and "the amount of correction": using larger adjustment sets leads to less power.*

Implementing IWT as permutation tests using uniform sampling

In order to apply spherical IWT, we propose to use permutation tests based on

$$T^B = \int_B t_x \, dx, \quad B \in \mathcal{B}_{S^2} \tag{2.33}$$

where t_x is an appropriate test statistic, defined pointwise on the sphere. Due to the infinite amount of balls in $\mathcal{B}_{S^2}$ we naturally have to rely on a finite approximation. However unlike Euclidian domains, spheres have positive curvature. This gives some challenge when implementing IWT on spheres since we cannot use a uniform grid for approximating the test statistics, which would be the natural choice in rectangular and one-dimensional cases.

Algorithm Assume we have M smooth observations from the sphere, $\xi_1, \dots, \xi_M : S^2 \to \mathbb{R}$, such that $\xi_1 \overset{D}{=} \cdots \overset{D}{=} \xi_M$ under the global null hypothesis. Let $t : \mathbb{R}^M \to \mathbb{R}_+$ be a suitable test statistic. We propose to use the following algorithm:

1. Sample N points uniformly, but not necessarily independently, on the sphere. Denote these points $P = \{s_1, \dots, s_N\}$.

2. Define the point sets $D_{ij} = \{s \in P : |s_i - s| \leq |s_i - s_j|\}$. These sets approximate the balls in S^2, and will be used as the sets on which to evaluate the test statistics.

3. Calculate the observed test statistics t_{ij}^{obs} on observed data $y_{D_{ij}1}, \dots y_{D_{ij}M}$, using only points in D_{ij} for $i, j = 1, \dots, N$:

$$t_{ij}^{obs} = \sum_{s \in D_{ij}} t(\xi_1(s), \dots, \xi_M(s)) \tag{2.34}$$

4. For each permutation $r^k = (r_1, \dots, r_M)$ of $1, \dots, M$, do:

   Calculate test statistic $t_{ij}^{r^k}$ on permuted data $y_{D_{ij}r_1}, \dots, y_{D_{ij}r_M}$, using only those points in D_{ij} for $i, j = 1, \dots, N$:

$$t_{ij}^{r^k} = \sum_{s \in D_{ij}} t(\xi_{r_1}(s), \dots \xi_{r_M}(s))$$

5. For every approximating ball D_{ij} compare test statistics of permutations with test statistic of observation to obtain ball-wise p-values $p^{D_{ij}}$, ie.

$$p^{D_{ij}} = (\#\text{permutations})^{-1} \sum_{k \in \text{permutations}} 1(t_{ij}^{obs} \leq t_{ij}^{r^k})$$

6. Finally, set $p(s_i) = p_i = p^{D_{ii}}$ and

$$\tilde{p}(s_i) = \tilde{p}_i = \max_{D_{ij} \ni s_i} p^{D_{ij}}$$

   to obtain adjusted and unadjusted p-values.

This approximation will give a good picture of the spherical IWT. It avoids using grids and similar, and has a simple implementation.

Significance level	Unadjusted p-values	IWT	fBH
0.10	0.410	0.098	0.252
0.05	0.323	0.047	0.169
0.01	0.185	0.011	0.064
0.001	0.085	0.0001	0.022

TABLE 2.2: Areas of significance at various significance levels as percentage of Earth total

Remark: The aggregated test statistic (2.34) (and similar for the permutations) can be replaced by any meaningful test. However, the use of a sum in (2.34) has an obvious computational advantage when evaluating the test across all D_{ij}.

2.2.3 Application to Earth climate data with comparison to fBH procedure

In this section we apply the spherical IWT to Earth climate data of Paper III. We refer to the manuscript for a description of the data set and setup.

IWT We sampled 10000 points uniformly on the sphere. For each of the 25 years in the data set we applied a local linear smoother using the kernel $K(x, y) = \max(\frac{\pi}{180} - d(x, y), 0)$, where d is geodesic distance on the sphere, measured in radians.

We used the algorithm outlined in section 2.2.2 with $B = 2000$ permutations. As test statistic we used $t(x_1, \ldots, x_{25}) = \max(T(x_1, \ldots, x_{25}), 0)$, where T is the t-statistic from the linear regression $y_{ij} = \alpha + \beta_j \text{year}_i + \text{noise}$.

fBH To perform the fBH procedure, we mapped the sphere into $\mathbf{T} = (-\pi, \pi) \times (-\pi/2, \pi/2)$ by (scaled) polar coordinates, ie. longitude and latitude. This mapping gives rise to a measure $f \cdot \mu$ on $\mathbf{T}$ where f is proportional to cos(latitude). This measure gives uniform weights to all points on Earth, assuming Earth to be a perfect sphere.

To ensure that we got a fair comparison between IWT and fBH approaches, we applied the BH procedure to the same points that were used in the IWT scheme. Since the chance of sampling a point is proportional to cos(latitude), this sampling scheme was a good approximation of the fBH procedure as defined in Paper III. One-sided t-tests were used for obtaining unadjusted p-values.

Note that due to different implementations, the numbers here are slightly different from those in Paper III.

Results

The coverage areas at various significance levels are provided in Table 2.2.

There were large differences between adjusted p-value functions and subsequent inference between fBH procedure and IWT. Interval-wise testing was much more conservative than the

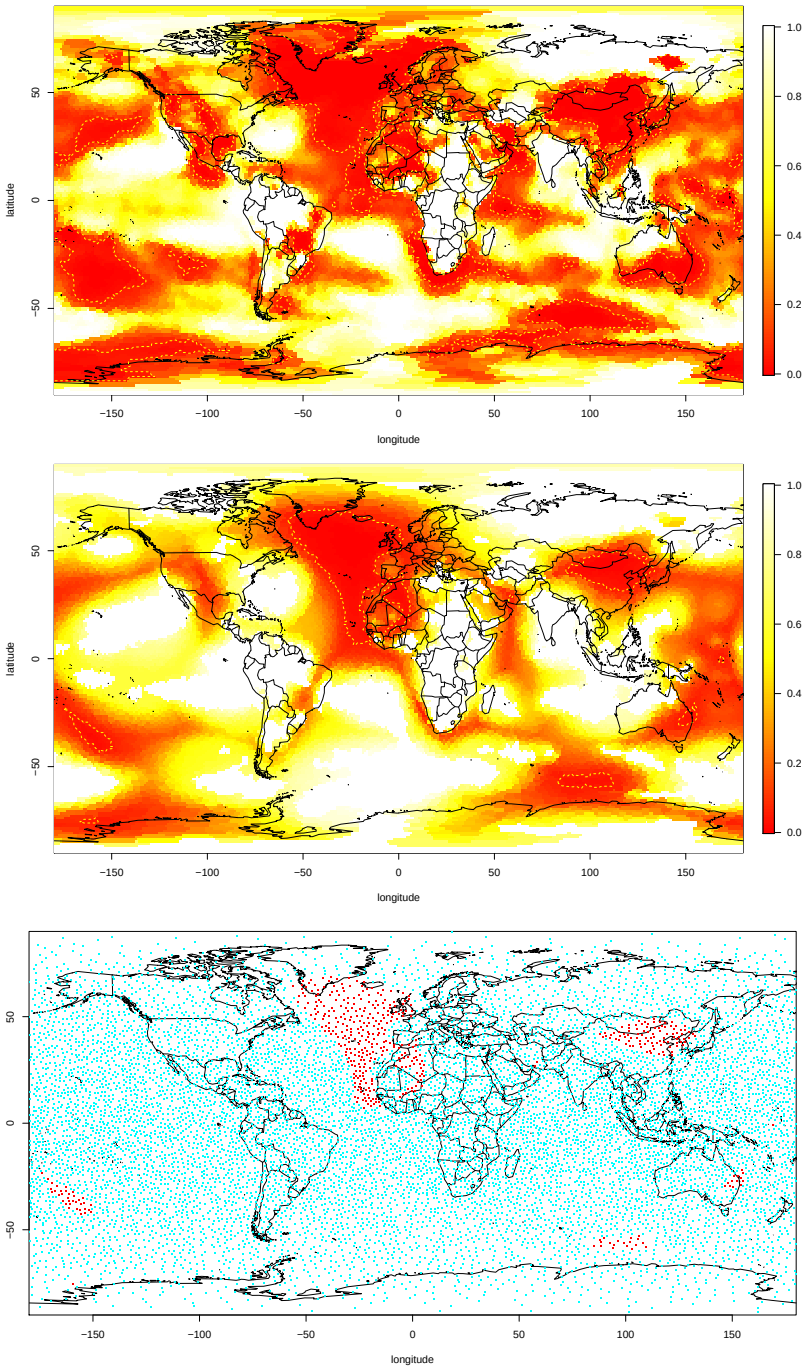


FIGURE 2.1: Upper plot: FDR-adjusted p-values. Middle plot: IWT-adjusted p-values. Dashed lines indicate 5% significance levels. Lower plot: Sampling points for IWT, color-coded at 5% significance level for IWT.

fBH procedure in selecting significant areas, and the IWT-selected regions were also much smoother. A notable reason for this is the fact that the fBH procedure is a global procedure that does not take proximity into account, while IWT is a local procedure based on proximity of points. If we take look at the map(s), the North Atlantic Ocean and northern China stands out; it is evidentr that these regions have experienced temperatures far above the normal in the latest years with the adverse weather effects this may cause.

2.2.4 Discussion

To the author's best knowledge, this is the first published extension of the the interval-wise testing framework to non-interval domains. We devised an algorithm for implementing IWT on S^2, and applied it to the Earth climate data set. No doubt the IWT-procedure can be extended to other other domains such as 2D and 3D Euclidian domains. As Euclidian domains have zero curvature, implementation should be easier than for spherical domains, but choosing "correction sets" (cf. Remark 2.2.3) will remain an issue.

In our application, IWT turned out to be quite conservative in comparison to the fBH procedure. For future studies, a comparison with other FWER-controlling procedures would be interesting. For this particular data set, the inclusion of covariates such land/sea would be interesting perspectives.

Bibliography

Abramowicz, K., Häger, C. K., Pini, A., Schelin, L., Sjöstedt de Luna, S. & Vantini, S. (2018), 'Nonparametric inference for functional-on-scalar linear models applied to knee kinematic hop data after injury of the anterior cruciate ligament', *Scandinavian Journal of Statistics* .

Adler, R. J. & Taylor, J. E. (2009), *Random Fields and Geometry*, Springer Science & Business Media.

Daubechies, I. (1992), *Ten lectures on wavelets*, Vol. 61, Siam.

Diggle, P. J., Heagerty, P., Liang, K.-Y. & Zeger, S. L. (2002), *Analysis of Longitudinal Data*, Oxford University Press.

Grimme, B. (2014), Nachweis und Analyse elementarer Invarianten als Bausteine menschlicher Armbewegungen, PhD thesis, Internationalen Graduiertenschule Biowissenschaften, Ruhr-Universität Bochum.

Hadjipantelis, P. Z., Aston, J. A., Müller, H.-G. & 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.

Horváth, L. & Kokoszka, P. (2012), *Inference for Functional Data with Applications*, Vol. 200, Springer Science & Business Media.

Koch, I. (2013), *Analysis of Multivariate and High-Dimensional Data*, Vol. 32, Cambridge University Press.

Markussen, B. (2013), 'Functional data analysis in an operator-based mixed-model framework', *Bernoulli* **19**(1), 1–17.

Marron, J. S., Ramsay, J. O., Sangalli, L. M. & Srivastava, A. (2015), 'Functional data analysis of amplitude and phase variation', *Statistical Science* **30**(4), 468–484.

Olsen, N. L., Markussen, B. & Raket, L. L. (n.d.), 'Simultaneous inference for misaligned multivariate functional data', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **67**(5), 1147–1176.

Olsen, N. L. & Tolver, A. (2017), 'Dynamical modelling of functional data using warped solutions of odes'. Poster presented at International Workshop for Functional and Operatorial Statistics, La Coruña, Spain.

Pini, A. & Vantini, S. (2017), 'Interval-wise testing for functional data', *Journal of Nonparametric Statistics* **29**(2), 407–424.

Raket, L. L., Grimme, B., Schöner, G., Igel, C. & Markussen, B. (2016), 'Separating timing, movement conditions and individual differences in the analysis of human movement', *PLoS Computational Biology* **12**(9), e1005092.

Raket, L. L., Sommer, S. & Markussen, B. (2014), 'A nonlinear mixed-effects model for simultaneous smoothing and registration of functional data', *Pattern Recognition Letters* **38**, 1–7.

Ramsay, J. O. (1982), 'When the data are functions', *Psychometrika* **47**(4), 379–396.

Ramsay, J. O. (1996), 'Principal differential analysis: Data reduction by differential operators', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 495–508.

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

Ramsay, J. O., Wickham, H., Graves, S. & Hooker, G. (2018), *fda: Functional Data Analysis*. R package version 2.4.8.
URL: *cran. r-project. org/ web/ packages/ fda/*

Sangalli, L. M., Secchi, P., Vantini, S. & Veneziani, A. (2009), 'A case study in exploratory functional data analysis: geometrical features of the internal carotid artery', *Journal of the American Statistical Association* **104**(485), 37–48.

Sørensen, H., Tolver, A., Thomsen, M. H. & Andersen, P. H. (2012), 'Quantification of symmetry for functional data with application to equine lameness classification', *Journal of Applied Statistics* **39**(2), 337–360.

Srivastava, A. & Klassen, E. P. (2016), *Functional and Shape Data Analysis*, Springer.

Thomsen, M. H., Jensen, A. T., Sørensen, H., Lindegaard, C. & Andersen, P. H. (2010), 'Symmetry indices based on accelerometric data in trotting horses', *Journal of biomechanics* **43**(13), 2608–2612.

Vantini, S. (2012), 'On the definition of phase and amplitude variability in functional data analysis', *Test* **21**(4), 676–696.

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

I

Simultaneous inference for misaligned multivariate functional data

Niels Lundtorp Olsen
Department of Mathematical Sciences
University of Copenhagen

Bo Markussen
Department of Mathematical Sciences
University of Copenhagen

Lars Lau Raket
Department of Mathematical Sciences
University of Copenhagen

Publication details
Published in *Journal of Royal Statistical Society, Series C*. doi:10.1111/rssc.12276

Simultaneous inference for misaligned multivariate functional data

Niels Lundtorp Olsen, Bo Markussen and Lars Lau Raket
Department of Mathematical Sciences, University of Copenhagen

Abstract

We consider inference for misaligned multivariate functional data that represents the same underlying curve, but where the functional samples have systematic differences in shape. In this paper we introduce a class of generally applicable models where warping effects are modelled through nonlinear transformation of latent Gaussian variables and systematic shape differences are modelled by Gaussian processes. To model cross-covariance between sample coordinates we propose a class of low-dimensional cross-covariance structures suitable for modeling multivariate functional data. We present a method for doing maximum-likelihood estimation in the models and apply the method to three data sets. The first data set is from a motion tracking system where the spatial positions of a large number of body-markers are tracked in three-dimensions over time. The second data set consists of longitudinal height and weight measurements for Danish boys. The third data set consists of three-dimensional spatial hand paths from a controlled obstacle-avoidance experiment. We use the developed method to estimate the cross-covariance structure, and use a classification set-up to demonstrate that the method outperforms state-of-the-art methods for handling misaligned curve data.

Keywords: functional data analysis, curve alignment, nonlinear mixed-effects models, template estimation

1 Introduction

While the literature and available methods for statistical analysis of univariate functional data have been rapidly increasing during the last two decades, multivariate functional data has been a largely overlooked topic. Extension of univariate methodology to multivariate functional data is often considered a trivial task, but is rarely done in practice. As a result, the non-trivial parts of extending methodology, such as temporal modeling of cross-covariance or warping of misaligned multidimensional signals, have only received little attention.

A wide range of methods for aligning curves are available. For general reviews of the literature on curve alignment, we refer to Ramsay & Silverman (2005), Kneip & Ramsay (2008), and Wang et al. (2015). Curve alignment is a nonlinear problem, so for the vast majority of methods, one can not generally expect to align data in a globally optimal way. In the multitude of available methods for univariate functional data, the quality of the results obtained with the available implementations is very variable. Often, good implementations of simple methods outperform far more advanced methods with less polished implementations, even if the advanced methods should be more suitable to the data at hand. From the perspective of multivariate functional data, a major issue is that only very few methods with publicly available implementations support alignment of multivariate curves.

While misaligned multivariate functional data have been underrepresented in the statistics literature, similar problems have had a central role in other fields. Analysis of misaligned curves in multiple dimensions is fundamental in the shape analysis literature (Younes 1998, Sebastian et al. 2003, Manay et al. 2006), where for example closed planar shapes can be thought of as functions $f : [0, 1] \to \mathbb{R}^2$ with $f(0) = f(1)$. In much shape data, one do not observe the parametrization of these functions, and for closed shapes the start and end points (0 and 1) of the parametrization are arbitrary in terms of the observed data. As an example, consider data consisting of cells' outlines obtained from 2D images that have been manually annotated. Here the first annotated point on a cell does not bear any significance—in fact the orientation of the cell is most likely completely random in the image. For this reason, a fundamental direction of theory in the shape analysis literature is built around invariance to parametrization of the function (Younes 1998) as well as other classical shape invariances such as translation, scaling and rotation (Kendall 1989, Dryden & Mardia 1998).

In recent years, the idea of using invariances similar to the shape analysis literature has been introduced as a general tool to analyze functional data (Vantini 2012). The most notable class of methods are based on elastic distances for functional data analysis (Srivastava et al. 2011, Kurtek et al. 2012, Tucker et al. 2013, Srivastava & Klassen 2016). The fundamental idea underlying these methods is to represent data in terms of square-root velocity functions and take advantage of the invariance properties of distance on the associated function space, in particular that distances are not affected by warping of the domain in the observed representation. An elastic distance between two curves f_1 and f_2 can be defined as the minimal distance between the square-root velocity functions associated to f_1 and $f_2 \circ v$ where the

minimum is taken over all possible warps v of f_2 (in the original representation). This approach has proven very successful compared to many conventional approaches, and efficient high-quality implementations for various data types and types of analyses are available (FSU n.d., Tucker 2017).

The vast majority of available methods for handling misaligned functional data are heuristic in the sense that they are based on some choice of data similarity measure that is typically not chosen because it fits well with important characteristics of the data. Rather, the typical rationale is computational convenience and/or incremental improvements over other methods. In the shape literature, methods are perhaps less heuristic and more idealistic, in the sense that they are derived from principles of how a distance between shapes should ideally be. This ideal behaviour is typically specified through invariance properties such as the ones described above. In contrast to these approaches for handling misalignment, we propose a full simultaneous statistical model for the fundamental types of variation in misaligned multivariate curves. In particular, we propose to treat amplitude variation and warping variation equally by modeling them as random effects on their respective domains.

Only few works have previously considered the idea of simultaneously modeling amplitude and warping as random effects. An early example of an integrated statistical model that modelled curve shifts as random Gaussian effects is presented in Rønn (2001). The simultaneous inference in the model allows data-driven regularization of the magnitude of the shifts through the estimated variance parameters. The idea has been extended to more general warping functions that are modelled by polynomials (Gervini & Gasser 2005, Rønn & Skovgaard 2009), and lately also to include serially correlated noise within the observations of an individual curve (Raket et al. 2014). In addition to the data-driven regularization of the predicted random effects achieved through estimation of variance parameters, the use of likelihood-based inference naturally relate the discrete observation points and the underlying continuous model. This relation avoids many common issues that arise when developing methods for continuous data in the form of pre-smoothed curves. In particular, the pinching problem, where areas with large deviations are compressed by warping to minimize the integrated residual, does not exist for these methods. Furthermore, the simultaneous modeling of amplitude and warping effects introduces an explicit maximum likelihood criterion for resolving the identifiability problems related to separating warp and amplitude effects (Marron et al. 2015). The maximum-likelihood estimates induce a separation of the two effects, namely the most likely given the variation observed in the data.

A related class of models with random affine transformations of both warping and amplitude variation have become popular in growth curve analysis (Beath 2007, Cole et al. 2010). Hadjipantelis et al. (2014, 2015) provide an extension to this in term of a simultaneous mixed-effects model for the scores in separate functional principal component analyses of the amplitude and the warping effects. The simultaneous model allows not only for cross-correlation within the amplitude and warping scores, but also across these two modes of variation. The estimation procedure used in Hadjipantelis et al. (2014, 2015), however, relies

on a pre-alignment of the curves that separates the vertical and the horizontal variation.

The major contribution of this paper is a new class of multivariate models that both eliminates the need for pre-smoothing and -alignment of samples and also allows for estimation of cross-correlation between the coordinates of the amplitude effect. In the proposed framework, even if we do not assume any cross-correlation of the amplitude effects, the prediction of warping functions will still take the full multivariate sample into account, and the alignment will thus typically be superior to alignment of the individual coordinates.

2 Modeling and inference for misaligned multivariate functional data

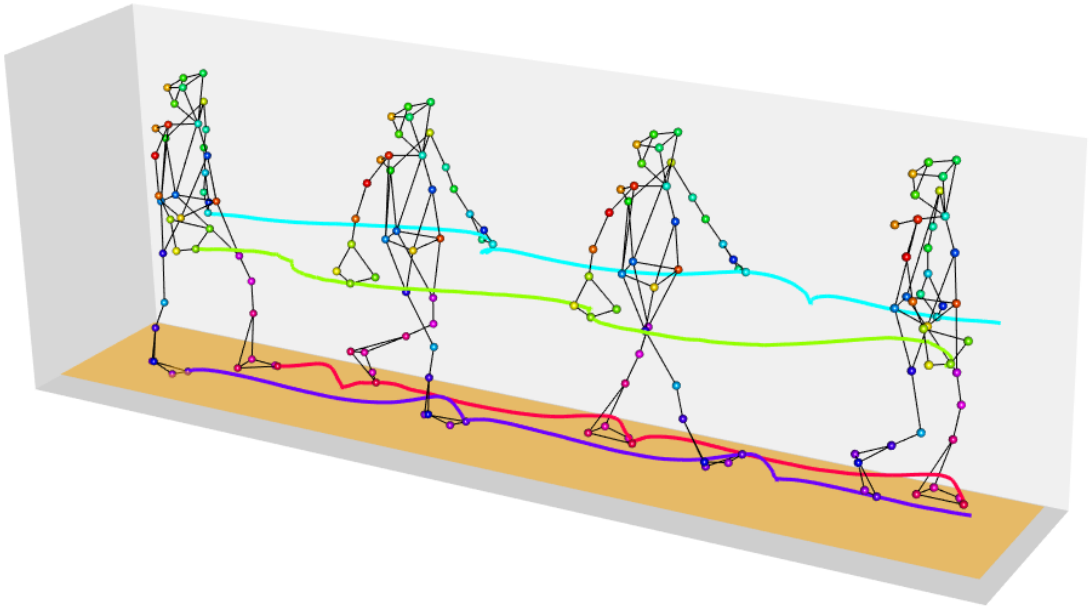


Figure 1: Data from a motion tracking system where the spatial positions of 41 physical markers are tracked in three-dimensions over time. A skeleton model based on the markers is displayed at four temporally equidistant points. The three-dimensional paths of hand and foot markers are displayed.

Consider the multivariate functional observation in Figure 1. The figure displays a walking sequence in three-dimensional space of a person equipped with 41 markers from the *CMU Graphics Lab Motion Capture Database* (n.d.). The observation is a curve in $\mathbb{R}^{123}$ recorded at 301 time points with a total of 36,963 observed values (20 marker positions missing due to occlusion).

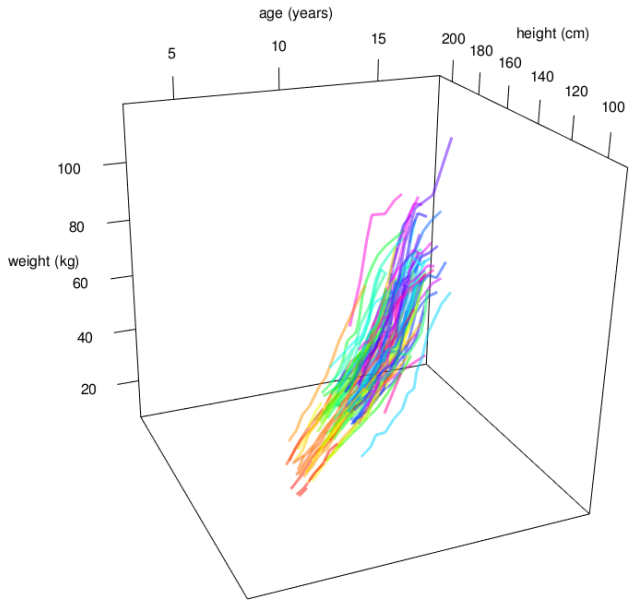


Figure 2: Height and weight measurements over time for 106 healthy boys from the Copenhagen Puberty Study. Each individual curve indicates a subject.

This sample illustrates some of the challenges in analyzing multivariate functional data. Firstly, a repetition of the walking cycle would in all likelihood produce a trajectory that is visually very similar to the sample, but it would differ in two aspects, the movement timing and the movement path would be slightly different. Such differences in timing and path are random perturbations around the person's ideal walking cycle. A natural model for such data is thus a nonlinear mixed-effects model where movement timing is modelled as a random effect whose effect is only observed through the nonlinear transformation of the movement path as a function of time, and the movement path variation is modelled as a stochastic process in $\mathbb{R}^{123}$. However, the very large number of observations in a single functional sample puts strong restrictions on the types of models that can be used. For example, the covariance matrix between the 41 markers at a single time point is 123×123, which in practice makes the problem of estimating a single unstructured covariance (7626 parameters) impossible.

Another example of multivariate functional data is longitudinal measurements of children's height and weight. Figure 2 displays such data from the Copenhagen Puberty Study (Aksglaede et al. 2009, Sørensen et al. 2010). The data reflects the fact that height and weight are generally increasing functions during childhood and adolescence. Again, there will be a nonlinear timing effect; observed age is a proxy for a biological or developmental age process of the child, and there will be systematic differences in observation values; taller and heavier children tend to stay taller and heavier than their peers. For height and weight data, one

would typically have few observations per child, but the possibility of many children. Thus, the cross-covariance at a given time point could easily be estimated, and one could have a natural interest in inferring possible changes in the correlation between height and weight over time.

The two above examples illustrate that the challenges of multivariate functional data can be very different. In the following we will introduce a class of models to analyze functional data containing both warp and amplitude variation. To make the model sufficiently flexible, we will introduce generic models for random warping functions and dynamic cross-correlation structures that can approximate arbitrary structures, and whose resolution of approximation can be coarsened by reducing the number of free parameters.

2.1 Statistical model

We consider a set of N discrete observations of q-dimensional curves $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N \colon [0,1] \to \mathbb{R}^q$ from J subjects. The curves are assumed to be generated according to the following model

$$\boldsymbol{y}_n(t) = \boldsymbol{\theta}_{f(n)}(v_n(t)) + \boldsymbol{x}_n(t), \quad n = 1, \ldots, N. \tag{1}$$

Here $f \colon \{1, \ldots, N\} \to \{1, \ldots, J\}$ is a known function that maps sample number to subject number. The unknown fixed effects are subject specific mean value functions $\boldsymbol{\theta}_j \colon [0,1] \to \mathbb{R}^q$ for $j = 1, \ldots, J$ that are modelled using a spline basis assumed to be continuously differentiable. Typical choices are B-spline bases and Fourier bases. The phase variation is modelled by random warping functions $v_n = v(\cdot, \boldsymbol{w}_n) \colon [0,1] \to [0,1]$, which are parametrized by independent latent zero-mean Gaussian variables $\boldsymbol{w}_n \in \mathbb{R}^{m_{\boldsymbol{w}}}$ for $n = 1, \ldots, N$ with a common covariance matrix $\sigma^2 C$. Here $v \colon [0,1] \times \mathbb{R}^{m_{\boldsymbol{w}}} \to [0,1]$ is a pre-specified function, that is assumed to be continuously differentiable in its second argument, and $m_{\boldsymbol{w}} \in \mathbb{N}$ is the dimension of the latent variable. The amplitude variation is modelled by independent zero-mean Gaussian processes $\boldsymbol{x}_n \colon [0,1] \to \mathbb{R}^q$ for $n = 1, \ldots, N$ with a common covariance function $\sigma^2 \mathcal{S}$. The unknown variance parameters are thus a scalar $\sigma^2 > 0$, a positive definite matrix $C \in \mathbb{R}^{m_{\boldsymbol{w}} \times m_{\boldsymbol{w}}}$ and a positive definite function $\mathcal{S} \colon [0,1] \times [0,1] \to \mathbb{R}^{q \times q}$. In sections 2.2 and 2.3 we discuss models for the warping functions and the cross-covariance of the amplitude variation that are highly expressive, while the number of parameters to be estimated is kept at a moderate level.

We assume that the nth curve is observed at $m_n \in \mathbb{N}$ prefixed time points t_{nk}, which neither need to be equally spaced in time nor to be shared by the N samples. Stacking the m_n temporally discrete observations into a vector we have

$$\vec{\boldsymbol{y}}_n = \{\boldsymbol{y}_n(t_{nk}) + \boldsymbol{\varepsilon}_{nk}\}_{k=1}^{m_n} \in \mathbb{R}^{q m_n}, \quad n = 1, \ldots, N, \tag{2}$$

where the observation noise is given by independent zero-mean Gaussian variables $\boldsymbol{\varepsilon}_{nk} \in \mathbb{R}^q$ with a common variance $\sigma^2 \boldsymbol{I}_q$. Here $\boldsymbol{I}_q \in \mathbb{R}^{q \times q}$ denotes the identity matrix.

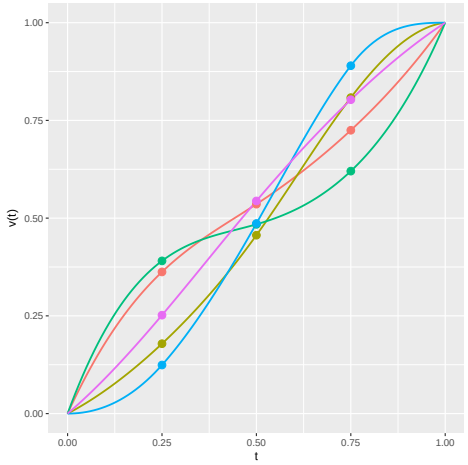


Figure 3: Simulated warping functions with the covariance given by (4). The warp values at the three interior anchor points are marked by points.

The major structural difference of model (1) compared to conventional functional mixed-effects models (Guo 2002) is the inclusion of a warping effect. When compared to conventional methods for curve alignment, the proposed model differs by having a random amplitude effect, by modeling warping functions as random effects, and by handling all effects simultaneously.

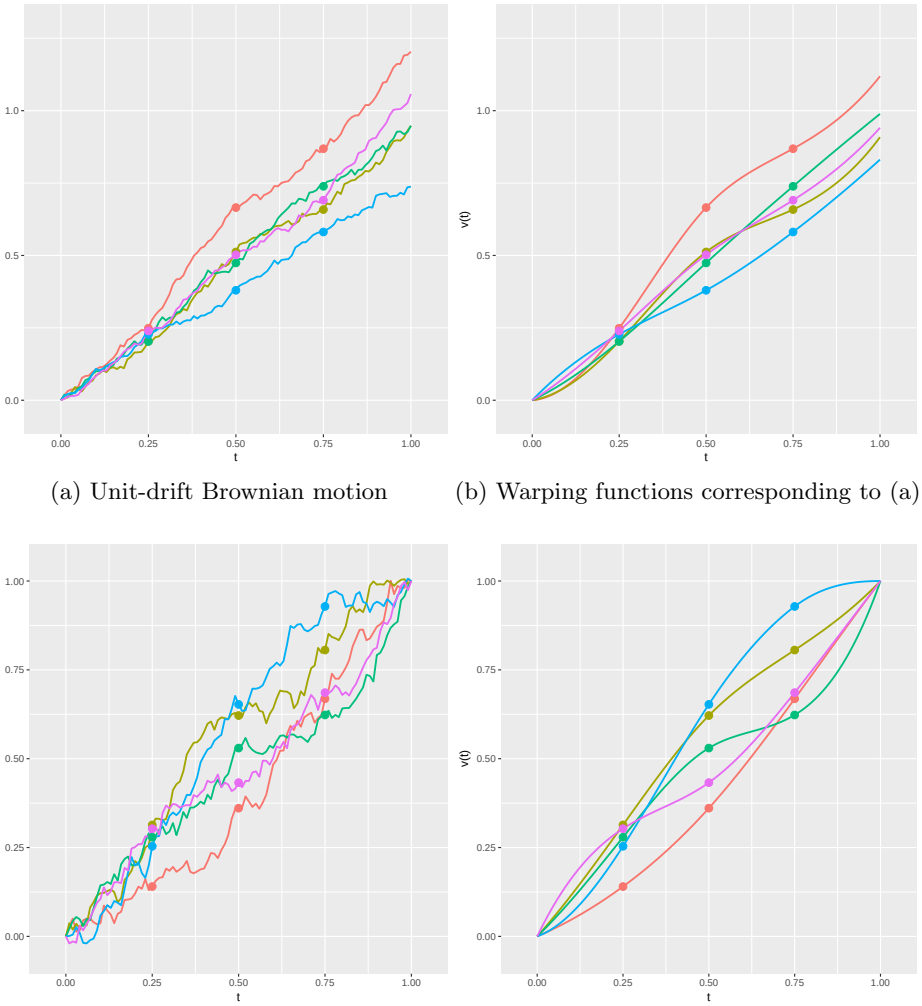
2.2 Modeling warping functions

The success of the model relies on its ability to approximate the realizations of the true warping functions. To accomplish this, the warping functions v_n must be sufficiently versatile and able to approximate a large array of different warps. We achieve this by modeling warping functions as the identity mapping plus a deformation modelled by interpolating latent warp variables $\boldsymbol{w}_n \in \mathbb{R}^{m_{\boldsymbol{w}}}$ at pre-specified (e.g. equidistant) anchor points t_k for $k = 1, \ldots, m_{\boldsymbol{w}}$

$$v_n(t) = v(t, \boldsymbol{w}_n) = t + \mathcal{E}_{\boldsymbol{w}_n}(t), \tag{3}$$

where the interpolation function $\mathcal{E}_{\boldsymbol{w}}$ can, for example, be a linear or a cubic spline.

The behavior of the predicted warping functions will be determined by the combination of interpolation method (and corresponding boundary conditions) and the estimated covariance of the latent variables $\boldsymbol{w}_n$. Throughout this paper we will use cubic spline interpolation of the latent variables. If we think of the parametrization of the nth sample, $v_n(t)$, as the internal time of the sample, it is often natural to assume that the internal time is always moving forward. To ensure this, we will predict the latent variables $\boldsymbol{w}_n$ using constrained optimization such that the sequence will be increasing along the corresponding anchor points. But for cubic interpolation, a sequence of increasing values at the interpolation points is not sufficient to ensure a monotone interpolation function. To force increasing warping functions we will use



(a) Unit-drift Brownian motion (b) Warping functions corresponding to (a)

(c) Unit-drift Brownian bridge (d) Warping functions corresponding to (c)

Figure 4: Constructions of warping functions from stochastic processes with parametric covariances. (a) simulated trajectories of a unit-drift Brownian motion with scale 0.1, (b) warping functions using a unit-drift Brownian motion model with $m_{\boldsymbol{w}} = 3$ interior equidistant anchor points, fixed interpolation at the left boundary and extrapolation of the rightmost deviation at the right endpoint, (c) simulated trajectories of a unit-drift Brownian bridge with scale 0.2, (d) warping functions using the unit-drift Brownian bridge model with $m_{\boldsymbol{w}} = 3$ interior equidistant anchor points and fixed interpolation at the boundary.

the Hyman filter (Hyman 1983) to ensure that the entire warping function is increasing. For some types of data, it may be meaningful to have warps that can go backwards in time, or it may be useful to include this option to account for uncertainty in the model if the observed signals contain features where the matching is highly ambiguous. Such types of warp models will not be considered in this paper.

The covariance matrix of the latent variables will determine the regularity of the predicted warping functions. When the number of latent variables $m_{\boldsymbol{w}}$ is small compared to the number of functional samples N and the number of sampling points $m_1, \ldots, m_N$ for the functional samples, one can assume an unstructured covariance and estimate the corresponding $(m_{\boldsymbol{w}}^2 + m_{\boldsymbol{w}})/2$ variance parameters. If the structure of the warping functions are of key interest, one may be able to study the underlying mechanism by estimating an unstructured covariance matrix. Consider for example the simulated warping functions shown in Figure 3. These warping functions use the increasing cubic spline construction detailed above with $m_{\boldsymbol{w}} = 3$ interior equidistant anchor points, fixed boundary points and covariance matrix

$$\begin{pmatrix} 0.005 & 0 & -0.004 \\ 0 & 0.001 & 0 \\ -0.004 & 0 & 0.005 \end{pmatrix}. \tag{4}$$

The interpretation of the strong negative covariance between first and third anchor point suggest a burnout type of process where samples that are ahead initially slow down toward the end and vice versa. The low variance of the middle anchor point suggest that the individual samples are largely synchronized around this time.

In many cases, one can choose a specific interpolation method and specify a reasonable parametric covariance for the latent variables based on properties of the data. It is, for example, often natural to think of warping processes as accumulations of small errors causing desynchronization of the set of observed trajectories that all started in the same state. Thinking of Gaussian processes, Brownian motion with linear unit drift would offer a simple model for phenomena where errors are accumulating and increasing the desynchronization of samples over time. Simulations of unit-drift Brownian motions are shown in Figure 4 (a) and the corresponding simulations of warping functions from $m_{\boldsymbol{w}} = 3$ interior equidistant anchor points, fixed left boundary point and linear extrapolation of the deviation of the rightmost anchor point at the right boundary point are shown in Figure 4 (b).

Suppose we are analyzing longitudinal data of children's heights where we could think of the warping function as the developmental (height) age of the child. At conception (approximately -9 months of age), where the child is merely a fertilized egg, all children are the size of a grain of sand and their developmental ages are synchronized. As the children become older the desynchronization of their developmental ages increases. This can, for example, be seen by the vast variation between the age of onset of puberty. The unit-drift Brownian motion warp model seems like a very suitable model for this desynchronization.

Other types of data may give rise to other models. Consider an experiment that records repetitions of a walking sequence such as the data in Figure 1, and assume that all sequences start from the same pose and end after two completed gait cycles. For such data, the desynchronization is not increasing over time since beginning and end poses are synchronized, but we would expect maximum desynchronization around the middle of the gait cycle window. In this setting, a more suitable model would be a unit-drift Brownian bridge as illustrated in Figure 4 (c) and (d).

Like other hyperparameters, the number of anchor points is a choice of modelling. However, a low number of anchor points (e.g. 3-5) will generate a class of warp functions that is sufficiently flexible for many applications; we used $m_w = 3$ in all applications presented in this paper. If, however, local variation is very strong and complex and the observed functional samples carry sufficiently clear information about the systematic shapes to recover such complex warps, a higher number of anchor points should be used.

2.3 Dynamic covariance structures

In the previous section we modelled the covariance structure of smooth warping functions and saw how one could use domain-specific knowledge of the data to choose models with few parameters. Even though the nature of the additive amplitude variation components $\boldsymbol{x}_n$ from model (1) is different, we can extend these ideas to construct parametric, low-dimensional cross-covariance structures that are sufficiently expressive to model a wide array of cross-covariance structures over time.

Proposition 1. Let $f : [0,1] \times [0,1] \to \mathbb{R}_+$ be a positive definite function on the temporal domain $[0,1]$. Let $0 = t_1 < \cdots < t_\ell = 1$ be anchor points, let $A_1, \ldots, A_\ell \in \mathbb{R}^{q\times q}$ be a set of symmetric positive definite matrices, and for each $t \in [0,1]$ define $B_t \in \mathbb{R}^{q\times q}$ as the unique positive definite matrix satisfying

$$B_t^\top B_t = \frac{t_{k+1} - t}{t_{k+1} - t_k} A_k + \frac{t - t_k}{t_{k+1} - t_k} A_{k+1} \quad \text{for } t \in [t_k, t_{k+1}]. \tag{5}$$

For all $s, t \in [0,1]$, define $K(s,t) = f(s,t) B_s^\top B_t \in \mathbb{R}^{q\times q}$. Then the function $K : [0,1]\times[0,1] \to \mathbb{R}^{q\times q}$ is positive definite.

Proof. First we remark that since the space of positive definite matrices is a convex cone, the linear interpolation $B_t^\top B_t$ is also positive definite, and we may take B_t as the positive square root. To prove that $K : [0,1] \times [0,1] \to \mathbb{R}^{q\times q}$ is positive definite it suffices to show that the associated finite dimensional marginal matrices are positive definite. Thus, given $s_1, \ldots, s_m \in [0,1]$ we let the block matrix $\mathbf{V} \in \mathbb{R}^{qm\times qm}$ be defined by

$$\mathbf{V} = \begin{pmatrix} B_{s_1}^\top f(s_1,s_1) B_{s_1} & B_{s_1}^\top f(s_1,s_2) B_{s_2} & \cdots & B_{s_1}^\top f(s_1,s_m) B_{s_m} \\ B_{s_2}^\top f(s_2,s_1) B_{s_1} & B_{s_2}^\top f(s_2,s_2) B_{s_2} & \cdots & \\ \vdots & & \ddots & \\ B_{s_m}^\top f(s_m,s_1) B_{s_1} & B_{s_m}^\top f(s_m,s_2) B_{s_2} & \cdots & B_{s_m}^\top f(s_m,s_m) B_{s_m} \end{pmatrix}. \tag{6}$$

By straightforward calculations we have $\mathbf{V} = \mathbf{B}^\top(\mathbf{F} \otimes \mathbf{I}_q)\mathbf{B}$, where $\mathbf{B} \in \mathbb{R}^{qm \times qm}$ is the block-diagonal matrix of $\{B_{s_1}, \ldots, B_{s_m}\}$ and

$$\mathbf{F} = \begin{pmatrix} f(s_1, s_1) & \cdots & f(s_1, s_m) \\ \vdots & \ddots & \vdots \\ f(s_m, s_1) & \cdots & f(s_m, s_m) \end{pmatrix}. \tag{7}$$

For $z \in \mathbb{R}^{qm} \setminus \{0\}$ we must show that $z^\top \mathbf{V} z > 0$. Setting $u = \mathbf{B}z \neq 0$ and using that $\mathbf{F}$ is positive definite by assumption we have $z^\top \mathbf{V} z = u^\top(\mathbf{F} \otimes \mathbf{I}_q)u > 0$. □

The above proposition gives a general framework for constructing dynamical covariance functions, and it is simple to construct parametric models that allow for estimation of time-varying cross-correlations in a statistical setting. In the statement of the proposition we assumed a common marginal covariance function f along all coordinates. The idea of modeling a cross-covariance structure by linearly interpolating cross-covariances at specific points seamlessly extends to multivariate diagonal covariance functions (i.e. no cross-covariances), such that the individual coordinates of the functional samples may be modelled using different types covariance functions or different parameters.

3 Estimation

Direct likelihood inference in the model (1) is not feasible as the model contains nonlinear latent variables in combination with possible very large data sizes. Instead we propose a maximum-likelihood estimation procedure based on iterative local linearization (Lindstrom & Bates 1990). The procedure is a multivariate extension of the estimation procedure described in Raket et al. (2014), however with an improved estimation of fixed effects.

The estimation procedure consists of alternating steps of (1); estimating fixed effects (i.e. spline coefficents) and predicting the most likely warp variables given the data and current parameter estimates, (2); estimating variance parameters from the locally linearized likelihood function around the maximum a posteriori predictions $\boldsymbol{w}_1^0, \ldots, \boldsymbol{w}_N^0$ of the warp variables. The linearization in the latent Gaussian warp parameters $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_N$ means that we approximate the nonlinearly transformed probability density by the density of a linear combination of multivariate Gaussian variables. The estimation procedure is thus a Laplace approximation of the likelihood, and the quality of the approximation is approximately second order (Wolfinger 1993).

Predicting warps In the first step of the estimation procedure we want to predict the most likely warps from model (1) given the current parameter estimates. The negative log posterior for a single functional sample is proportional to

$$(\vec{\boldsymbol{\gamma}}_{\boldsymbol{w}_n} - \vec{\boldsymbol{y}}_n)^\top (\boldsymbol{I}_{qm_n} + S_n)^{-1} (\vec{\boldsymbol{\gamma}}_{\boldsymbol{w}_n} - \vec{\boldsymbol{y}}_n) + \boldsymbol{w}_n^\top C^{-1} \boldsymbol{w}_n \tag{8}$$

where $\vec{\boldsymbol{\gamma}}_{\boldsymbol{w}_n} \in \mathbb{R}^{qm_n}$ is the stacked vector $\{\boldsymbol{\theta}_{f(n)}(v(t_{nk}, \boldsymbol{w}_n))\}_{k=1}^{m_n}$ and $S_n \in \mathbb{R}^{qm_n \times qm_n}$ is the amplitude covariance $\{\mathcal{S}(t_{nj}, t_{nk})\}_{j,k=1,\dots,m_n}$ at the sample points. The issue of predicting warps is thus a nonlinear least squares problem that can be solved by conventional methods.

Estimating variance parameters Since $\boldsymbol{\theta}_{f(n)} \circ v(t_{nk}, \cdot)$ are smooth functions for all $n = 1, \dots, N$, $k = 1, \dots, m_n$ we can linearize model (1) around a given prediction $\boldsymbol{w}_n^0$ using the first-order Taylor expansion. The linearization is given by

$$\boldsymbol{\theta}_{f(n)}(v(t_{nk}, \boldsymbol{w}_n)) \approx \boldsymbol{\theta}_{f(n)}(v(t_{nk}, \boldsymbol{w}_n^0)) + \partial_t \boldsymbol{\theta}_{f(n)}(v(t_{nk}, \boldsymbol{w}_n^0))(\nabla_w v(t_{nk}, \boldsymbol{w}_n^0))^\top (\boldsymbol{w}_n - \boldsymbol{w}_n^0). \quad (9)$$

For the discrete observation of the nth curve this gives a linearization of model (1) as a vectorized linear mixed-effects model on the form

$$\vec{\boldsymbol{y}}_n \approx \vec{\boldsymbol{\gamma}}_{\boldsymbol{w}_n^0} + Z_n(\boldsymbol{w}_n - \boldsymbol{w}_n^0) + \vec{\boldsymbol{x}}_n + \vec{\boldsymbol{\varepsilon}}_n, \quad n = 1, \dots, N, \quad (10)$$

where $\vec{\boldsymbol{\gamma}}_{\boldsymbol{w}_n^0}, \vec{\boldsymbol{x}}_n, \vec{\boldsymbol{\varepsilon}}_n \in \mathbb{R}^{qm_n}$ are the stacked vectors

$$\vec{\boldsymbol{\gamma}}_{\boldsymbol{w}_n^0} = \{\boldsymbol{\theta}_{f(n)}(v(t_{nk}, \boldsymbol{w}_n^0))\}_{k=1}^{m_n}, \qquad \vec{\boldsymbol{x}}_n = \{\boldsymbol{x}_n(t_{nk})\}_{k=1}^{m_n}, \qquad \vec{\boldsymbol{\varepsilon}}_n = \{\boldsymbol{\varepsilon}_{nk}\}_{k=1}^{m_n},$$

and $Z_n \in \mathbb{R}^{qm_n \times m_{\boldsymbol{w}}}$ is the row-wise stacked matrix

$$Z_n = \{\partial_t \boldsymbol{\theta}_{f(n)}(v(t_{nk}, \boldsymbol{w}_n^0)) \nabla_{\boldsymbol{w}} v(t_{nk}, \boldsymbol{w}_n^0)\}_{k=1}^{m_n}.$$

In the approximative model (10) twice the negative profile log-likelihood $l(\sigma^2, C, \mathcal{S})$ for the variance parameters is given by

$$\sum_{n=1}^{N} \left(qm_n \log \sigma^2 + \log \det V_n + \sigma^{-2} (\vec{\boldsymbol{y}}_n - \vec{\boldsymbol{\gamma}}_{\boldsymbol{w}_n^0} + Z_n \boldsymbol{w}_n^0)^\top V_n^{-1} (\vec{\boldsymbol{y}}_n - \vec{\boldsymbol{\gamma}}_{\boldsymbol{w}_n^0} + Z_n \boldsymbol{w}_n^0) \right), \quad (11)$$

where $V_n = Z_n C Z_n^\top + S_n + \boldsymbol{I}_{qm_n}$ with $S_n = \{\mathcal{S}(t_{nj}, t_{nk})\}_{j,k=1,\dots,m_n}$. In particular, the profile maximum-likelihood estimate for σ^2 is given by

$$\hat{\sigma}^2 = \frac{1}{qm} \sum_{n=1}^{N} (\vec{\boldsymbol{y}}_n - \vec{\boldsymbol{\gamma}}_{\boldsymbol{w}_n^0} + Z_n \boldsymbol{w}_n^0)^\top V_n^{-1} (\vec{\boldsymbol{y}}_n - \vec{\boldsymbol{\gamma}}_{\boldsymbol{w}_n^0} + Z_n \boldsymbol{w}_n^0)$$

where $m = \sum_{n=1}^{N} m_n$ is the total number of observations. Estimation of the variance parameters C and $\mathcal{S}$ related to the warping and amplitude effects is done using the profile likelihood $l(\hat{\sigma}^2, C, \mathcal{S})$.

Estimating fixed effects As the fixed effects are given by spline bases, estimation of these can be handled within the framework of linear Gaussian models, remembering that basis functions should be evaluated at warped time points $v_n(t_{nk})$. Since $v_n(t_{nk}) = v(t_{nk}, \boldsymbol{w}_n)$ changes with $\boldsymbol{w}_n$, we are required to recalculate the spline basis matrix for each new prediction of $\boldsymbol{w}_n$. This estimation improves that of Raket et al. (2014), which used a point-wise estimation based on the inverse warp that ignored the amplitude variance of the curves.

There is no closed-form expression for the maximum-likelihood estimator of the fixed effects in the linearized model, since spline coefficients also enter the variance terms through the matrices Z_n, as can be seen in equation (11). However, by construction Z_n is linear in the spline coefficients so estimation can be done using an EM algorithm. The details of these calculations can be found in the supplementary material.

In practice, the estimation in the linearized model can be approximated by estimating from the posterior likelihood (8) which gives a computationally efficient closed-form solution. The difference between these two approaches is that the EM algorithm takes the uncertainty in prediction of $\boldsymbol{w}_n$ into account and is guaranteed to decrease the linearized likelihood (11). However, for a moderate number of warp parameters, there should only be a small conditional variance on $\boldsymbol{w}_n$.

In the data applications presented in the following sections, we estimated fixed effects from the posterior likelihood. In the last application on hand movements, these posterior likelihood estimates were used to initialized the likelihood optimization which were subsequently fine-tuned by the EM algorithm with a single update per warp prediction. This was done to evaluate if improved likelihood estimates could be obtained, but the EM algorithm offered only a very slight improvement in linearized likelihood.

4 Applications

4.1 Motion capture data

Data and model Data consists of four 12-dimensional functional objects. The curves consist of a total of 1284 temporal observations in $\mathbb{R}^{12}$. As can be seen in Figure 5, the trajectories start and end at different places during the gait cycle. To handle this structure, time was scaled to the interval $[0, 1]$ such that all samples began at 0.1, and such that the temporally longest trajectory ended at 0.9. We included random shift parameters s_n in our warping functions to model these different temporal onsets of the gait cycle. The shifts s_n were modelled as Gaussian random variables. The full model is

$$\boldsymbol{y}_n(t) = \boldsymbol{\theta}(v(t, \boldsymbol{w}_n, s_n)) + \boldsymbol{x}_n(t) \tag{12}$$

where $\boldsymbol{\theta} : [0, 1] \to \mathbb{R}^{12}$ is the mean curve for the observations (modelled using a 3-dimensional B-spline basis with 30 interior anchor points) and the warping function v is given by

$$v(t, \boldsymbol{w}_n, s_n) = t + s_n + \mathcal{E}_{\boldsymbol{w}_n}(t)$$

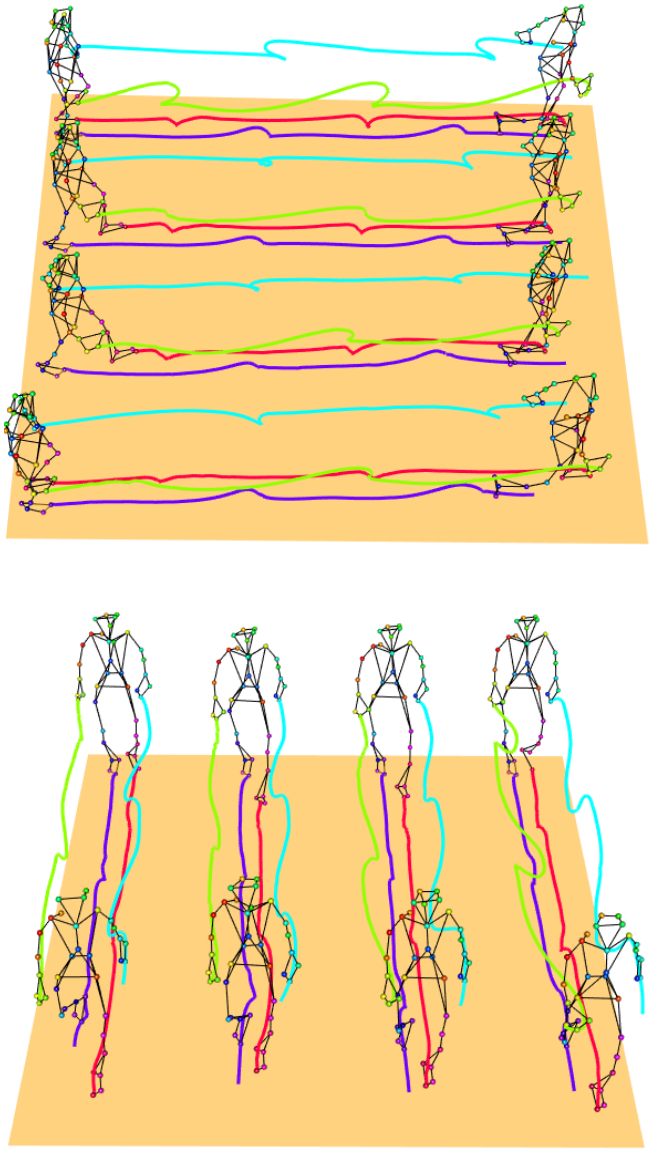


Figure 5: Side and frontal view of the motion trajectories of four walking sequences performed by the same participant.

where $\mathcal{E}_{\boldsymbol{w}_n}$ is an increasing cubic spline interpolation (Hyman filtered) of $\boldsymbol{w}_n$ at $m_{\boldsymbol{w}} = 3$ equidistant anchor points. No subject-specific effects were included as all responses were recorded from the same individual. The amplitude effect $\boldsymbol{x}_n$ was modelled as a Gaussian process with a Matérn covariance $f_{\text{Matérn}(2,\kappa)}(s,t)$ with second order smoothness, assuming independent coordinates and a common range parameter κ (see equation (16) in the supplement). We assumed different scaling parameters for each of the 12 coordinates of $\boldsymbol{x}_n$. Since the data is roughly cut to include two gait cycles, one would expect high synchronization of start and end poses in percentual time when corrected for the different onsets. Therefore, latent variables $\boldsymbol{w}_n$ were modelled as discretely observed Brownian bridges with a single scale parameter.

Results The predicted warping functions are shown in Figure 6, and the corresponding aligned samples are shown in Figure 7. The samples are nicely aligned, in particular, the regular elevation profiles of the left and right feet seems very well aligned. The remaining signals have their key-features aligned, with the residual variation evenly spread out across the coordinates. This is a feature of the simultaneous multivariate fitting, where the best alignment given the variation in the different coordinates is found. Individual alignment of the coordinates would produce warping functions that overfitted the individual aspects of the movement. In Figure 8, we have displayed the estimated mean trajectories $\boldsymbol{\theta}$ and illustrated the uncertainty after alignment by 95% prediction ellipsoids for the amplitude effect $\boldsymbol{x}_n$.

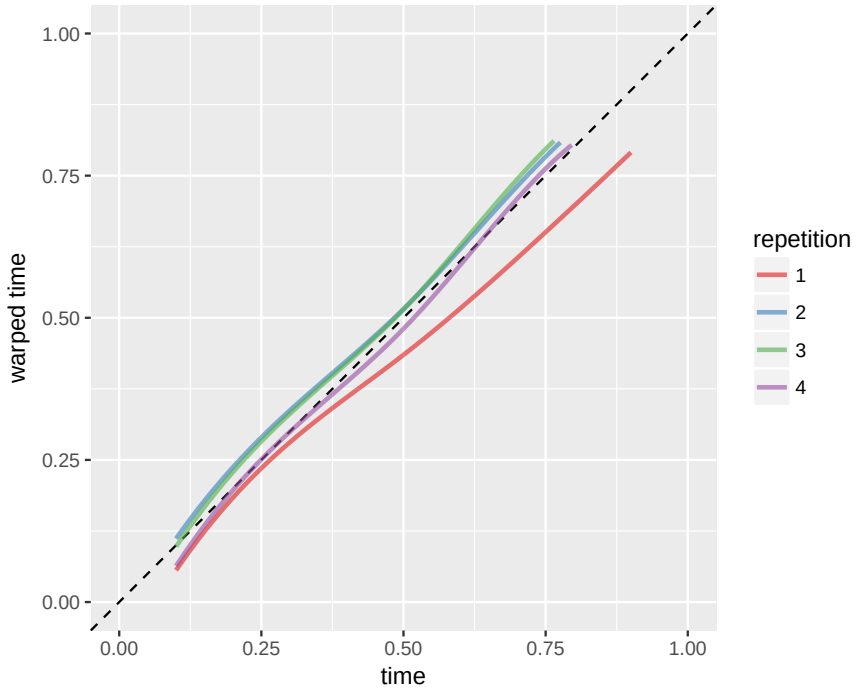


Figure 6: Predicted warping functions for the motion capture data

4.2 Height and weight data

Consider the height and weight measurements from the Copenhagen Puberty Study (Aksglaede et al. 2009, Sørensen et al. 2010) shown in Figure 2. The data contains 960 pairs of height and weight measurements for 106 healthy Danish boys. The individual amplitude effects in the data set are clearly visible in the form of systematic deviations from the mean. The data also contain warping variation in the sense that age is a proxy for developmental age; each boy has his own internal clock that determines, for example, the onset of puberty. Alignment for this warping effect would then align the pubertal growth spurts visible as steep height increase in the individual boys occuring in the period 11 to 14 years.

Modeling While height is a naturally increasing function of age, weight is not necessarily. However, looking at the 2014 Danish weight reference Tinggaard et al. (2014), we see a convex increase in the cross-sectional mean weight curve in the relevant age interval. Based on this, we modelled $\boldsymbol{\theta}$ using an increasing spline (integrated quadratic B-splines) basis with 20 equidistant internal knots in the age interval $[5, 17]$ in both dimensions. The warping functions (3) were modelled as increasing cubic (Hyman filtered) splines with $m_{\boldsymbol{w}} = 3$ equidistant internal anchor points in the age interval $[5, 20]$ and extrapolation at the right boundary point as in Figure 4(b). The latent variables $\boldsymbol{w}_n$ were modelled as discretely observed Brownian

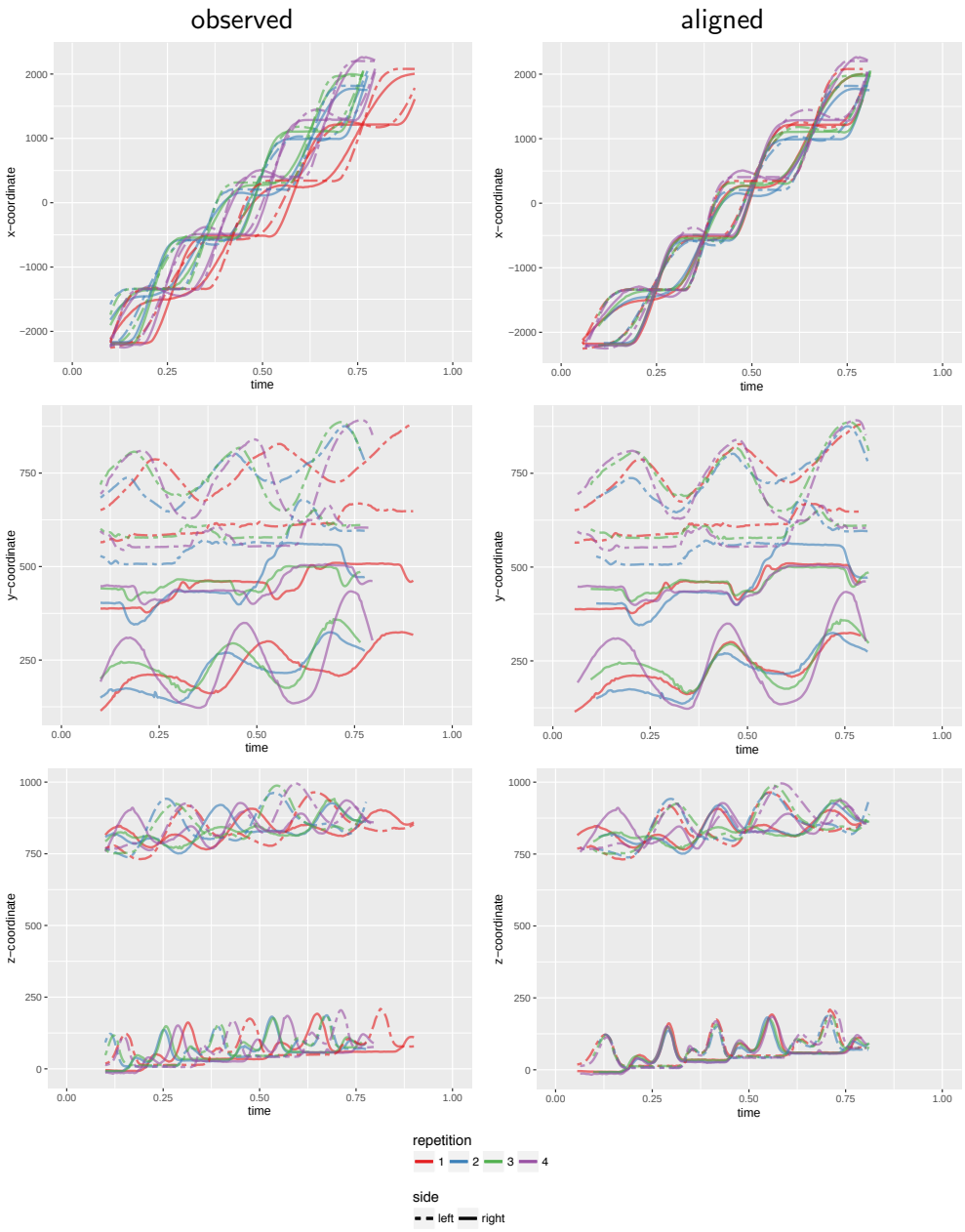


Figure 7: Observed and aligned curves from the motion capture data. Data values are the raw values from the tracking system.

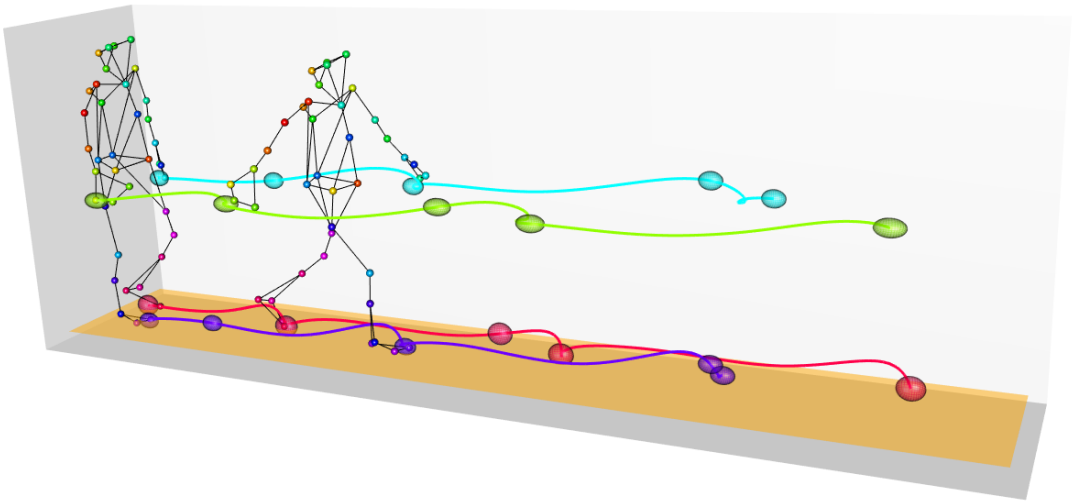


Figure 8: Estimated mean trajectories with five temporally equidistant ellipsoids indicating 95% (marginal) confidence areas. Two of the intermediate body poses of the fourth sample have been added as a reference.

motions with a single scale parameter. The temporally increasing variance of the Brownian motion seems as a good model for developmental age where one would expect high initial synchronization, and up to several years desynchronization at the onset of puberty.

To model the amplitude variation, we used a dynamic cross-covariance with equidistant knots at $\{5, 10, 15, 20\}$ years as described in Proposition 1, that is,

$$\mathcal{S}(s,t) = f_{\text{Matérn}(2,\kappa)}(s,t)B_s^\top B_t.$$

The temporal covariance structure $f_{\text{Matérn}(2,\kappa)}(s,t)$ is the Matérn covariance function with fixed smoothness parameter $\alpha = 2$ and unknown range parameter κ, see equation (16) in the supplement. This implies twice differentiable sample paths of $\boldsymbol{x}_i$, which is a reasonable assumption given the nature of the data. Furthermore, since we expected heterogeneous variances of the measurement error $\boldsymbol{\varepsilon}_{nk}$ on height and weight in equation (2), we extended the model with a parameter $\rho > 0$ such that

$$\text{Var}(\boldsymbol{\varepsilon}_{nk}) = \sigma^2 \begin{pmatrix} 1 & 0 \\ 0 & \rho \end{pmatrix}.$$

This gives a total of 14 parameters describing the cross-covariance model.

Results The aligned samples and estimated means are displayed in the right-side panels of Figure 9, and the corresponding predicted warping functions can be found in Figure 10. We see that the individual growth curves are now aligned more tightly than before, in particular the pubertal height spurts seem to be well aligned. Although the shapes of the curves are well aligned, the model still allowed for considerable amplitude variation to be left after warping. This is as it should be; for increasing curves such as these a perfect fit could be achieved by warping, but the result would be meaningless and indicate that developmental age could be perfectly determined from a single measurement of a child's height. Given the proposed model-based separation of amplitude and warping effects induced by the maximum likelihood estimates, the information contained in a child's longitudinal data about the child's developmental age can be quantified through the posterior distribution of the warping effects.

The estimated covariance structure is shown in Figure 11. As one would expect, height and weight variances increase with age. The covariance increases at a slower rate and has a slight decrease after 15 years, giving a correlation of 0.42 at 16.5 years.

4.3 Arm movement data

Our third example is an analysis of human arm movements in obstacle avoidance tasks. Hand-movement paths in two experimental conditions are displayed in Figure 12. In each experimental condition, a wooden cylindrical object (pink) located at a starting position (green cylinder) was to be moved 60 centimeters forward and placed on a target cylinder. Between

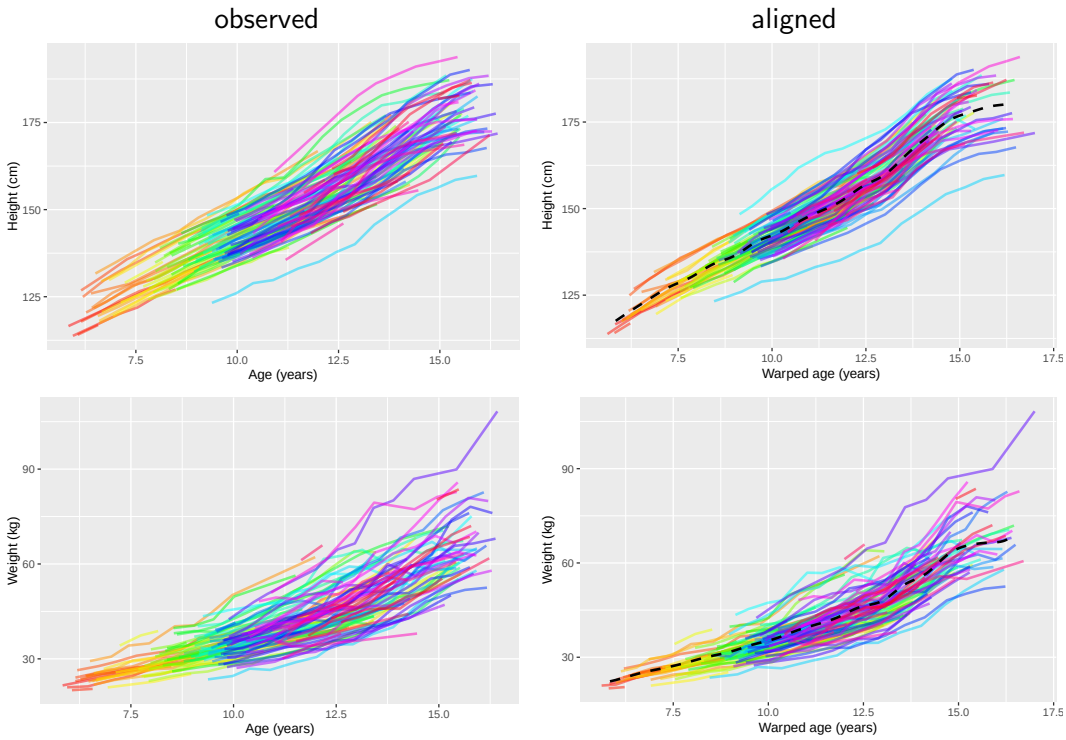


Figure 9: Observed and aligned height and weight curves from the Copenhagen Puberty Study. The estimated template curves are displayed as dashed black lines.

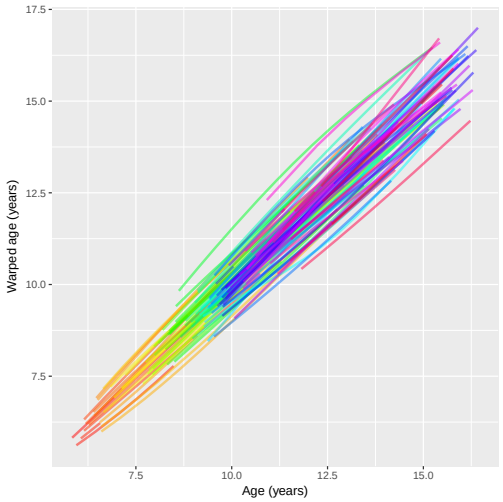


Figure 10: Predicted warping functions corresponding to the data in Figure 9.

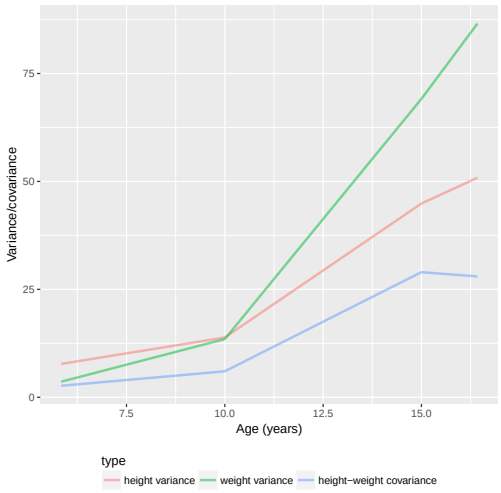


Figure 11: Estimated marginal variances and cross-covariance functions of age for the height and weight data in Figure 9. The marginal variances also include the error variance.

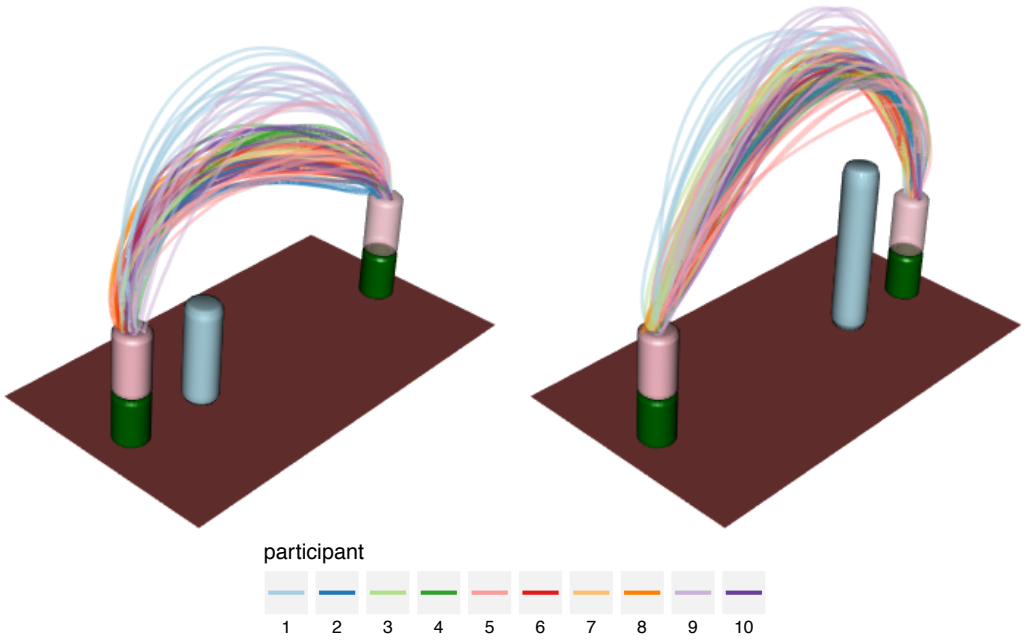


Figure 12: Recorded movement paths in experiment with small obstacle 15 cm from starting position (left) and tall obstacle 45 cm from starting position (right).

the starting and target positions, a cylindrical obstacle was placed. The obstacle height (small, medium, tall) and obstacle position (five equidistant positions between starting and target positions) varied with experimental condition. A total of 15 obstacle avoidance conditions were performed plus a control condition with no obstacle. Ten right-handed participants performed ten repetitions of each experimental condition, and the spatial position of the hand was recorded at a sampling rate of 110 Hz. The data set thus consists of 1600 functional samples with a total of $m = 175,535$ three-dimensional sampling points giving a total sample size of $526,605$ observations. The present data set is described in detail in Grimme (2014), and the experiment is a refined version of the experiment described in Grimme et al. (2012). The data set is available through a public repository.[1]

Data processing and modeling We analyzed the data separately for the 16 experimental conditions. Following the convention for modeling human motor control data, time was modelled as percentual time rather than observed time. This means that all movement time intervals were scaled to $[0, 1]$, such that 0 corresponds to the onset of the movement and 1 corresponds to the end of the movement. We used model (1) to model the data separately for the 16 different experimental conditions. The mean path $\boldsymbol{\theta}_j$ for the jth participants was modelled in a cubic B-spline basis with 21 interior knots. We modelled the warping functions (3) as increasing cubic spline interpolations (Hyman filtered) with $m_{\boldsymbol{w}} = 3$ equidistant anchor points. The choice of three knots was evaluated, and found optimal, in terms of the cross-validation set-up described in the classification study below. The latent variables $\boldsymbol{w}_n$ were modelled as discretely observed Brownian bridges with a single scale parameter, because of the fixed endpoints of the data.

The amplitude variation was modelled using a dynamic cross-correlation model with knots at $\{0, 0.4, 0.6, 1\}$ as described in Proposition 1, that is,

$$\mathcal{S}(s,t) = f_{\text{mixture}(a)}(s,t) f_{\text{Matérn}(\alpha,\kappa)}(s,t) B_s^\top B_t.$$

The temporal covariance structure is given as a combination of stationary and bridge Matérn serial correlation with mixture parameter a, smoothness parameter α, and range parameter κ. The details of this covariance structure are described in equations (15) and (16) in the supplement. This dynamic cross-correlation structure has 27 free parameters.

The knot positions $\{0, 0.4, 0.6, 1\}$ were chosen such that we were able to model a change in cross-correlation structure around the middle of the movement in percentual time, in particular the change that happens when the movement progresses from lift to descend. The concept of isochrony (Grimme et al. 2012) suggests that the times where the peak heights are reached are largely invariant to obstacle height and placement, and for the given data the peak heights generally occur for $t \in (0.4, 0.6)$, see for example Figure 13.

The left column of Figure 13 displays the observed x-, y- and z-coordinates in a single experimental condition as functions of percentual time. The right column displays the coordinates in

[1] `https://github.com/larslau/Bochum_movement_data`

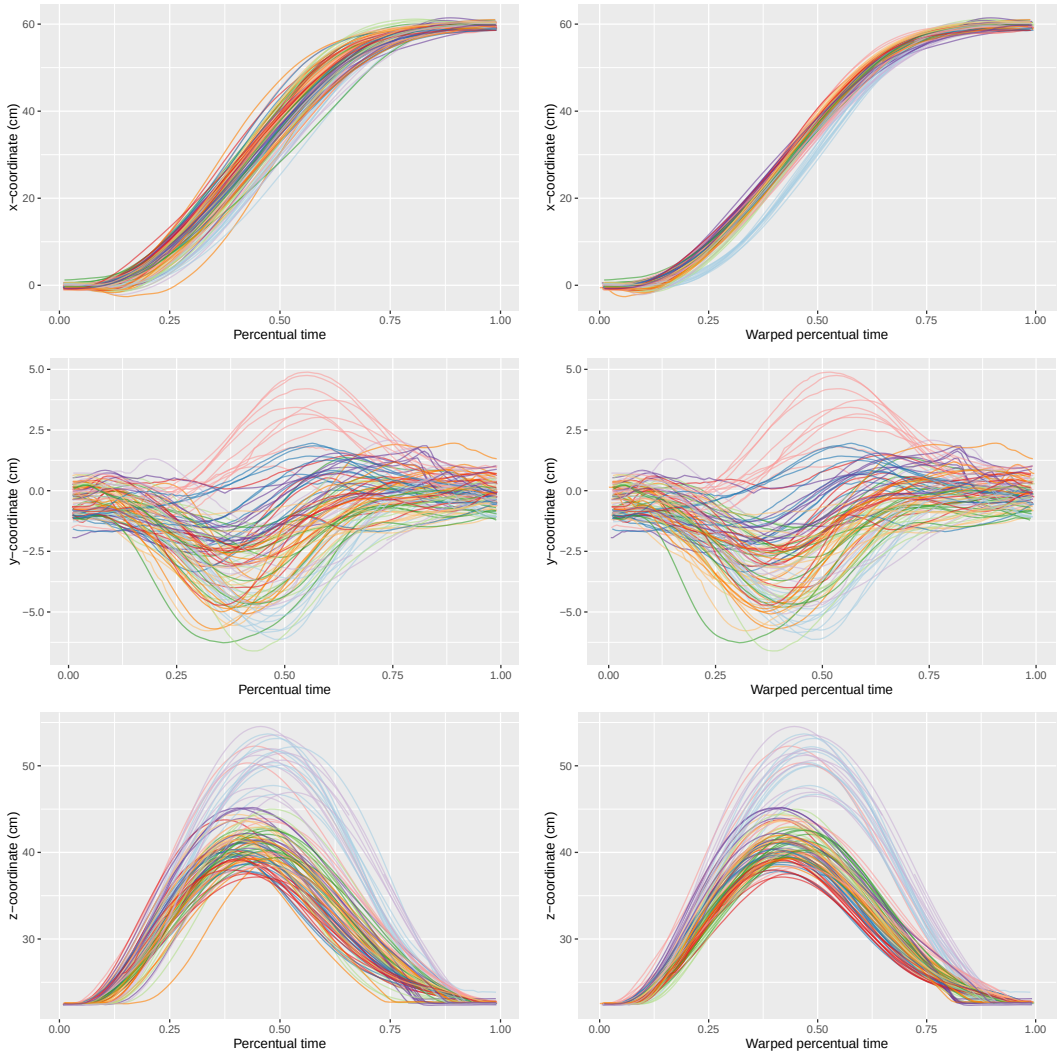


Figure 13: Data from the experiment with a small obstacle 30 cm from starting position plotted in percentual time (left column) and warped percentual time (right column). Coloring follows the coloring in Figure 12.

predicted warped percentual time. We see that the x- and z-coordinates are very well aligned within participant, and that the alignment of the y-coordinate seems to contain a relatively larger proportion of amplitude variation after alignment than the x- and z-coordinates. We note that the alignment procedure does not change the movement path in (x, y, z)-space. The predicted maximum-a-posteriori warping functions are displayed in Figure 14.

Parameter estimates The common variance parameter σ and the Matérn parameters α and κ varied little with experiment. On the other hand the relative weight, a, of the stationary

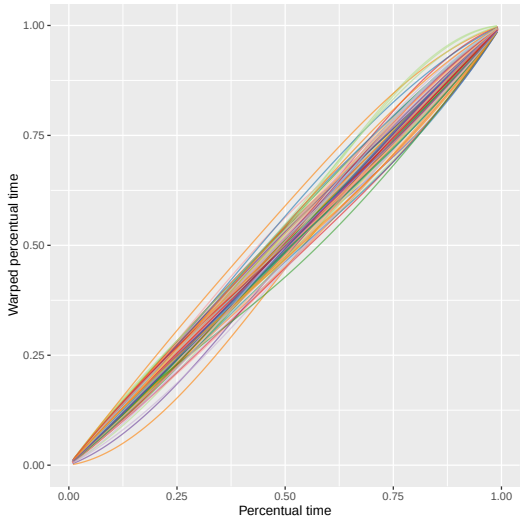


Figure 14: Predicted warping functions corresponding to the alignment in Figure 13.

covariance and the bridge covariance varied considerably across experiments. However a was large in all cases meaning that a large majority of the variance is captured by the stationary part. We refer to Table 2 in the supplementary material for all parameter estimates.

Variance and cross-correlations The amplitude variation was assumed to be generated from Gaussian processes $\boldsymbol{x}_n$ and white noise $\boldsymbol{\varepsilon}_n \sim N(0, \sigma^2 \boldsymbol{I}_{3m_n})$. Since the observed curves are very smooth the estimated contributions from the white noise terms were very small.

Figure 15 show the ratios of systematic amplitude variance to linearized systematic variance (amplitude and linearized warp) as estimated by the model. At the endpoints all variance was captured by the serially correlated amplitude effect. In the y-direction almost all variation was captured by the amplitude variance which fits well with the aligned y-coordinates of the movement path in Figure 13. The warp-related variance accounted for a larger part of the variation in the x- and z-directions. The temporal structure of the x-coordinate reveals that the warp effect explained the majority of the variance around the middle of the movement, while for the z-coordinate it explained the majority of the variance during lift and descend. Thus, the model predicted warping functions using a trade-off where the (percentual) temporal midpoints of the transport component and the lift and descend components had highest influence when measuring the alignment of samples.

The individual participant's estimated mean trajectories and the systematic amplitude variation are illustrated in Figure 16. In the right-hand illustration, the prediction ellipsoids in the middle are relatively small considering that this is the region with most variation. This is because most of the variation was captured by the participant-specific mean curves and the warping effect, as one would expect. The amplitude variance around the endpoints seems

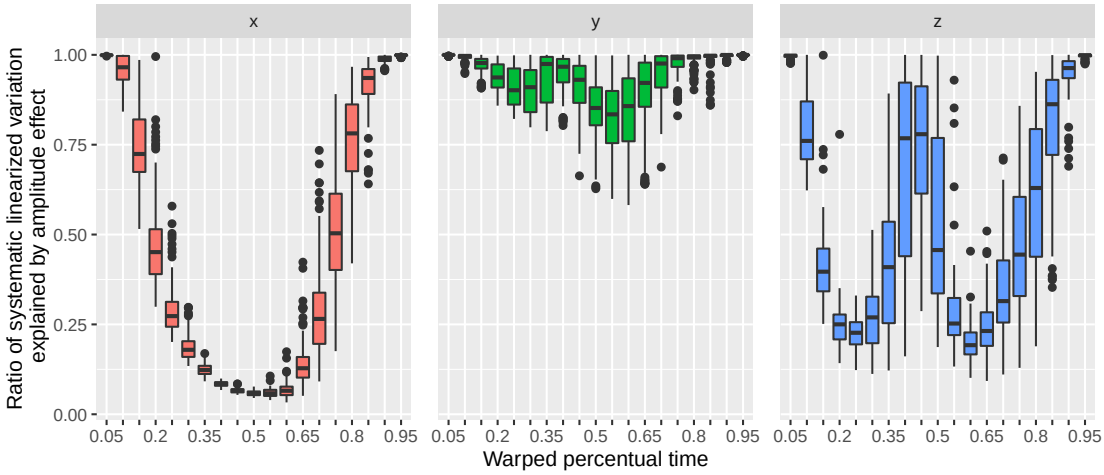


Figure 15: Coordinatewise boxplot of the temporal development of the ratio of S_n to $S_n + Z_n C Z_n^\top$ for the 100 samples in the experiment with a small obstacle 30 cm from starting position.

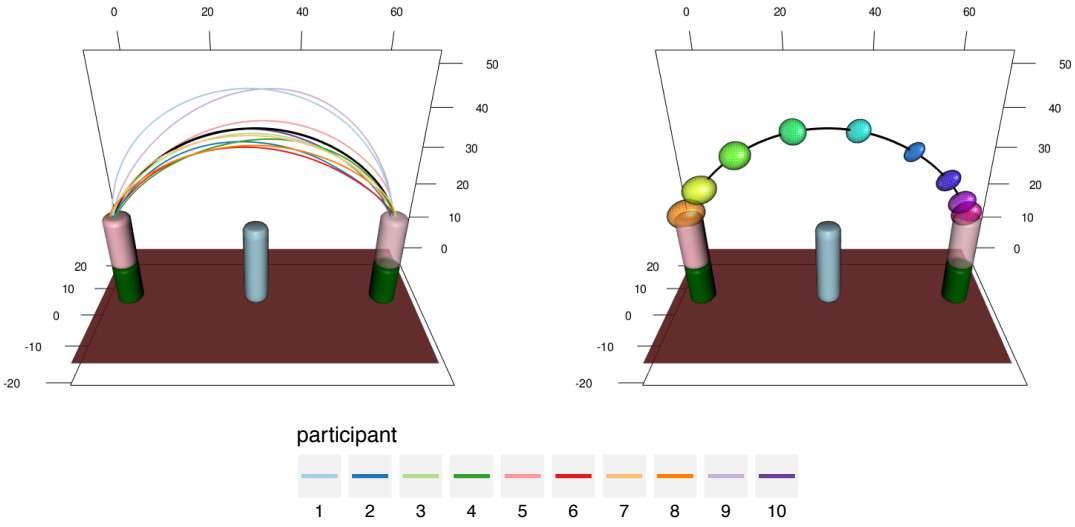


Figure 16: Estimated experiment-specific curve (black) and participant-specific curves for the experimental set-up with small obstacle 30 cm from starting position (left) and estimated 95% predictions ellipsoids for the systematic amplitude effect in the same set-up (right). The ellipsoids are displayed temporally equidistant around the mean trajectory for the experimental set-up.

somewhat overestimated, which suggests that the chosen anchor points provided a too coarse model for the dynamics of the true covariance function around the endpoints.

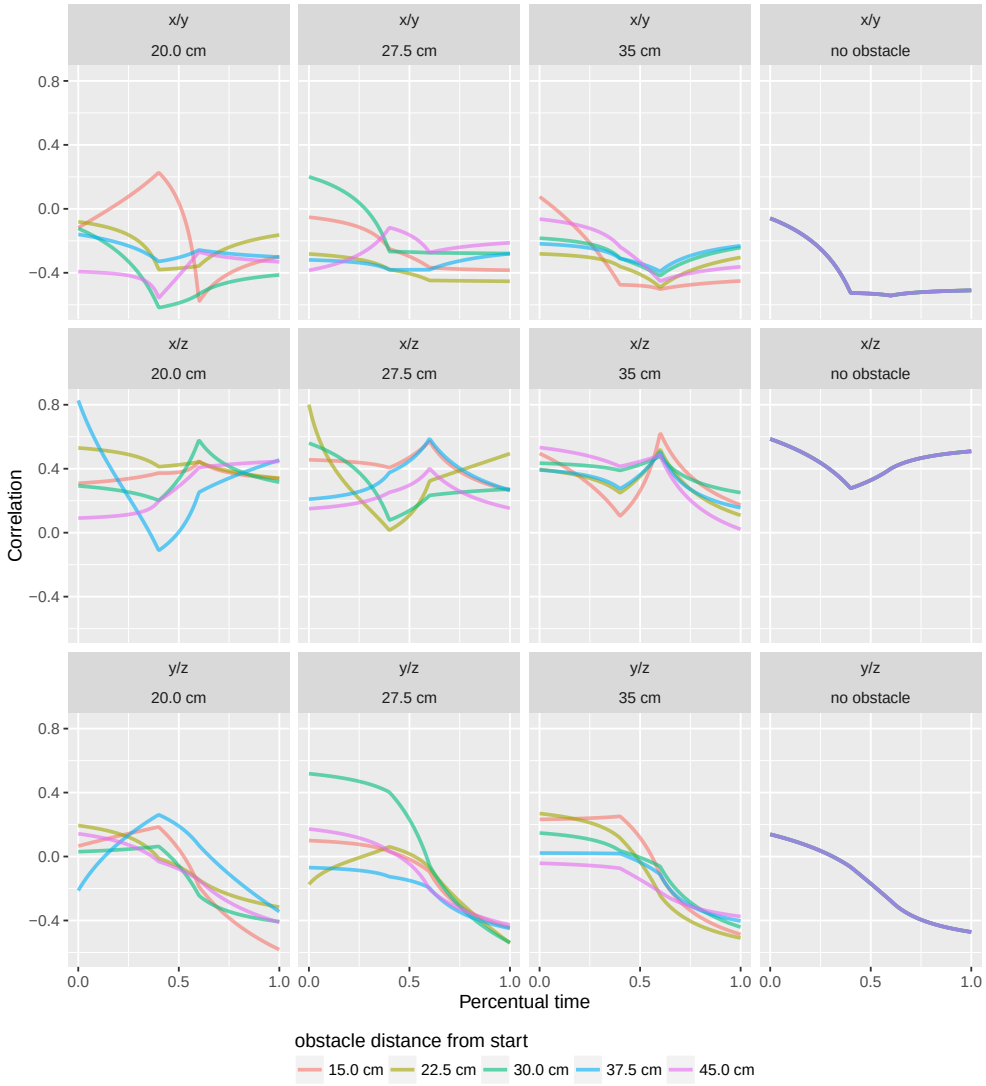


Figure 17: Correlation functions over time as estimated by the proposed model in all 16 experimental set-ups.

Of particular interest is the correlation for the three axes (i.e. x/y, x/z and y/z) and how it varies over time as seen in Figure 17. From the results, it is clear that the correlations vary over time, which Figure 16 also illustrates. The variation of correlation with respect to time is moderate for the x/y- and x/z-correlations, but for the y/z-correlations there is a clear trend for all experimental set-ups that the correlation goes from positive values to negative values. This is a surprising and perhaps unexpected feature since all experimental set-ups are symmetric in the y-coordinate. A plausible explanation is that lifting a centrally placed object with the right hand is generally associated with moving that hand to the right (in our set-up, a positive y-value). When the object is raised we observe a positive correlation in the y/z-plane (faster initial movement timing amplifies the effect), and when the object is lowered again we observe corresponding negative correlation.

Classification To objectively compare different models, one can fit the models to a subset of the samples and compare their fits in terms of their classification accuracies of participant on the remaining data. That is, for a given functional sample that was not used to fit the model, we wish to determine which of the participants performed the movement. The primary objective of such an exercise is to compare similar generative models, but not as such to get the highest possible classification accuracy—a higher score could probably be achievable by standard machine learning methods that would reveal little about the structure of the problem. A similar classification-based approach was used to evaluate the hierarchical "pavpop" model described in Raket et al. (2016), which was applied to the 1-dimensional acceleration magnitude profiles of the 3-dimensional arm movement data set.

The present classification was done in a chronological 5-fold cross-validation set-up (first fold consisted of the two first repetitions for each person, second fold of the third and fourth and so forth). Different models were fitted on the five training sets, each leaving out one of the folds (test set). For each test set, the samples were classified using the model estimates from the corresponding training set. The classification accuracy was then computed as the average classification accuracy across the five folds for each experiment.

In the following, the proposed method is denoted by SIMM (Simultaneous Inference for Misaligned Multivariate curves). The following models were used in the comparison:

Nearest centroid (NC) The centroids for each person were estimated as the pointwise means in the training set. The classification was done using minimal Euclidean distance to the estimated centroid (using linear interpolation).

Nearest centroid weighted (NC-W) The centroids were computed similarly to the NC method, but the classification was done using a distance with weighted coordinates, the weights for the x-, y- and z-coordinates were 0.1/0.7/0.2.

Fisher-Rao L^2 (FR-L^2) Pointwise template functions were estimated using group-wise elastic function alignment and PCA extraction for modeling amplitude variation (Tucker

et al. 2013, Tucker 2017). The standard setting of using 3 principal components was used. The elastic curve approach for functional data is widely considered the state-of-the-art framework for handling misaligned functional data (Marron et al. 2015). The template functions were estimated separately for each of the three value coordinates of the trajectories. Classification was done using minimal Euclidean distance to the estimated template functions.

Fisher-Rao elastic ($\mathbf{FR_E}$) Template functions were estimated similarly to FR-L^2, but classification was done using an elastic distance that both measures coordinate-wise distances as a sum of phase (Tucker et al. 2013, Section 3.1) and amplitude directions (Tucker et al. 2013, Definition 1). The weighting between phase and amplitude distances was 0.16/0.84.

Fisher-Rao elastic weighted ($\mathbf{FR_E}$-W) Template functions and classification was done similarly to $\mathrm{FR_E}$, except that we include a weighting of the three elastic distances corresponding to each value coordinate. The weighting between phase and amplitude distances was 0.14/0.86 and the weights for the x-, y- and z-components of the elastic distance were 0.3/0.2/0.5.

Elastic curve metric (EM) Multivariate elastic distance between curves is defined as geodesic distance on $L^2([0,1];\mathbb{R}^3)/\Gamma$, where Γ is the closure of the set of positive diffeomorphisms on $[0,1]$. In the quotient space $L^2([0,1];\mathbb{R}^3)/\Gamma$, all temporal features are removed and comparison of curves is done using only their image in $\mathbb{R}^3$, but in a way that is consistent with reparametrizations of the original curves (Srivastava & Klassen 2016). Templates were estimated as the pointwise averages of samples aligned to the Karcher mean in $L^2([0,1];\mathbb{R}^3)/\Gamma$ computed using the `fdasrvf` R-package (Tucker 2017). Classification was done using a weighted sum of multivariate elastic distance and phase distance (defined as for the $\mathrm{FR_E}$ method). The weighting between elastic and phase distances was 0.24/0.76.

SIMM The person-specific templates are estimated using the proposed model with a diagonal cross-covariance structure (i.e. no cross-covariance). Classification is done using nearest posterior distance under the maximum likelihood estimates as a function of the unknown sample.

SIMM-CC Estimation and classification are done similarly to the SIMM method, but using the full dynamic cross-covariance structure described in the previous sections.

All weights described in the above methods were chosen by cross-validation on the accuracies for the three experimental set-ups with $d = 30.0$ cm. The grids used for determining the parameters are given in the supplementary material.

The classification accuracies are available in Table 1. If we first consider the NC-type methods that do not model any warping effect, we see a marked increase in accuracy when weighting

the different coordinates in the classification, and thus emulating a constant diagonal cross-covariance structure. If we consider the basic elastic model FR-L^2 based on the Fisher-Rao metric, we see similar results to the simple NC model, even though the FR-L^2 method also accounts for a warping effect when estimating the template. When classifying using an elastic distance, as was done in FR_E, we see a great increase in classification accuracy. The phase distance contributes considerably to these improvements. When only considering elastic amplitude distance (i.e. weighting phase/amplitude distances 0/1) the average classification accuracy is 0.576. Taking the deformation distance into account in the classification, and thus paying a price for warping the templates, we see a great increase in classification accuracy. The heuristic idea of having to pay a price for large warps in many ways emulates the proposed idea of modeling the warping functions as random effects. Finally, the FR_E-W method includes a weighting of the combined phase and amplitude distances across the x-, y- and z-coordinates of the observed trajectories, which again increases the accuracy.

The elastic metric has many similarities with the Fisher-Rao metric, but is multivariate in nature. The EM method has higher accuracies than the similar FR_E and FR_E-W methods. Exploratory comparison of results suggested that this was caused by more appropriate warping across all coordinates leading to both better estimates of templates and in turn more accurate phase distances.

The SIMM model is the proposed model described above, but without a dynamic cross-correlation structure. Instead we have three scale parameters that describe the weighting of the marginal variances in the three value coordinates. The model is thus both comparable to FR_E-W and EM, both of which are outperformed in terms of accuracy. It is important to note that while FR_E-W and EM required cross-validation on a subset of the test data to estimate the parameters, the SIMM model estimates all variance parameters used in the weighting of the different aspect of the movement from the training data. The final model, SIMM-CC, includes a full dynamic cross-covariance structure. Even though one could anticipate that this model was much more prone to overfitting to the training data (the model includes 27 free amplitude variance parameters compared to the 6 parameters of the SIMM model), we see a slight increase in accuracy of the method. We remark that the EM, SIMM and SIMM-CC methods, which make a joint warp of the three spatial coordinates, had the best accuracies among the methods in consideration. This strongly supports the idea of modeling multivariate signals with a joint warping of all value coordinates.

5 Discussion

In this paper we have proposed a new class of models for simultaneous inference for misaligned multivariate functional data. We fitted these types of models to three different data sets and applied it in one classification scenario.

The idea behind the approach is to simultaneously model the predominant effects in func-

d	obstacle	NC	NC-W	FR-L^2	$\mathrm{FR_E}$	$\mathrm{FR_E}$-W	EM	SIMM	SIMM-CC
	S	0.62	0.71	0.58	0.77	0.79	0.77	0.80	**0.85**
15.0 cm	M	0.60	0.63	0.62	0.64	0.68	0.77	0.80	**0.83**
	T	0.52	0.57	0.54	0.58	0.58	0.77	**0.84**	0.81
	S	0.51	0.58	0.50	0.68	0.66	**0.77**	0.69	**0.77**
22.5 cm	M	0.52	0.64	0.56	0.62	0.73	0.70	**0.75**	0.72
	T	0.50	0.62	0.49	0.64	0.73	0.73	0.74	**0.79**
	S	0.53	*0.59*	0.53	*0.69*	*0.72*	***0.76***	0.70	**0.76**
30.0 cm	M	0.45	*0.47*	0.48	*0.65*	*0.68*	*0.70*	**0.79**	0.75
	T	0.58	*0.63*	0.56	*0.65*	*0.73*	*0.78*	**0.86**	0.83
	S	0.51	0.55	0.52	0.67	0.72	0.70	**0.77**	0.76
37.5 cm	M	0.45	0.50	0.43	0.68	0.65	0.69	**0.68**	**0.68**
	T	0.50	0.53	0.54	0.67	0.73	0.72	**0.80**	**0.80**
	S	0.49	0.54	0.51	0.66	0.71	0.75	0.69	**0.76**
45.0 cm	M	0.48	0.53	0.44	0.66	0.70	0.71	**0.78**	0.73
	T	0.50	0.54	0.50	0.71	0.75	0.74	0.82	**0.83**
NA	-	0.48	0.56	0.52	0.68	0.72	**0.80**	0.64	0.70
average		0.515	0.574	0.520	0.666	0.705	0.741	0.761	**0.773**

Table 1: Classification accuracies of various methods. **Bold** indicates best result(s), *italic* indicates that the given experiments were used for training.

tional data sets, misalignment and amplitude variation, as random effects. The simultaneous modeling allows separation of these effects in a data-driven manner, namely by maximum likelihood estimation. In particular, we saw that this separation resulted in nicely behaving warping functions that did not seem to over-align the functional samples.

The models enable estimation of dynamic correlation functions between the individual coordinates of the amplitude variation. We demonstrated that one can achieve superior fits and better classification using the parametric construction from Proposition 1, even when the number of free parameters is high relative to the number of functional samples. By fitting the model to two large functional data sets related to human movement, we also demonstrated the computational feasibility of maximum likelihood inference with such models.

The proposed parametric model class for dynamic covariance structures is very general, but other modeling approaches could be better suited in some situations. For example, instead of using a fixed number of parameters to describe each marginal variance and cross-covariance function, one would often prefer to do this in a data-driven manner. One possibility could be to model the multivariate amplitude covariance function using a multivariate functional factor analysis model, for example a multivariate extension of the rank reduced model of James et al. (2000), where the number of parameters describing the covariance is fixed, and the covariance is described in terms of functional principal components. However, such amplitude effects cannot be effectively fitted using conventional optimizers for the likelihood, and would require development of specialized efficient fitting methods (e.g. generalizing the methods of Peng & Paul 2009). Another relevant approach would be simultaneous warping of fixed effects and amplitude variation, and one could also consider extending the domain of feasible warping functions by modelling the latent warp variables w as more general functional objects (e.g. stochastic processes) instead of elements belonging to $\mathbb{R}^{m_{\boldsymbol{w}}}$ for some $m_{\boldsymbol{w}}$. We will leave these extensions as future work.

SUPPLEMENTARY MATERIAL

Cross-validation grids

The cross-validation used to determine the parameters of the methods NC-W, FR_E, FR_E-W and EM in Section 4.3 were given as follows. The possible weights between the three value coordinates were $\{\boldsymbol{w} \in \mathbb{R}^3 : w_i \in \{0, 0.1, \ldots, 1\}, w_1 + w_2 + w_3 = 1\}$ and the possible weights between amplitude and phase distance were $\{\boldsymbol{w} \in \mathbb{R}^2 : w_i \in \{0, 0.02, \ldots, 1\}, w_1 + w_2 = 1\}$. NC-W only uses weighting between value coordinates and FR_E and EM only use weighting between the amplitude and phase distance.
For the SIMM-CC model we explored adding more than three knots to the warp model ($m_{\boldsymbol{w}} = 3, 4, 5$), but $m_{\boldsymbol{w}} = 3$ gave the best cross-validation score.

Covariance functions

Below we list the covariance functions that are used in the three data examples.

Schur's theorem states that the pointwise product of covariance functions yields a valid covariance function (Schur 1911). This property is used in the arm movement example.

Brownian bridge The covariance function for the Brownian bridge defined on the temporal domain $[0, 1]$ is given by:

$$f_{\text{bridge}}(s,t) = \tau^2 \min(s,t) \cdot (1 - \max(s,t)) = \tau^2(\min(s,t) - st), \quad s,t \in [0,1]. \tag{13}$$

where $\tau > 0$ is a scale parameter.

Brownian motion The covariance function for the Brownian motion defined on the domain $[0, \infty)$ is given by:

$$f_{\text{motion}}(s,t) = \tau^2 \min(s,t), \quad s,t \geq 0. \tag{14}$$

where $\tau > 0$ is a scale parameter.

Mixing stationary and bridge covariances The combination of a stationary and bridge covariance with mixtures a and b is given by

$$f_{\text{mixture}(a,b)}(s,t) = a + b \cdot (\min(s,t) - st)$$

In our analysis the parameter b is redundant, so we use

$$f_{\text{mixture}(a)}(s,t) = a + \min(s,t) - st \tag{15}$$

Note that the bridge covariance is not the same construction as when conditioning a stochastic process X on its endpoint value.

Matérn covariance function The covariance function for the Matérn covariance with smoothness parameter α and range parameter κ is given by:

$$f_{\text{Matérn}(\alpha,\kappa)}(s,t) = \frac{2^{1-\alpha}}{\Gamma(\alpha)}(|s-t|/\kappa)^\alpha K_\alpha(|s-t|/\kappa), \quad s,t \in \mathbb{R}. \tag{16}$$

Here K_α is the modified Bessel function of the second kind. A Gaussian process with Matérn covariance is stationary, and conversely any stationary continuous Gaussian process with mean zero has a covariance function that up to scale is given by a Matérn covariance function (Rasmussen & Williams 2006).

Parameter estimates for Arm movement data

d	obstacle	σ	α	κ	a	$\sigma\tau$
15.0 cm	S	0.0012	1.432	0.157	19.56	0.0519
	M	0.0012	1.749	0.120	24.60	0.0525
	T	0.0013	1.627	0.124	22.54	0.0502
22.5 cm	S	0.0013	1.788	0.128	25.13	0.0531
	M	0.0011	1.638	0.139	58.20	0.1177
	T	0.0012	1.679	0.121	26.37	0.0528
30.0 cm	S	0.0012	1.773	0.121	20.96	0.0549
	M	0.0014	1.663	0.139	21.31	0.0518
	T	0.0012	1.687	0.128	26.63	0.0643
37.5 cm	S	0.0012	1.481	0.155	17.69	0.0622
	M	0.0013	1.658	0.125	19.80	0.0596
	T	0.0010	1.633	0.121	34.29	0.0563
45.0 cm	S	0.0013	1.761	0.123	19.10	0.0504
	M	0.0016	1.760	0.119	13.09	0.0668
	T	0.0010	1.670	0.121	37.46	0.0548
NA	-	0.0009	1.786	0.142	47.04	0.0561

Table 2: Parameter estimates for the arm movement data.

EM algorithm for the spline coefficients in the linearized model

First note that by assumption the mean curves $\boldsymbol{\theta}$ are the same, expect for warping, for trajectories belonging to the same subject groups and are independent of other subject groups. Thus, in order to simplify notation and ease argumentation, we will assume that all trajectories belong to the same subject group.

Let $f = \{f_k\}$ be the spline base function for $\boldsymbol{\theta}$ and let $\boldsymbol{c}$ be the spline coefficients, i.e. $\boldsymbol{\theta}(t) = f(t) \cdot \boldsymbol{c}$. Consider the linearized model from Equation (10):

$$\vec{\boldsymbol{y}}_n \approx \vec{\boldsymbol{\gamma}}_{\boldsymbol{w}_n^0} + Z_n(\boldsymbol{w}_n - \boldsymbol{w}_n^0) + \vec{\boldsymbol{x}}_n + \vec{\boldsymbol{\varepsilon}}_n, \quad n = 1, \ldots, N$$

with log-likelihood

$$\sum_{n=1}^{N} \left(q m_n \log \sigma^2 + \log\det V_n + \sigma^{-2} (\vec{\boldsymbol{y}}_n - \vec{\boldsymbol{\gamma}}_{\boldsymbol{w}_n^0} + Z_n \boldsymbol{w}_n^0)^\top V_n^{-1} (\vec{y}_n - \vec{\boldsymbol{\gamma}}_{\boldsymbol{w}_n^0} + Z_n \boldsymbol{w}_n^0) \right).$$

For the remainder we assume that $\boldsymbol{w}_n^0 = \{\boldsymbol{w}_{nl}^0\}_{l=1}^{m_w}$ and all variance parameters $(\mathcal{S}, C, \sigma^2)$ are fixed, and that we have a current estimate of the spline coefficients, $\boldsymbol{c}_0$. The conditional

expectation and variance of $\boldsymbol{w}_n$ given the observations $\boldsymbol{y}$ under the current parameters will be denoted by $\bar{\boldsymbol{w}}_n = \{\bar{\boldsymbol{w}}_{nl}\}_{l=1}^{m_w} \in \mathbb{R}^{m_w}$ and $\bar{\bar{\boldsymbol{w}}}_n = \{\bar{\bar{\boldsymbol{w}}}_{nl_1l_2}\}_{l_1,l_2=1}^{m_w} \in \mathbb{R}^{m_w \times m_w}$, respectively. Using this notation the conditional log-likelihood of $\vec{\boldsymbol{y}}_n$ given $\boldsymbol{w}_n$ is

$$l_{\vec{\boldsymbol{y}}_n|\boldsymbol{w}_n} = (\vec{\boldsymbol{y}}_n - \vec{\boldsymbol{\gamma}}_{\boldsymbol{w}_n^0} - Z_n(\boldsymbol{w}_n - \boldsymbol{w}_n^0))^\top S_n^{-1}(\vec{\boldsymbol{y}}_n - \vec{\boldsymbol{\gamma}}_{\boldsymbol{w}_n^0} - Z_n(\boldsymbol{w}_n - \boldsymbol{w}_n^0)) + \log\det S_n.$$

The term $\log\det S_n$ does not influence the estimation of $\boldsymbol{c}$, and hence it will be removed in the following. The conditional expectation $E[l_{\vec{\boldsymbol{y}}_n|\boldsymbol{w}_n}|\vec{\boldsymbol{y}}_n]$ given the observation hence equals

$$(\vec{\boldsymbol{y}}_n - \vec{\boldsymbol{\gamma}}_{\boldsymbol{w}_n^0} - Z(\bar{\boldsymbol{w}}_n - \boldsymbol{w}_n^0))^\top S_n^{-1}(\vec{\boldsymbol{y}}_n - \vec{\boldsymbol{\gamma}}_{\boldsymbol{w}_n^0} - Z(\bar{\boldsymbol{w}}_n - \boldsymbol{w}_n^0)) + \operatorname{tr}(S_n^{-1}Z_n\bar{\bar{\boldsymbol{w}}}_nZ_n^\top). \tag{17}$$

Defining $R_n = f(v(t_k, \boldsymbol{w}_n^0))$ and $R_{nl} = \partial_t f(v(t_k, \boldsymbol{w}_n^0))\partial_{\boldsymbol{w}_l} v(t_k, \boldsymbol{w}_n^0)$ for $l = 1, \ldots, m_w$ we have that $Z_n = \{R_{nl} \cdot \boldsymbol{c}\}_{l=1}^{m_w}$ and thus $Z_n\boldsymbol{w}_n = (\sum_{l=1}^{m_w} \boldsymbol{w}_{nl}R_{nl}) \cdot \boldsymbol{c}$. Using this the trace from (17) can be expanded as a double sum

$$\operatorname{tr}(S_n^{-1}Z\bar{\bar{\boldsymbol{w}}}_nZ^\top) = \sum_{l_1,l_2=1}^{m_w} \bar{\bar{\boldsymbol{w}}}_{nl_1l_2} \operatorname{tr}\left(S_n^{-1}R_{nl_1}\boldsymbol{c}\boldsymbol{c}^\top R_{nl_2}^\top\right).$$

Calculating the gradient of (17) now gives that $\nabla_{\boldsymbol{c}} E[l_{\vec{\boldsymbol{y}}_n|\boldsymbol{w}_n}|\vec{\boldsymbol{y}}_n]$ is proportional to

$$-K_n^\top S_n^{-1}(\vec{\boldsymbol{y}}_n - K_n\boldsymbol{c}) + \sum_{l_1,l_2=1}^{m_w} \bar{\bar{\boldsymbol{w}}}_{nl_1l_2}R_{nl_2}^\top S_n^{-1}R_{nl_1}\boldsymbol{c},$$

where $K_n = R_n + \sum_{l=1}^{m_w}(\bar{\boldsymbol{w}}_{nl} - \boldsymbol{w}_{nl}^0)R_{nl}$. From this it follows that the M-step of the EM algorithm for the spline coefficients $\boldsymbol{c}$ is given by

$$\boldsymbol{c}_{\text{new}} = \left[\sum_{n=1}^{N} K_n^\top S_n^{-1}K_n + \sum_{l_1,l_2=1}^{m_w} \bar{\bar{\boldsymbol{w}}}_{nl_1l_2}R_{nl_1}S_n^{-1}R_{nl_2}^\top\right]^{-1} \sum_{n=1}^{N} K_n^\top S_n^{-1}y_n.$$

References

Aksglaede, L., Sørensen, K., Petersen, J. H., Skakkebæk, N. E. & Juul, A. (2009), 'Recent decline in age at breast development: the Copenhagen Puberty Study', *Pediatrics* **123**(5), e932–e939.

Beath, K. J. (2007), 'Infant growth modelling using a shape invariant model with random effects', *Statistics in Medicine* **26**(12), 2547–2564.

CMU Graphics Lab Motion Capture Database (n.d.), `http://mocap.cs.cmu.edu/`.

Cole, T. J., Donaldson, M. D. & Ben-Shlomo, Y. (2010), 'SITAR—a useful instrument for growth curve analysis', *International Journal of Epidemiology* **39**, 1558–1566.

Dryden, I. L. & Mardia, K. V. (1998), *Statistical shape analysis*, Vol. 4, J. Wiley Chichester.

FSU (n.d.), 'Statistical Shape Analysis & Modeling Group software available for free public use', `http://ssamg.stat.fsu.edu/software/`.

Gervini, D. & Gasser, T. (2005), 'Nonparametric maximum likelihood estimation of the structural mean of a sample of curves', *Biometrika* **92**(4), 801–820.

Grimme, B. (2014), Analysis and identification of elementary invariants as building blocks of human arm movements, PhD thesis, International Graduate School of Biosciences, Ruhr-Universität Bochum. (In German).

Grimme, B., Lipinski, J. & Schöner, G. (2012), 'Naturalistic arm movements during obstacle avoidance in 3D and the identification of movement primitives', *Experimental Brain Research* **222**(3), 185–200.

Guo, W. (2002), 'Functional mixed effects models', *Biometrics* **58**(1), 121–128.

Hadjipantelis, P. Z., Aston, J. A., Müller, H.-G. & 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.

Hadjipantelis, P. Z., Aston, J. A., Müller, H.-G., Moriarty, J. et al. (2014), 'Analysis of spike train data: A multivariate mixed effects model for phase and amplitude', *Electronic Journal of Statistics* **8**(2), 1797–1807.

Hyman, J. M. (1983), 'Accurate monotonicity preserving cubic interpolation', *SIAM Journal on Scientific and Statistical Computing* **4**(4), 645–654.

James, G. M., Hastie, T. J. & Sugar, C. A. (2000), 'Principal component models for sparse functional data', *Biometrika* **87**(3), 587–602.

Kendall, D. G. (1989), 'A survey of the statistical theory of shape', *Statistical Science* **18**, 87–99.

Kneip, A. & Ramsay, J. O. (2008), 'Combining registration and fitting for functional models', *Journal of the American Statistical Association* **103**(483), 1155–1165.

Kurtek, S., Srivastava, A., Klassen, E. & Ding, Z. (2012), 'Statistical modeling of curves using shapes and related features', *Journal of the American Statistical Association* **107**(499), 1152–1165.

Lindstrom, M. J. & Bates, D. M. (1990), 'Nonlinear mixed effects models for repeated measures data', *Biometrics* **46**(3), 673–687.

Manay, S., Cremers, D., Hong, B.-W., Yezzi, A. J. & Soatto, S. (2006), 'Integral invariants for shape matching', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28**(10), 1602–1618.

Marron, J., Ramsay, J. O., Sangalli, L. M. & Srivastava, A. (2015), 'Functional data analysis of amplitude and phase variation', *Statistical Science* **30**(4), 468–484.

Peng, J. & Paul, D. (2009), 'A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data', *Journal of Computational and Graphical Statistics* (4), 995–1015.

Raket, L. L., Grimme, B., Schöner, G., Igel, C. & Markussen, B. (2016), 'Separating timing, movement conditions and individual differences in the analysis of human movement', *PLoS Computational Biology* **12**(9), e1005092.

Raket, L. L., Sommer, S. & Markussen, B. (2014), 'A nonlinear mixed-effects model for simultaneous smoothing and registration of functional data', *Pattern Recognition Letters* **38**, 1–7.

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

Rasmussen, C. E. & Williams, C. K. I. (2006), *Gaussian Processes for Machine Learning*, The MIT Press.

Rønn, B. B. (2001), 'Nonparametric maximum likelihood estimation for shifted curves', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63**(2), 243–259.

Rønn, B. B. & Skovgaard, I. M. (2009), 'Nonparametric maximum likelihood estimation of randomly time-transformed curves', *Brazilian Journal of Probability and Statistics* **23**(1), 1–17.

Schur, J. (1911), 'Bemerkungen zur Theorie der beschränkten Bilinearformen mit unendlich vielen Veränderlichen.', *Journal für die reine und Angewandte Mathematik* **140**, 1–28.

Sebastian, T. B., Klein, P. N. & Kimia, B. B. (2003), 'On aligning curves', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25**(1), 116–125.

Sørensen, K., Aksglaede, L., Petersen, J. H. & Juul, A. (2010), 'Recent changes in pubertal timing in healthy danish boys: associations with body mass index', *The Journal of Clinical Endocrinology & Metabolism* **95**(1), 263–270.

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

Srivastava, A. & Klassen, E. P. (2016), *Functional and Shape Data Analysis*, Springer.

Tinggaard, J., Aksglaede, L., Sørensen, K., Mouritsen, A., Wohlfahrt-Veje, C., Hagen, C. P., Mieritz, M. G., Jørgensen, N., Wolthers, O. D., Heuck, C. et al. (2014), 'The 2014 Danish references from birth to 20 years for height, weight and body mass index', *Acta Paediatrica* **103**(2), 214–224.

Tucker, J. D. (2017), *fdasrvf: Elastic Functional Data Analysis*. R package version 1.8.3. **URL:** *https://github.com/jdtuck/fdasrvf_R/*

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

Vantini, S. (2012), 'On the definition of phase and amplitude variability in functional data analysis', *Test* **21**(4), 676–696.

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

Wolfinger, R. (1993), 'Laplace's approximation for nonlinear mixed models', *Biometrika* **80**(4), 791–795.

Younes, L. (1998), 'Computable elastic distances between shapes', *SIAM Journal on Applied Mathematics* **58**(2), 565–586.

II

Markov component analysis for multivariate functional data

Bo Markussen
Department of Mathematical Sciences
University of Copenhagen

Niels Lundtorp Olsen
Department of Mathematical Sciences
University of Copenhagen

Publication details
Draft.

Markov component analysis for multivariate functional data

Bo Markussen, Niels Lundtorp Olsen
Department of Mathematical Sciences, University of Copenhagen

Abstract

Multivariate functional data analysis has got little attention in literature, and there are few dedicated methods for this. In this paper we introduce a class of factorizable matrices and corresponding Cholesky calculus that allow for computations with linear complexity in the data length. This class of matrices and its calculus can also be formulated in the continuous setting using linear operators. The proposed class of factorizable matrices is used to construct a model framework for multivariate, possibly misaligned functional data. Amplitude and phase variation is modelled by the same underlying Gaussian process, which allow for correlation between phase and amplitude. We devise a method for doing maximum-likelihood estimation using the EM algorithm, and the proposed model is applied to a data set on horse lameness. Whereas the mean structure is easily identified, our current implementation of the estimation procedure is too slow for identifying phase variation.

1 Introduction

Principal component analysis [1], *partial least squares* [2], *factor analysis* [3], and *canonical correlation analysis* [4] are among the most frequently applied statistical analyses of multivariate response data $Y \in \mathbb{R}^D$. In recent years increasingly more data is not only possibly multivariate, but also collected over time. The analysis of such data $X\colon [a,b] \to \mathbb{R}^D$ was named *functional data analysis* (FDA) by Ramsay [6]. Most research on FDA has been concerned with univariate functional data, i.e. when $D = 1$. However, even univariate functional data can be seen as an instance of multivariate data. This viewpoint arise naturally in two different ways. Either because the sampling of the functional data at discrete time points $t_1 < \ldots < t_J$ is seen as J-dimensional data $\{X(t_j)\}_{j=1}^J \in \mathbb{R}^J$. Or since the functional data is represented by a finite dimensional basis expansion $X(t) = \sum_{i=1}^K \phi_i(t)c_i$. The latter viewpoint is by far most prominent in the literature, and many papers suggest that basis representations constitute an intrinsic character of functional data analysis. In such a set-up the approximation dimension K and the basis functions ϕ_i must be selected, after which the data is identified with the K-dimensional data $\{c_i\}_{i=1}^K \in \mathbb{R}^K$. One way of doing that is by *functional principal component analysis*, which in its most simple variant consists in selecting the basis functions as some functional interpolation of the loadings from a principal components analysis (PCA) of the discretely sampled data $\{X(t_j)\}_{j=1}^J \in \mathbb{R}^J$.

The literature contains different variants of functional PCA, which may be used for selecting a basis representation as discussed above. Beside dimension reduction these methods are also extensively used for smoothing and denoising. Concerning functional variants of the three other multivariate analysis methods listed above, i.e. partial least squares (PLS), factor analysis (FA), and canonical correlation analysis (CCA), the existing literature is comparably sparse (see e.g. [7, 8, 9, 10]). One explanation for this might be that the literature on other aspects of analysis of multivariate data $X\colon [a,b] \to \mathbb{R}^D$, such as alignment and phase variation, in general is comparable sparse, and that methods like PLS and CCA only make sense when $D > 1$. But the explanation might also lie in the mathematical formulation of the multivariate methods. The original and still mostly used formulations of PCA, PLS and CCA are in terms of maximization of variance and correlations, respectively. However, in particular the maximization of raw correlation does not carry any information in the functional set-up, where perfect correlation may be achieved when the dimension of the response is larger than the number of observations [7].

To circumvent the caveat imposed by maximization of variances and correlations we propose to extend the probabilistic formulations of PCA [11] and CCA [12] to the functional setting. These probabilistic formulations place PCA and CCA in the same model framework as the original formulation of FA. In Section 2 we review the probabilistic interpretations of PCA and CCA. Furthermore, we will argue that PLS can be given a similar formulation, in which we have a natural ordering of the methods as PCA $\subset$ PLS $\subset$ FA $\subset$ CCA.

The generalization of the multivariate models to the functional setting that we propose in

Section 2.1 leads to the usage of Gaussian Markov processes. The variance-covariance function $K\colon [a,b]\times[a,b]\to\mathbb{R}^{D\times D}$ of such processes can be shown to have the structure

$$K(t,s)=\begin{cases} g(t)h(s)^\top & \text{for } t\geq s,\\ h(t)g(s)^\top & \text{for } t<s \end{cases} \tag{1}$$

with $g,h\colon [a,b]\to\mathbb{R}^{D\times q}$ for some $q\in\mathbb{N}$. Kernels on the form Eq. (1) are called *factorizable* in [13], and such kernels arise frequently in analysis (ref?) as well as probability theory [16, 17]. In Section 3.1 we devise an efficient computational framework for such covariance structures, which is employed in Section 4 to do likelihood inference in the proposed probabilistic models.

Finally, in Section 5 we provide further applications including covariance estimation and a PLS alternative to standard functional regression [18].

2 Probabilistic models for multivariate data

Mathematically, FA as well as the probabilistic model formulations of PCA, PLS and CCA all can be given the graphical representation

$$Z \xrightarrow{\beta} \boxed{Y} \longleftarrow U, \tag{2}$$

that is $Y_n=\beta Z_n+U_n$ if $n=1,\ldots,N$ denotes the observation index. Here $Y\in\mathbb{R}^D$ is the observed D-dimensional data, $Z\in\mathbb{R}^q$ in an unobserved q-dimensional latent variable, $U\in\mathbb{R}^D$ is the error term, and $\beta\in\mathbb{R}^{D\times q}$ is the matrix defining the linear transformation from the latent variable to the observation space.

In the wording of PCA, $Z_1,\ldots,Z_N$ are the *scores* and β is the matrix of *loadings*. The difference between classical PCA and the probabilistic formulation given by [11] may be understood in terms of the modelling of the scores within the framework of *mixed models* (refs?). Here the scores $\{Z_n\}_{n=1}^N\in\mathbb{R}^{N\times q}$ are fixed effects for classical PCA, but independent random effects $Z_n\sim\mathcal{N}(0,\mathrm{I}_q)$ for probabilistic PCA. But in both cases the error terms may be interpreted as independent normally distributed random variables $U_n\sim\mathcal{N}(0,\sigma^2\,\mathrm{I}_D)$, and the loading matrix $\beta\in\mathbb{R}^{D\times q}$ as well as the error variance $\sigma^2>0$ are parameters that may be found by maximum likelihood estimation. In this formulation probabilistic PCA is a linear random effects model, where β is a common but unknown design matrix for the random effects. Similarly, classical PCA is a bilinear fixed effects model.

The extension from probabilistic PCA to the other probabilistic multivariate models lies in the variance assumption on the error terms $U_n\sim\mathcal{N}(0,\Upsilon)$. For FA the extension is from homogeneous variances $\Upsilon=\sigma^2\,\mathrm{I}_D$ to heterogeneous variance, that is $\Upsilon=\mathrm{diag}(\sigma_1^2,\ldots,\sigma_D^2)$. For CCA the standard set-up is to have two multivariate responses $Y^{[1]}\in\mathbb{R}^{D_1}$ and $Y^{[2]}\in\mathbb{R}^{D_2}$, and the graphical representation

$$U^{[1]}\longrightarrow\boxed{Y^{[1]}}\xleftarrow{\beta_1} Z\xrightarrow{\beta_2}\boxed{Y^{[2]}}\longleftarrow U^{[2]}.$$

However, if the two response vectors are stacked into a single multivariate response $Y = \begin{pmatrix} Y^{[1]} \\ Y^{[2]} \end{pmatrix} \in \mathbb{R}^D$ with $D = D_1 + D_2$, then [12] shows that classical and probabilistic CCA finds the same maximum likelihood estimates for the loadings $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \in \mathbb{R}^{D \times q}$, where the variance assumption on the error terms is

$$U_n = \begin{pmatrix} U_n^{[1]} \\ U_n^{[2]} \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} \Upsilon_1 & 0_{D_1 \times D_2} \\ 0_{D_2 \times D_1} & \Upsilon_2 \end{pmatrix}\right)$$

with general positive definite matrices $\Upsilon_1 \in \mathbb{R}^{D_1 \times D_1}$ and $\Upsilon_2 \in \mathbb{R}^{D_2 \times D_2}$. For PLS the standard set-up is that of regressing a multivariate response $Y^{[1]} \in \mathbb{R}^{D_1}$ on a multivariate regressor $Y^{[2]} \in \mathbb{R}^{D_2}$. The idea in PLS is to impose a trade-off between explaining as much variation as possible on $Y^{[1]}$ and using as much variation as possible from $Y^{[2]}$. This may be done in the CCA set-up by assuming $\Upsilon_1 = \sigma^2 \mathrm{I}_{D_1}$ and $\Upsilon_2 = \xi\sigma^2 \mathrm{I}_{D_2}$, where the trade-off between variances is parametrized by $\xi > 0$. This parameter may either be inferred from data, or be preselected by the user. Although very natural we have not seen this formulation of PLS in the literature.

In summary, if the trade-off in PLS is inferred from data, then PCA, PLS, FA, and CCA have the same model structure Eq. (2), with increasingly more general error variances $\Upsilon \in \mathbb{R}^{D \times D}$. Namely,

$$\begin{aligned} &\text{PCA: } \Upsilon = \sigma^2 \mathrm{I}_D, && \text{PLS: } \Upsilon = \begin{pmatrix} \sigma_1^2 \mathrm{I}_{D_1} & 0_{D_1 \times D_2} \\ 0_{D_2 \times D_1} & \sigma_2^2 \mathrm{I}_{D_2} \end{pmatrix}, \\ &\text{FA: } \Upsilon = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_D^2 \end{pmatrix}, && \text{CCA: } \Upsilon = \begin{pmatrix} \Upsilon_1 & 0_{D_1 \times D_2} \\ 0_{D_2 \times D_1} & \Upsilon_2 \end{pmatrix}. \end{aligned} \tag{3}$$

In the following we will only use the probabilistic versions of these models, i.e. with independent random scores $Z_n \sim \mathcal{N}(0, \mathrm{I}_q)$. Beside estimation and interpretation of the model parameters $\beta \in \mathbb{R}^{D \times q}$ and $\Upsilon \in \mathbb{R}^{D \times D}$, the latter subject to the variance restrictions in the different models, various predictions are also of interest in applications. The interpretation of PCA and FA is often done via predictions of the scores, i.e. the conditional means $\mathrm{E}[Z_n | Y_n]$. PLS is often used in a regression set-up, which amounts to $\mathrm{E}[Y_n^{[1]} | Y_n^{[2]}]$. And being the most general CCA often is interpreted and used in both ways.

2.1 Extension to multivariate functional data

To extend FA and probabilistic PCA, PLS, and CCA to multivariate functional data we propose to extend the q-dimensional random score vector to a stochastic process $Z \colon [a, b] \to$

$\mathbb{R}^q$, which we will call the *score process* throughout this paper. This concept is not to be confused with the notion of the score process for likelihood analysis of stochastic processes. The score process is inserted in the generalization of Eq. (2) to the functional setting, which we propose as

$$Z(t) \xrightarrow{\beta(t)} \boxed{X(t)} \longleftarrow U(t), \tag{4}$$

that is $X(t_j) = \beta(t_j)Z(t_j) + U_j$ with independent error terms $U_1, \ldots, U_J$.

In the multivariate setting the score vector was normally distributed, and so we will assume that the score process is a Gaussian process. Furthermore, to encompass the ordering of time inherent in functional data we will assume that the scores process is Markov. [5] gives a characterization of univariate Gaussian Markov processes. The generalization of this to multivariate processes states that a continuous Gaussian process $Z\colon [a,b] \to \mathbb{R}^q$ is Markov if and only if there exists a non-vanishing continuous function $\tilde{\beta}\colon [a,b] \to \mathbb{R}^{q\times q}$, a matrix $\gamma_0 \in \mathbb{R}^{q\times q}$, and a continuous function $\gamma\colon [a,b] \to \mathbb{R}^{q\times q}$ such that

$$\operatorname{Cov}\big(Z(t), Z(s)\big) = \tilde{\beta}(t)\bigg(\gamma_0\gamma_0^\top + \int_a^{\min\{t,s\}} \gamma(u)\gamma(u)^\top \,\mathrm{d}u\bigg)\tilde{\beta}(s)^\top.$$

Since the factor $\tilde{\beta}(t)$ may be adsorbed by the matrix of loadings $\beta(t)$ provided by Eq. (4), we may without loss of generality for the present usage assume that $\tilde{\beta}(t) = \mathrm{I}_q$. At this point in the construction a few remarks are in place.

Remark 1. *We may without loss of generality assume that γ_0 and $\gamma(t)$ are lower triangular. Doing this implies a Choleskey decomposition of the increasing function $f(t) = \gamma_0\gamma_0^\top + \int_a^t \gamma(u)\gamma(u)^\top du$ taking values in the cone of positive definite matrices. Although the characterization of Gaussian Markov processes can be stated in terms of such increasing functions, we prefer the Choleskey decomposition as it will become computationally convenient later on.*

Remark 2. *In applications, data is typically sampled at discrete time points $t_1 < \ldots < t_J$. In this setting the full function $\gamma\colon [a,b] \to \mathbb{R}^{q\times q}$ is non-identifiable, so inference will be done for the terms $\gamma_1, \ldots, \gamma_J$ given by the Cholesky decompositions*

$$\gamma_1\gamma_1^\top = \gamma_0\gamma_0^\top + \int_a^{t_1} \gamma(u)\gamma(u)^\top \, du, \qquad \gamma_j\gamma_j^\top \overset{j\geq 1}{=} \int_{t_{j-1}}^{t_j} \gamma(u)\gamma(u)^\top \, du.$$

From this interpolation may be done to other time points $t \in [a,b]$.

Remark 3. *The bilinear structure of scores and loadings impose an over-parametrization. In the multivariate setting this over-parametrization is conveniently resolved by assuming that the coordinates of the score vector are independent with a standard normal distribution, i.e. $Z \sim \mathcal{N}(0, I_q)$. In the functional setting it is no longer natural to assume independence between the coordinates of the score process. However, we may still assume that $\gamma_1 = I_q$.*

The last thing that remains in order to specify what we will understand by functional probabilistic PCA, PLS, and CCA, and by functional FA, is to specify the variance structure of the error terms. For the error terms we will assume independent and constant variance $\Upsilon \in \mathbb{R}^{D\times D}$ given by Eq. (3). Furthermore, to emphasize the underlying Markov assumption we propose the following names and acronyms *Markov Multivariate Functional Principal Component Analysis* (MMF-PCA), *Markov Multivariate Functional Partial Least Squares* (MMF-PLS), *Markov Multivariate Functional Factor Analysis* (MMF-FA), and *Markov Multivariate Functional Canonical Correlation Analysis* (MMF-CCA).

3 Factorizable matrices

The probabilistic models introduced in Section 2.1 are special cases of Gaussian state space models. And the variance-covariance matrices for Gaussian state space models are exactly the positive definite matrices with a factorizable structure in the sense of [13]. Motivated by this we will parametrize lower triangular block matrices with a factorizable structure, which will be called $\mathrm{L}_{\mathrm{mat}}$-matrices. In Section 3.1 we develop a so-called Cholesky calculus for such matrices. By this name we simply mean a matrix calculus that works based on the Cholesky decomposition of positive semi-definite matrices S, i.e. a lower triangular matrix L such that $S = LL^\top$.

The Cholesky calculus has linear computational complexity in the length of the time series, like the well-known Kalman smoother from time series analysis. Seen from this perspective the Cholesky calculus simply provides an alternative computational framework for doing likelihood inference for Gaussian state space models. However, we prefer to use the Cholesky calculus for three reasons; it directly links to continuous time Gaussian processes via the factorizable structures, it is easily implemented, and being an ordinary matrix calculus it is more transparent and versatile. As an example of the latter we in Section 4 implement the combination of functional and multivariate observations.

3.1 Cholesky calculus

Let dimensions $d_1, \ldots, d_p \in \mathbb{N}$, $b_1, \ldots, b_p \in \mathbb{N}$, and $q \in \mathbb{N}$ be given. For parameters $\alpha = \{\alpha_j\}_{j=1}^p \in \oplus_{j=1}^p \mathbb{R}^{d_j\times b_j}$, $\beta = \{\beta_j\}_{j=1}^p \in \oplus_{j=1}^p \mathbb{R}^{d_j\times q}$, $\gamma = \{\gamma_j\}_{j=1}^p \in \oplus_{j=1}^p \mathbb{R}^{b_j\times q}$ we introduce the lower triangular block matrix $\mathrm{L}_{\mathrm{mat}}(\alpha,\beta,\gamma)$ by

$$\mathrm{L}_{\mathrm{mat}}(\alpha,\beta,\gamma) = \left\{1_{i=j}\alpha_j + 1_{i>j}\beta_i\gamma_j^\top\right\}_{i,j=1}^p \in \mathbb{R}^{(\sum_{j=1}^p d_j)\times(\sum_{j=1}^p b_j)}.$$

The remaining of this section collects results providing a toolbox for matrix manipulations and computations within the class of $\mathrm{L}_{\mathrm{mat}}$-matrices. These computations have linear complexity in the length p of the time series, and use standard matrix computations in dimension d_j of

the response. Thus, the computations are very efficient for long time series in relatively low dimensions. The results are formulated in terms of parameters

$$\alpha = \{\alpha_j\}_{j=1}^p \in \oplus_{j=1}^p \mathbb{R}^{d_j \times b_j}, \quad \beta = \{\beta_j\}_{j=1}^p \in \oplus_{j=1}^p \mathbb{R}^{d_j \times q}, \quad \gamma = \{\gamma_j\}_{j=1}^p \in \oplus_{j=1}^p \mathbb{R}^{b_j \times q},$$
$$\tilde{\alpha} = \{\tilde{\alpha}_j\}_{j=1}^p \in \oplus_{j=1}^p \mathbb{R}^{\tilde{d}_j \times \tilde{b}_j}, \quad \tilde{\beta} = \{\tilde{\beta}_j\}_{j=1}^p \in \oplus_{j=1}^p \mathbb{R}^{\tilde{d}_j \times \tilde{q}}, \quad \tilde{\gamma} = \{\tilde{\gamma}_j\}_{j=1}^p \in \oplus_{j=1}^p \mathbb{R}^{\tilde{b}_j \times \tilde{q}}.$$

If the dimensions of the stacked matrices comply with a given matrix operation, then this operation is defined element-wise in the obvious way, e.g. $\beta\gamma^\top = \{\beta_j\gamma_j^\top\}_{j=1}^p \in \oplus_{j=1}^p \mathbb{R}^{d_j \times b_j}$. Similarly, if the row dimensions match up, then element-wise column concatenation $\beta|\tilde{\beta}$ is defined by

$$\beta|\tilde{\beta} = \big\{\beta_j|\tilde{\beta}_j\big\}_{j=1}^p \in \oplus_{j=1}^p \mathbb{R}^{d_j \times (q+\tilde{q})}.$$

In the arithmetic hierarchy, column concatenation is performed after multiplication and addition unless specified otherwise by inserting parenthesis. The definition of the $\mathrm{L}_{\mathrm{mat}}$-matrices implies that $\mathrm{L}_{\mathrm{mat}}(\alpha, 0, 0)$ is block diagonal, and for matrices with block-wise conformable dimensions we immediately have multiplication formulae like

$$\mathrm{L}_{\mathrm{mat}}(\alpha, 0, 0)\, \mathrm{L}_{\mathrm{mat}}(\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}) = \mathrm{L}_{\mathrm{mat}}(\alpha\tilde{\alpha}, \alpha\tilde{\beta}, \tilde{\gamma}).$$

The following proposition motivates the naming *Cholesky calculus*. Thus, the proposition provides a recursive computation for the Cholesky decomposition of sum of positive definite matrices. As a consequence of this we can keep all computations within the class of lower triangular factorizable matrices.

Proposition 1. *Assume that α_j is invertible for every j. Then we have*

$$L_{mat}(\alpha\alpha^\top, 0, 0) + L_{mat}(\beta\gamma^\top, \beta, \gamma)\, L_{mat}(\beta\gamma^\top, \beta, \gamma)^\top = L_{mat}(\alpha\delta, \beta, \tilde{\gamma})\, L_{mat}(\alpha\delta, \beta, \tilde{\gamma})^\top,$$

where recursively for $j = 1, \ldots, p$, with $u_0 = \tilde{u}_0 = 0_{q\times q} \in \mathbb{R}^{q\times q}$,

$$u_j = u_{j-1} + \gamma_j^\top \gamma_j,$$
$$\delta_j = cholesky\Big(I_{d_j} + \alpha_j^{-1}\beta_j\big(u_j - \tilde{u}_{j-1}\big)\beta_j^\top \alpha_j^{-1,\top}\Big), \qquad \tilde{\gamma}_j = \delta_j^{-1}\alpha_j^{-1}\beta_j\big(u_j - \tilde{u}_{j-1}\big),$$
$$\tilde{u}_j = \tilde{u}_{j-1} + \tilde{\gamma}_j^\top \tilde{\gamma}_j.$$

Similarly,

$$L_{mat}(\alpha^\top\alpha, 0, 0) + L_{mat}(\beta\gamma^\top, \beta, \gamma)^\top\, L_{mat}(\beta\gamma^\top, \beta, \gamma) = L_{mat}(\zeta\alpha, \tilde{\beta}, \gamma)^\top\, L_{mat}(\zeta\alpha, \tilde{\beta}, \gamma),$$

where recursively for $j = p, \ldots, 1$, with $v_{p+1} = \tilde{v}_{p+1} = 0_{q\times q} \in \mathbb{R}^{q\times q}$,

$$v_j = v_{j+1} + \beta_j^\top \beta_j,$$
$$\zeta_j = \Big(cholesky\Big(I_{d_j} + \alpha_j^{-1,\top}\gamma_j\big(v_j - \tilde{v}_{j+1}\big)\gamma_j^\top \alpha_j^{-1}\Big)\Big)^\top, \quad \tilde{\beta}_j = \zeta_j^{-1,\top}\alpha_j^{-1,\top}\gamma_j\big(v_j - \tilde{v}_{j+1}\big),$$
$$\tilde{v}_j = \tilde{v}_{j+1} + \tilde{\beta}_j^\top \tilde{\beta}_j.$$

In these recursions $u_j - \tilde{u}_j$ and $v_j - \tilde{v}_j$ are positive semidefinite for every j. In particular, δ_j and ζ_j are well defined and invertible for every j.

Proof. Define $u_0 = \tilde{u}_0 = 0_{q\times q}$ and

$$u_j = \sum_{k=1}^{j} \gamma_k^\top \gamma_k, \qquad \tilde{u}_j = \sum_{k=1}^{j} \tilde{\gamma}_k^\top \tilde{\gamma}_k.$$

Identifying the (j,j)'th, respectively, the $(j,j+1)$'th block elements of the left and right hand side of the first equation give

$$\begin{aligned}
\alpha_j\alpha_j^\top + \beta_j u_j \beta_j^\top &= \alpha_j\alpha_j^\top + \beta_j\Big(\sum_{k=1}^{j}\gamma_k^\top\gamma_k\Big)\beta_j^\top \\
&= \alpha_j\delta_j\delta_j^\top\alpha_j + \beta_j\Big(\sum_{k=1}^{j-1}\tilde{\gamma}_k^\top\tilde{\gamma}_k\Big)\beta_j^\top = \alpha_j\delta_j\delta_j^\top\alpha_j^\top + \beta_j\tilde{u}_{j-1}\beta_j^\top, \\
\beta_j u_j \beta_{j+1}^\top &= \beta_j\Big(\sum_{k=1}^{j}\gamma_k^\top\gamma_k\Big)\beta_{j+1}^\top \\
&= \alpha_j\delta_j\tilde{\gamma}_j\beta_{j+1}^\top + \beta_j\Big(\sum_{k=1}^{j-1}\tilde{\gamma}_k^\top\tilde{\gamma}_k\Big)\beta_{j+1}^\top = \alpha_j\delta_j\tilde{\gamma}_j\beta_{j+1}^\top + \beta_j\tilde{u}_{j-1}\beta_{j+1}^\top,
\end{aligned}$$

which implies the formulae for δ_j and $\tilde{\gamma}_j$. For the second equation we define $v_{p+1} = \tilde{v}_{p+1} = 0_{q\times q}$ and

$$v_j = \sum_{k=j}^{p} \beta_k^\top \beta_k, \qquad \tilde{v}_j = \sum_{k=j}^{p} \tilde{\beta}_k^\top \tilde{\beta}_k.$$

Identifying the (j,j)'th, respectively, the $(j,j-1)$'th block elements of the left and right hand side of the first equation give

$$\begin{aligned}
\alpha_j^\top\alpha_j + \gamma_j v_j \gamma_j^\top &= \alpha_j^\top\alpha_j + \gamma_j\Big(\sum_{k=j}^{p}\beta_k^\top\beta_k\Big)\gamma_j^\top \\
&= \alpha_j^\top\zeta_j^\top\zeta_j\alpha_j + \gamma_j\Big(\sum_{k=j+1}^{p}\tilde{\beta}_k^\top\tilde{\beta}_k\Big)\gamma_j^\top = \alpha_j^\top\zeta_j^\top\zeta_j\alpha_j + \gamma_j\tilde{v}_{j+1}\gamma_j^\top, \\
\gamma_j v_j \gamma_{j-1}^\top &= \gamma_j\Big(\sum_{k=j}^{p}\beta_k^\top\beta_k\Big)\gamma_{j-1}^\top \\
&= \alpha_j^\top\zeta_j^\top\tilde{\beta}_j\gamma_{j-1}^\top + \gamma_j\Big(\sum_{k=j+1}^{p}\tilde{\beta}_k^\top\tilde{\beta}_k\Big)\gamma_{j-1}^\top = \alpha_j^\top\zeta_j^\top\tilde{\beta}_j\gamma_{j-1}^\top + \gamma_j\tilde{v}_{j+1}\gamma_{j-1}^\top,
\end{aligned}$$

which implies the formulae for ζ_j and $\tilde{\beta}_j$.

The formulae stated above require that δ_j and ζ_j are well defined and invertible for every j. This follows if $u_j - \tilde{u}_j$ and $v_j - \tilde{v}_j$ are positive semidefinite for every j. Since the formulae are

derived recursively we are allowed to prove this by induction using the stated formulae. Below we do this for the u's, and the proof for the v's is similar going backwards from $j = p+1$ to $j = 1$.

We have that $u_0 - \tilde{u}_0 = 0_{q\times q}$ is positive semidefinite. Now let j be given and assume that $u_{j-1} - \tilde{u}_{j-1}$ is positive semidefinite. Then $u_j - \tilde{u}_{j-1}$ also is positive semidefinite. Let $(u_j - \tilde{u}_{j-1})^{1/2}$ be the positive semidefinite square root, and let $UDV^\top$ be the singular value decomposition of $\beta_j(u_j - \tilde{u}_{j-1})^{1/2}$. Then we have

$$\begin{aligned}
\tilde{\gamma}_j^\top \tilde{\gamma}_j &= (u_j - \tilde{u}_{j-1})\beta_j^\top \alpha_j^{-1,\top}\delta_j^{-1,\top}\delta_j^{-1}\alpha_j^{-1}\beta_j(u_j - \tilde{u}_{j-1}) \\
&= (u_j - \tilde{u}_{j-1})^{1/2} V D^\top U^\top \alpha_j^{-1,\top}\Big(\mathrm{I}_{d_j} + \alpha_j^{-1}UDV^\top VD^\top U^\top \alpha_j^{-1,\top}\Big)^{-1}\alpha_j^{-1}UDV^\top (u_j - \tilde{u}_{j-1})^{1/2} \\
&= (u_j - \tilde{u}_{j-1})^{1/2} V D^\top \Big(U^\top \alpha_j \alpha_j^\top U + DD^\top\Big)^{-1} DV^\top (u_j - \tilde{u}_{j-1})^{1/2} \\
&\leq u_j - \tilde{u}_{j-1},
\end{aligned}$$

where the inequality is in the sense of positive semidefinite matrices. This gives $\tilde{u}_j = \tilde{u}_{j-1} + \tilde{\gamma}_j^\top \tilde{\gamma}_j \leq u_j$ as required for the induction step. □

Remark 4. *If Proposition 1 is stated in terms of α^{-1}, then we see that this also works in a limit with singular α_j's. This will be used later when computing the conditional variances in a Gaussian state space model.*

The following proposition shows that the class of lower triangular factorizable matrices constitutes a matrix algebra. We also show how the application of this matrix algebra on stacked matrices $x = \{x_j\}_{j=1}^p \in \oplus_{j=1}^p \mathbb{R}^{b_j\times N}$ can be computed in linear time. Here $N \in \mathbb{N}$ can be interpreted as the number of observations.

Proposition 2. *For fixed dimensions $d_1, \ldots, d_p \in \mathbb{N}$ and $b_1, \ldots, b_p \in \mathbb{N}$, the class of lower triangular factorizable matrices with free $q \in \mathbb{N}$ is an additive matrix group with*

$$\begin{aligned}
-L_{mat}(\alpha, \beta, \gamma) &= L_{mat}(-\alpha, -\beta, \gamma), \\
L_{mat}(\alpha, \beta, \gamma) + L_{mat}(\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}) &= L_{mat}(\alpha + \tilde{\alpha}, \beta|\tilde{\beta}, \gamma|\tilde{\gamma}).
\end{aligned}$$

The left application of $L_{mat}(\alpha, \beta, \gamma)$ on $x = \{x_j\}_{j=1}^p \in \oplus_{j=1}^p \mathbb{R}^{b_j\times N}$, that is

$$y = \{y_j\}_{j=1}^p = L_{mat}(\alpha, \beta, \gamma)x \in \oplus_{j=1}^p \mathbb{R}^{d_j\times N},$$

can be computed recursively for $j = 1, \ldots, p$ by

$$u_0 = 0_{q\times N}, \qquad y_j = \alpha_j x_j + \beta_j u_{j-1}, \qquad u_j = u_{j-1} + \gamma_j^\top x_j.$$

The right application of $L_{mat}(\alpha, \beta, \gamma)$ on $x = \{x_j\}_{j=1}^p \in \oplus_{j=1}^p \mathbb{R}^{d_j\times N}$, that is

$$y^\top = \{y_j^\top\}_{j=1}^p = x^\top L_{mat}(\alpha, \beta, \gamma) \in \oplus_{j=1}^p \mathbb{R}^{N\times b_j},$$

can be computed recursively for $j = p, \dots, 1$ *by*

$$v_{p+1} = 0_{q\times N}, \qquad y_j = \alpha_j^\top x_j + \gamma_j v_{j+1}, \qquad v_j = v_{j+1} + \beta_j^\top x_j.$$

In the remaining statements we assume that $b_j = d_j$ *for* $j = 1, \dots, p$*. Then the matrix group is closed under matrix multiplication. In particular,* $L_{mat}(\alpha, \beta, \gamma)\, L_{mat}(\tilde\alpha, \tilde\beta, \tilde\gamma)$ *equals*

$$L_{mat}\left(\alpha\tilde\alpha, \beta|\alpha\tilde\beta - \beta\gamma^\top\tilde\beta - \beta\, B(\gamma^\top\tilde\beta), \tilde\alpha^\top\gamma + \tilde\gamma\, B(\tilde\beta^\top\gamma)|\tilde\gamma\right),$$

where the backward cumulative sum operator B *is defined by*

$$B(z) = \left\{1_{j<p}\sum_{i=j+1}^{p} z_i\right\}_{j=1}^{p}, \quad z = \{z_j\}_{j=1}^p.$$

The matrix $L_{mat}(\alpha, \beta, \gamma)$ *is invertible if and only if* $\alpha_j \in \mathbb{R}^{d_j\times d_j}$ *is invertible for* $j = 1, \dots, p$*. In this case, the left inverse application of* $L_{mat}(\alpha, \beta, \gamma)$ *on* $y = \{y_j\}_{j=1}^p \in \oplus_{j=1}^p \mathbb{R}^{d_j\times N}$*, that is*

$$x = \{x_j\}_{j=1}^p = L_{mat}(\alpha, \beta, \gamma)^{-1} y \in \oplus_{j=1}^p \mathbb{R}^{d_j\times N},$$

can be computed recursively for $j = 1, \dots, p$ *by*

$$u_0 = 0_{q\times N}, \qquad x_j = \alpha_j^{-1}\big(y_j - \beta_j u_{j-1}\big), \qquad u_j = u_{j-1} + \gamma_j^\top x_j.$$

And the right inverse application of $L_{mat}(\alpha, \beta, \gamma)$ *on* $y = \{y_j\}_{j=1}^p \in \oplus_{j=1}^p \mathbb{R}^{d_j\times N}$*, that is*

$$x^\top = \{x_j^\top\}_{j=1}^p = y^\top\, L_{mat}(\alpha, \beta, \gamma)^{-1} \in \oplus_{j=1}^p \mathbb{R}^{N\times d_j},$$

can be computed recursively for $j = p, \dots, 1$ *by*

$$v_{p+1} = 0_{q\times N}, \qquad x_j = \alpha_j^{-1,\top}\big(y_j - \gamma_j v_{j+1}\big), \qquad v_j = v_{j+1} + \beta_j^\top x_j.$$

Moreover, if we also have that $I_q - \beta_j^\top \alpha_j^{-1,\top}\gamma_j$ *is invertible for* $j = 2, \dots, p-1$*, then the matrix inverse is also a lower triangular factorizable matrix. In particular,* $L_{mat}(\alpha, \beta, \gamma)^{-1} = L_{mat}(\alpha^{-1}, \tilde\beta, \tilde\gamma)$*, where* $u_1 = I_q$ *and recursively for* $j = 2, \dots, p$*,*

$$\tilde\gamma_{j-1} = \alpha_{j-1}^{-1,\top}\gamma_{j-1}u_{j-1}^{-1}, \qquad u_j = u_{j-1}(I_q - \beta_j^\top\alpha_j^{-1,\top}\gamma_j), \qquad \tilde\beta_j = -\alpha_j^{-1}\beta_j u_{j-1}^\top.$$

Proof. The formulae for the additive structure follow by straightforward calculations. To compute the multiplicative structure we start by noting the immediate formulae

$$\begin{aligned}
\mathrm{L}_{\mathrm{mat}}(\alpha, 0, 0)\,\mathrm{L}_{\mathrm{mat}}(\tilde\alpha, 0, 0) &= \mathrm{L}_{\mathrm{mat}}(\alpha\tilde\alpha, 0, 0),\\
\mathrm{L}_{\mathrm{mat}}(\alpha, 0, 0)\,\mathrm{L}_{\mathrm{mat}}(0, \tilde\beta, \tilde\gamma) &= \mathrm{L}_{\mathrm{mat}}(0, \alpha\tilde\beta, \tilde\gamma),\\
\mathrm{L}_{\mathrm{mat}}(0, \beta, \gamma)\,\mathrm{L}_{\mathrm{mat}}(\tilde\alpha, 0, 0) &= \mathrm{L}_{\mathrm{mat}}(0, \beta, \tilde\alpha^\top\gamma).
\end{aligned}$$

Moreover, we have that $\mathrm{L}_{\mathrm{mat}}(0,\beta,\gamma)\,\mathrm{L}_{\mathrm{mat}}(0,\tilde{\beta},\tilde{\gamma})$ equals

$$\mathrm{L}_{\mathrm{mat}}(\beta\gamma^\top,\beta,\gamma)\,\mathrm{L}_{\mathrm{mat}}(0,\tilde{\beta},\tilde{\gamma}) - \mathrm{L}_{\mathrm{mat}}(0,\beta\gamma^\top\tilde{\beta},\tilde{\gamma}),$$

where the first term applied on $x \in \mathbb{R}^{(d\times q)}$ may be rewritten as

$$\begin{aligned}
&\mathrm{L}_{\mathrm{mat}}(\beta\gamma^\top,\beta,\gamma)\,\mathrm{L}_{\mathrm{mat}}(0,\tilde{\beta},\tilde{\gamma})x \\
&\quad= \left\{\beta_i \sum_{j=1}^{i} \gamma_j^\top \tilde{\beta}_j 1_{j>1} \sum_{k=1}^{j-1} \tilde{\gamma}_k^\top x_k\right\}_{i=1}^{p} \\
&\quad= \left\{1_{i>1}\beta_i \sum_{k=1}^{i-1}\sum_{j=k+1}^{p} \gamma_j^\top \tilde{\beta}_j \tilde{\gamma}_k^\top x_k\right\}_{i=1}^{p} - \left\{1_{1<i<p}\,\beta_i \sum_{k=1}^{i-1}\sum_{j=i+1}^{p} \gamma_j^\top \tilde{\beta}_j \tilde{\gamma}_k^\top x_k\right\}_{i=1}^{p} \\
&\quad= \mathrm{L}_{\mathrm{mat}}\left(0,\beta,\tilde{\gamma}\,\mathrm{B}(\tilde{\beta}^\top\gamma)\right)x - \mathrm{L}_{\mathrm{mat}}\left(0,\beta\,\mathrm{B}(\gamma^\top\tilde{\beta}),\tilde{\gamma}\right)x.
\end{aligned}$$

Here the second equality sign follows by interchanging and rewriting the double sums appearing in the cumulative sums. Thus, using linearity we have that $\mathrm{L}_{\mathrm{mat}}(\alpha,\beta,\gamma)\,\mathrm{L}_{\mathrm{mat}}(\tilde{\alpha},\tilde{\beta},\tilde{\gamma})$ equals

$$\mathrm{L}_{\mathrm{mat}}\left(\alpha\tilde{\alpha},\beta|\alpha\tilde{\beta} - \beta\gamma^\top\tilde{\beta} - \beta\,\mathrm{B}(\gamma^\top\tilde{\beta}),\tilde{\alpha}^\top\gamma + \tilde{\gamma}\,\mathrm{B}(\tilde{\beta}^\top\gamma)|\tilde{\gamma}\right).$$

For the left application $y = \{y_j\}_{j=1}^p = \mathrm{L}_{\mathrm{mat}}(\alpha,\beta,\gamma)\{x_j\}_{j=1}^p$ we define $u_0 = 0_{q\times q}$ and $u_j = \sum_{k=1}^{j}\gamma_k^\top x_k$. This gives the stated recursions, that is

$$y_j = \alpha_j x_j + 1_{j>1}\sum_{k=1}^{j-1}\beta_j\gamma_k^\top x_k = \alpha_j x_j + \beta_j u_{j-1}, \qquad\qquad u_j = u_{j-1} + \gamma_j^\top x_j.$$

In case that $b_j = d_j$ and α_j is invertible for all $j = 1,\ldots,p$ the same definition of the u_j's gives

$$x_j = \alpha_j^{-1}\left(y_j - \beta_j u_{j-1}\right), \qquad\qquad u_j = u_{j-1} + \gamma_j^\top x_j.$$

For the right application $y^\top = \{y_j^\top\}_{j=1}^p = \{x_j^\top\}_{j=1}^p\,\mathrm{L}_{\mathrm{mat}}(\alpha,\beta,\gamma)$ we define $v_{p+1}^\top = 0_{q\times q}$ and $v_j^\top = \sum_{k=j}^{p} x_k^\top\beta_k$. This gives the stated recursions, that is

$$y_j^\top = x_j^\top\alpha_j + 1_{j<p}\sum_{k=j+1}^{p} x_k^\top\beta_k\gamma_j^\top = x_j^\top\alpha_j + v_{j+1}^\top\gamma_j^\top, \qquad\qquad v_j^\top = v_{j+1}^\top + x_j^\top\beta_j.$$

In case that $b_j = d_j$ and α_j is invertible for all $j = 1,\ldots,p$ the same definition of the v_j's gives

$$x_j^\top = \left(y_j - v_{j+1}^\top\gamma_j^\top\right)\alpha_j^{-1}, \qquad\qquad v_j^\top = v_{j+1}^\top + x_j^\top\beta_j.$$

Concerning the multiplicative inverse, we observe that matrix elements in $\alpha = \{\alpha_j\}_{j=1}^p \in \oplus_{j=1}^p\mathbb{R}^{d_j\times d_j}$ are the diagonal elements in the block matrix representation of $\mathrm{L}_{\mathrm{mat}}(\alpha,\beta,\gamma)$.

Hence $\mathrm{L}_{\mathrm{mat}}(\alpha,\beta,\gamma)$ is invertible if and only if α_j are invertible for all $j=1,\dots,p$. However, to show that the inverse also is a lower triangular factorizable matrix and to derive the inversion formula requires more work.

Using [15, Theorem 10] the matrix inverse can also be stated in terms of a product integral solving an inhomogeneous Volterra integral equation. This leads to the representation

$$\mathrm{L}_{\mathrm{mat}}(\alpha,\beta,\gamma)^{-1} = \mathrm{L}_{\mathrm{mat}}(\alpha^{-1}, -\alpha^{-1}\beta u_\dagger^\top, \alpha^{-1,\top}\gamma u^{-1}),$$

where $u_\dagger = \{1_{i=1}\cdot \mathbb{I}_q + 1_{i>1}\cdot u_{i-1}\}_{i=1}^p$, and

$$u = \left\{ \prod_{j=1}^{i} \left(\mathbb{I}_q - \beta_j^\top \alpha_j^{-1,\top} \gamma_j \right) \right\}_{i=1}^p \in (\mathbb{R}^{q\times q})^p$$

with factors ordered from the right to the left with increasing indices. This representation is easily restated as the recursive formula stated in the proposition. □

The following two propositions are useful for computing conditional variances.

Proposition 3. *Suppose that $b_j = d_j$ for $j=1,\dots,p$. Then the i'th diagonal block element in the block matrix $L_{mat}(\alpha,\beta,\gamma)^{-1}\, L_{mat}(\alpha,\beta,\gamma)^{-1,\top}$ is given by*

$$\alpha_i^{-1}\alpha_i^{-1,\top} + \alpha_i^{-1}\beta_i\xi_{i-1}\beta_i^\top\alpha_i^{-1,\top},$$

where $\xi_0 = 0_{q\times q}$, and recursively for $i=1,\dots,p-1$,

$$\xi_i = \big(\, I_q - \gamma_i^\top\alpha_i^{-1}\beta_i\big)\xi_{i-1}\big(\, I_q - \gamma_i^\top\alpha_i^{-1}\beta_i\big)^\top + \gamma_i^\top\alpha_i^{-1}\alpha_i^{-1,\top}\gamma_i.$$

Proof. For every $i=1,\dots,p$ let $y^{[i]} = \{1_{j=i}\,\mathrm{I}_{d_j}\}_{j=1}^p$ and consider the corresponding right inverse application of $\mathrm{L}_{\mathrm{mat}}(\alpha,\beta,\gamma)$, that is $x^{[i]} = \{x_j^{[i]}\}_{j=1}^p = \mathrm{L}_{\mathrm{mat}}(\alpha,\beta,\gamma)^{-1,\top} y^{[i]}$. Using that $x_j^{[i]} = 0_{d_j}$ for $j>i$ we have, that the i'th diagonal block element of $\mathrm{L}_{\mathrm{mat}}(\alpha,\beta,\gamma)^{-1}\,\mathrm{L}_{\mathrm{mat}}(\alpha,\beta,\gamma)^{-1,\top}$ is given by

$$\sum_{j=1}^{i} x_j^{[i],\top} x_j^{[i]}.$$

Proposition 2 states a formula for the right inverse application. The quantities $v_j^{[i]}$, here also indexed by i, are given by $v_j^{[i]} = 0_q$ for $j>i$ and

$$\begin{aligned} v_j^{[i]} &= v_{j+1}^{[i]} + \beta_j^\top x_j^{[i]} \\ &= v_{j+1}^{[i]} + \beta_j^\top\alpha_j^{-1,\top}\big(y_j^{[i]} - \gamma_j v_{j+1}^{[i]}\big) \\ &= \beta_j^\top\alpha_j^{-1,\top} y_j^{[i]} + \big(\,\mathrm{I}_q - \beta_j^\top\alpha_j^{-1,\top}\gamma_j\big)v_{j+1}^{[i]}. \end{aligned}$$

This gives $v_i^{[i]} = \beta_i^\top \alpha_i^{-1,\top}$, and for $j < i$ we have

$$v_j^{[i]} = \Big(\prod_{k=j}^{i-1} \big(\mathrm{I}_q - \beta_k^\top \alpha_k^{-1,\top} \gamma_k \big) \Big) \beta_i^\top \alpha_i^{-1,\top}.$$

This gives $x_i^{[i]} = \alpha_i^{-1,\top}$, and for $j < i$ we have

$$x_j^{[i]} = -\alpha_j^{-1,\top} \gamma_j v_{j+1}^{[i]} = -\alpha_j^{-1,\top} \gamma_j \Big(\prod_{k=j+1}^{i-1} \big(\mathrm{I}_q - \beta_k^\top \alpha_k^{-1,\top} \gamma_k \big) \Big) \beta_i^\top \alpha_i^{-1,\top}.$$

Introducing $\xi_0 = 0_q$, the definition below, and the associated recursion

$$\begin{aligned} \xi_{i-1} &\stackrel{\text{def}}{=} \sum_{j=1}^{i-1} \Big(\prod_{k=j+1}^{i-1} \big(\mathrm{I}_q - \beta_k^\top \alpha_k^{-1,\top} \gamma_k \big) \Big)^\top \gamma_j^\top \alpha_j^{-1} \alpha_j^{-1,\top} \gamma_j \Big(\prod_{k=j+1}^{i-1} \big(\mathrm{I}_q - \beta_k^\top \alpha_k^{-1,\top} \gamma_k \big) \Big) \\ &= \big(\mathrm{I}_q - \beta_{i-1}^\top \alpha_{i-1}^{-1,\top} \gamma_{i-1} \big)^\top \xi_{i-2} \big(\mathrm{I}_q - \beta_{i-1}^\top \alpha_{i-1}^{-1,\top} \gamma_{i-1} \big) + \gamma_{i-1}^\top \alpha_{i-1}^{-1} \alpha_{i-1}^{-1,\top} \gamma_{i-1} \end{aligned}$$

we recover the claimed formula for the i'th diagonal block element

$$\sum_{j=1}^{i} x_j^{[i],\top} x_j^{[i]} = \alpha_i^{-1} \beta_i \xi_{i-1} \beta_i^\top \alpha_i^{-1,\top} + \alpha_i^{-1} \alpha_i^{-1,\top}.$$

□

Proposition 4. *Suppose that $d_j = b_j = q$ for $j = 1, \ldots, p$, and let A be the lower triangular matrix $L_{mat}\big(\{I_q\}_{j=1}^p, \{I_q\}_{j=1}^p, \{I_q\}_{j=1}^p\big)$. Let $\xi_0 = 0_{q\times q}$, and let $\xi_i \in \mathbb{R}^{q\times q}$ be the i'th diagonal block element in the block matrix*

$$A\, L_{mat}(\alpha, \beta, \{I_q\}_{j=1}^p)^{-1}\, L_{mat}(\alpha, \beta, \{I_q\}_{j=1}^p)^{-1,\top} A^\top.$$

Then we have, recursively for $i = 1, \ldots, p$,

$$\xi_i = \big(I_q - \alpha_i^{-1} \beta_i \big) \xi_{i-1} \big(I_q - \alpha_i^{-1} \beta_i \big)^\top + \alpha_i^{-1} \alpha_i^{-1,\top}.$$

Proof. We start by rewriting $A\, \mathrm{L}_{\mathrm{mat}}(\alpha, \beta, \{\mathrm{I}_q\}_{j=1}^p)^{-1}\, \mathrm{L}_{\mathrm{mat}}(\alpha, \beta, \{\mathrm{I}_q\}_{j=1}^p)^{-1,\top} A^\top$ as

$$\Big(\mathrm{L}_{\mathrm{mat}}(\alpha, \beta, \{\mathrm{I}_q\}_{j=1}^p) A^{-1} \Big)^{-1} \Big(\mathrm{L}_{\mathrm{mat}}(\alpha, \beta, \{\mathrm{I}_q\}_{j=1}^p) A^{-1} \Big)^{-1,\top}.$$

It is easily verified that A^{-1} is a lower triangular block matrix with I_q's on the diagonal, $-\mathrm{I}_q$'s on the first lower diagonal, and $0_{q\times q}$'s elsewhere. This implies

$$\tilde{A} \stackrel{\text{def}}{=} \mathrm{L}_{\mathrm{mat}}\big(\alpha, \beta, \{\mathrm{I}_q\}_{j=1}^p\big) A^{-1} = \Big\{ 1_{j=i} \alpha_i + 1_{j=i-1} (\beta_i - \alpha_i) \Big\}_{i,j=1}^p.$$

The proof now continuous along the lines of the proof for Proposition 3 using $x^{[i]} = \tilde{A}^{-1,\top} y^{[i]}$ with $y^{[i]} = \{1_{j=i}\,\mathrm{I}_q\}_{j=1}^p$. The relation between $x^{[i]}$ and $y^{[i]}$ is given by $x_j^{[i]} = 0_q$ for $j > i$, and

$$y_j^{[i]} \overset{j\leq i}{=} \alpha_j^\top x_j^{[i]} + 1_{j<p}\big(\beta_{j+1} - \alpha_{j+1}\big)^\top x_{j+1}^{[i]}.$$

Solving this we find $x_i^{[i]} = \alpha_i^{-1,\top}$, and

$$\begin{aligned} x_j^{[i]} &\overset{j\leq i}{=} \alpha_j^{-1,\top}\Big(y_j^{[i]} - \big(\beta_{j+1} - \alpha_{j+1}\big)^\top x_{j+1}^{[i]}\Big) \qquad \overset{j\leq i}{=} \alpha_j^{-1,\top}\big(\alpha_{j+1} - \beta_{j+1}\big)^\top x_{j+1}^{[i]} \\ &\overset{j\leq i}{=} \alpha_j^{-1,\top} \prod_{k=j+1}^{i} \big(\mathrm{I}_q - \beta_k^\top \alpha_k^{-1,\top}\big). \end{aligned}$$

Thus, we have $\xi_1 = x_1^{[1],\top} x_1^{[1]} = \alpha_1^{-1}\alpha_1^{-1,\top}$ and the recursion for $i = 2, \ldots, p$,

$$\begin{aligned} \xi_i &= \sum_{j=1}^{i} x_j^{[i],\top} x_j^{[i]} \\ &= \sum_{j=1}^{i-1} \Big(\prod_{k=j+1}^{i} \big(\mathrm{I}_q - \beta_k^\top \alpha_k^{-1,\top}\big) \Big)^\top \alpha_j^{-1}\alpha_j^{-1,\top} \Big(\prod_{k=j+1}^{i} \big(\mathrm{I}_q - \beta_k^\top \alpha_k^{-1,\top}\big) \Big) + \alpha_i^{-1}\alpha_i^{-1,\top} \\ &= \big(\mathrm{I}_q - \alpha_i^{-1}\beta_i\big)\xi_{i-1}\big(\mathrm{I}_q - \alpha_i^{-1}\beta_i\big)^\top + \alpha_i^{-1}\alpha_i^{-1,\top}. \end{aligned}$$

□

4 Simultaneous modelling of phase and amplitude variation

In the following we develop a simultaneous model for phase and amplitude variation in multivariate functional data $X_n(t) \in \mathbb{R}^D$ sampled at J discrete time points $t_1 < \ldots < t_J$. Before describing the details of this construction in section 4.1 we first list the parameters used in the model, and the random variables used as the stochastic basis. In section 4.3 we derive formulae for conditional means and variances, and develop an EM-algorithm for estimation of model parameters.

The population mean for $X_n(t)$ is modelled via a continuously differentiable functional basis $\Phi\colon \mathbb{R} \to \mathbb{R}^{D\times K}$ and design matrices $\Xi_n \in \mathbb{R}^{K\times p}$, and parametrized by $\theta \in \mathbb{R}^p$. In applications Φ could be a spline or Fourier basis as appropriate.

Parameters Let $q \in \mathbb{N}$ be the dimension of latent stochastic processes describing phase and amplitude variation (we allow phase and amplitude variation to be correlated). Let there be given jointly independent random variables

$$U_{nj} \sim \mathcal{N}_D(0, \alpha\alpha^\top), \qquad V_{nj} \sim \mathcal{N}_q(0, \gamma_j^\top \gamma_j)$$

for $n = 1, \ldots, N$ and $j = 1, \ldots, J$. Here $\alpha \in \mathbb{R}^{D\times D}$ and $\gamma_j \in \mathbb{R}^{q\times q}$ provides Cholesky parametrizations of variance terms. Finally, the model also includes regression parameters and the latent variables, $\beta_j^{\text{phase}} \in \mathbb{R}^{1\times q}$, $\beta_j^{\text{amp}} \in \mathbb{R}^{D\times q}$.

For notational convenience we introduce the cumulated processes

$$Z_{nj} = \sum_{i=1}^{j} V_{ni} \in \mathbb{R}^q.$$

Population means are modelled through the spline basis Φ, the known design matrix $\Xi \in \mathbb{R}^{K\times p}$, and unknown parameters $\theta \in \mathbb{R}^p$, such that the mean signal at time t_j is

$$\Phi(t_j + \beta_j^{\text{phase}} Z_{nj})\Xi_n\theta$$

4.1 Phase and amplitude in functional data

Olsen et al. [20] propose a simultaneous modelling of phase and amplitude variation given observations of multivariate Gaussian processes at J prefixed time points $t_1 < \ldots < t_J$, that is,

$$\text{signal}\big(\text{phase}_n(t_j)\big) + \text{amplitude}_n(t_j) + \text{error}_{nj}$$

for $n = 1, \ldots, N$ and $j = 1, \ldots, J$. Here the random phase functions should be increasing, and in [20] these functions are modelled to be stochastically independent of the serially correlated amplitude deviations. However, Hadjipantelis et al. [22, 23] argue that the phase and the amplitude variation may be correlated in some situations. Below we propose a simultaneous model for phase and amplitude variation following the approach of [20], but which allow for correlated phase and amplitude variations. We propose to estimate in the model following the approach of [20]. That is, a first order approximation around the max-posterior of $\text{phase}_n(t_j)$, with alternating of updating the approximation and updating parameters using the EM algorithm. We remark that the MCA framework allow us to do computations in linear time in the number of data points NJ.

For each $n = 1, \ldots, N$ let there be given a vector $\tau_n = \{\tau_{nj}\}_{j=1}^J \in \mathbb{R}^J$. The vectors τ_n parametrize the predicted warping functions via deviations from the time points $t = \{t_j\}_{j=1}^J$, that is for the n'th curve the time point t_j is shifted to $t_j + \tau_{nj}$. Restrictions on the warping functions are reflected by restrictions on the τ_n's. Thus, that the slope of warping functions lies in the interval $[s_{\min}, s_{\max}]$ with $s_{\min} < 1 < s_{\max}$ correspond to the following box-constraints for $j = 1, \ldots, J-1$,

$$(s_{\min} - 1) \cdot (t_{j+1} - t_j) \leq \tau_{n,j+1} - \tau_{nj} \leq (s_{\max} - 1) \cdot (t_{j+1} - t_j).$$

The phase and the amplitude at time t_j for the n'th curve are given by

$$\text{phase} = t_j + \beta_{nj}^{\text{phase}} Z_{nj}, \qquad \text{amplitude} = \beta_{nj}^{\text{amp}} Z_{nj}.$$

Using the same Z allows for correlation between phase and amplitude; in particular phase and amplitude must be correlated if $q = 1$. However if we let

$$Z_{nj} = \begin{pmatrix} Z_{nj}^{\text{phase}} \\ Z_{nj}^{\text{amp}} \end{pmatrix}, \quad \beta_{nj}^{\text{phase}} = \begin{pmatrix} \tilde{\beta}_{nj}^{\text{phase}} & 0 \end{pmatrix}, \quad \beta_{nj}^{\text{amp}} = \begin{pmatrix} 0 & \tilde{\beta}_{nj}^{\text{amp}} \end{pmatrix}$$

with $\text{Cov}(Z_{nj}^{\text{phase}}, Z_{nj}^{\text{amp}}) = 0$, then phase and amplitude are independent.

Concerning the observation $X_n(t_j) = x_{nj}$ we invoke the first order Taylor approximation of the signal part around $t_j + \tau_{nj}$, which gives

$$\begin{aligned} X_n(t_j) &= \Phi(t_j + \beta_j^{\text{phase}} Z_{nj})\Xi_n\theta + \beta_j^{\text{amp}} Z_{nj} + U_{nj} \\ &\approx \Phi(t_j + \tau_{nj})\Xi_n\theta + \dot{\Phi}(t_j + \tau_{nj})(\beta_j^{\text{phase}} Z_{nj}^{\text{phase}} - \tau_{nj})\Xi_n\theta + \beta_j^{\text{amp}} Z_{nj}^{\text{amp}} + U_{nj}, \end{aligned}$$

where $\dot{\Phi}$ is the derivative of Φ.

This approximation of the multivariate functional data gives:

$$\begin{aligned} X_{nj}^{\tau_n} &= \Phi(t_j + \tau_{nj})\Xi_n\theta + \dot{\Phi}(t_j + \tau_{nj})\Xi_n\theta\big(\beta_j^{\text{phase}} Z_{nj} - \tau_{nj}\big) + \beta_j^{\text{amp}} Z_{nj} + U_{nj} \\ &= [\Phi(t_j + \tau_{nj}) - \tau_{nj}\dot{\Phi}(t_j + \tau_{nj})]\Xi_n\theta + \big(\dot{\Phi}(t_j + \tau_{nj})\Xi_n\theta_X\beta_j^{\text{phase}} + \beta_j^{\text{amp}}\big)Z_{nj} + U_{nj} \end{aligned}$$

4.2 Prediction of phase deviation

The phase deviation parameters τ_{nj} are found by minimising the posterior likelihood, subject to the previously introduced box-constraints.

By concatenating the observations and the phase deviations we can formulate the posterior likelihood in an MCA framework for a given τ:

$$\tilde{X}_{nj} = \begin{pmatrix} X_{nj} \\ \tau_{nj} \end{pmatrix}, \quad \tilde{\beta}_j = \begin{pmatrix} \beta_j^{\text{amp}} & \beta_j^{\text{phase}} \end{pmatrix}, \quad \tilde{\gamma}_j = \gamma_j,$$

$$\tilde{\alpha}\tilde{\alpha}^\top = \begin{pmatrix} \alpha\alpha^\top & 0 \\ 0 & 0 \end{pmatrix}, \quad \text{E}[\tilde{X}_{nj}] = \begin{pmatrix} \Phi(t_j + \tau_{nj})\Xi_n\theta \\ 0 \end{pmatrix} \quad (5)$$

The likelihood of this can be calculated in linear time using relevant propositions. Note that this is a non-linear optimization problem as we optimize in τ.

4.3 Conditional means and variances

In order to apply the apply EM algorithm, expressions for conditional means and variances are needed. As the approximation is a linear Gaussian model, this can be done only using linear algebra.

Theorem 1. *Assume that α is invertible. Then the random matrices V_n and $X_n^{\tau_n}$ have a joint multivariate normal distribution.*

The mean structure is given by

$$E[V_{nj}] = 0_q, \qquad E[X_n^{\tau_n}] = \Xi_n^{\tau_n}\theta$$

with parameter $\theta \in \mathbb{R}^p$ and design matrix $\Xi_n^{\tau_n} \in \oplus_{j=1}^J \mathbb{R}^{D\times p}$ given by

$$\Xi_{nj}^{\tau_n} = \big(\Phi(t_j+\tau_{nj})\Xi_n - \dot\Phi(t_j+\tau_{nj})\Xi_n\tau_{nj}\big)$$

The covariance structure is given by

$$\begin{aligned}
\mathrm{Var}(V_n) &= L_{mat}(\gamma^\top\gamma, 0, 0),\\
\mathrm{Cov}(X_n^{\tau_n}, V_n) &= L_{mat}(\beta\gamma^\top\gamma, \beta, \gamma^\top\gamma),\\
\mathrm{Var}(X_n^{\tau_n}) &= L_{mat}(\alpha\delta, \beta, \tilde\gamma)\, L_{mat}(\alpha\delta, \beta, \tilde\gamma)^\top,
\end{aligned}$$

where we for notational convenience have suppressed the dependency of β, $\tilde\gamma$ and δ on θ, τ_n and n in the notation. Here the Cholesky decomposition $\alpha = \{\alpha_j\}_{j=1}^J \in \oplus_{j=1}^J \mathbb{R}^{D\times D}$ of the error variance is given by

$$\alpha_j = \alpha$$

the random effect design matrix $\beta = \{\beta_j\}_{j=1}^J \in \oplus_{j=1}^J \mathbb{R}^{D\times q}$ parametrized by $\beta_j^{phase} \in \mathbb{R}^{1\times q}$ and $\beta_j^{amp} \in \mathbb{R}^{D\times q}$ for $j = 1, \ldots, J$ is given by

$$\beta_j = \dot\Phi(t_j+\tau_{nj})\Xi_n\theta\beta_j^{phase} + \beta_j^{amp}.$$

Furthermore, $\delta = \{\delta_j\}_{j=1}^J \in \otimes_{j=1}^J \mathbb{R}^{D\times D}$ and $\tilde\gamma = \{\tilde\gamma_j\}_{j=1}^J \in \otimes_{j=1}^J \mathbb{R}^{D\times q}$ are given by the forward computations initiated by $u_0 = \tilde u_0 = 0_{q\times q}$ and recursively for $j = 1, \ldots, J$,

$$\begin{aligned}
u_j &= u_{j-1} + \gamma_j^\top\gamma_j,\\
\delta_j &= cholesky\Big(I_D + \alpha_j^{-1}\beta_j(u_j - \tilde u_{j-1})\beta_j^\top\alpha_j^{-1,\top}\Big),\\
\tilde\gamma_j &= \delta_j^{-1}\alpha_j^{-1}\beta_j(u_j - \tilde u_{j-1}),\\
\tilde u_j &= \tilde u_{j-1} + \tilde\gamma_j^\top\tilde\gamma_j.
\end{aligned}$$

Proof. The stochastic process $X_n^{\tau_n}$ can be represented as a Gaussian state space model

$$X_{nj}^{\tau_n} = \Xi_{nj}^{\tau_n}\theta + \beta_j Z_{nj} + U_{nj}. \tag{6}$$

The joint multivariate normal distribution follows since the random matrices V_n and $X_n^{\tau_n}$ are build by linear operations from the same underlying normal random variables. Hereby the

mean structure follows immediately, and concerning the covariance structure we have

$$\begin{aligned}
\mathrm{Var}(V_n) &= \Big\{1_{i=j}\gamma_i^\top\gamma_i\Big\}_{i,j=1}^J = \mathrm{L}_{\mathrm{mat}}(\gamma^\top\gamma, 0, 0),\\
\mathrm{Cov}(X_n^{\tau_n}, V_n) &= \Big\{1_{i\geq j}\beta_i\gamma_j^\top\gamma_j\Big\}_{i,j=1}^J = \mathrm{L}_{\mathrm{mat}}(\beta\gamma^\top\gamma, \beta, \gamma^\top\gamma),\\
\mathrm{Var}(X_n^{\tau_n}) &= \left\{1_{i=j}\alpha_i\alpha_i^\top + \beta_i\Big(\sum_{k=1}^{\min\{i,j\}}\gamma_k^\top\gamma_k\Big)\beta_j^\top\right\}_{i,j=1}^J\\
&= \mathrm{L}_{\mathrm{mat}}(\alpha\alpha^\top, 0, 0) + \mathrm{L}_{\mathrm{mat}}(\beta\gamma^\top, \beta, \gamma)\,\mathrm{L}_{\mathrm{mat}}(\beta\gamma^\top, \beta, \gamma)^\top\\
&= \mathrm{L}_{\mathrm{mat}}(\alpha\delta, \beta, \tilde\gamma)\,\mathrm{L}_{\mathrm{mat}}(\alpha\delta, \beta, \tilde\gamma)^\top,
\end{aligned}$$

where the formulae for δ and $\tilde\gamma$ follow by Proposition 1. □

Using the Cholesky decomposition we have the following result:

Corollary 1. *The log-likelihood l_n given the observation of $X_n^{\tau_n}$ may be computed in linear time via*

$$\begin{aligned}
l_n &= \sum_{j=1}^J \log\det(\alpha_j\delta_j) + \frac{1}{2}\sum_{j=1}^J z_{nj}^\top z_{nj},\\
\{z_{nj}\}_{j=1}^J &= L_{mat}(\alpha\delta, \beta, \tilde\gamma)^{-1}\big(X_n^{\tau_n} - \Xi_n^{\tau_n}\theta\big).
\end{aligned}$$

Theorem 2. *The conditional mean $E[V_n|X_n^{\tau_n}]$ of $V_n \in \mathbb{R}^{q\times J}$ given $X_n^{\tau_n} \in \oplus_{j=1}^J\mathbb{R}^D$ is given by*

$$L_{mat}(\beta\gamma^\top\gamma, \beta, \gamma^\top\gamma)^\top\, L_{mat}(\alpha\delta, \beta, \tilde\gamma)^{-1,\top}\, L_{mat}(\alpha\delta, \beta, \tilde\gamma)^{-1}(X_n^{\tau_n} - \Xi_n^{\tau_n}\theta).$$

Thereafter, the conditional mean of Z_n can be computed recursively by

$$E[Z_{nj}|X_n^{\tau_n}] = E[Z_{n,j-1}|X_n^{\tau_n}] + E[V_{nj}|X_n^{\tau_n}].$$

The conditional variances are given by

$$\mathrm{Var}[V_{nj}|X_n^{\tau_n}] = \gamma_j^\top\zeta_j^{-1}\tilde\beta_j\xi_{j-1}\tilde\beta_j^\top\zeta_j^{-1,\top}\gamma_j + \gamma_j^\top\zeta_j^{-1}\zeta_j^{-1,\top}\gamma_j, \qquad \mathrm{Var}[Z_{nj}|X_n^{\tau_n}] = \xi_j.$$

Here $\zeta_j, \tilde\beta_j \in \mathbb{R}^{q\times q}$ arise from the backward computation given by $v_{J+1} = \tilde v_{J+1} = 0_{q\times q}$, and recursively for $j = J, \ldots, 1$,

$$\begin{aligned}
v_j &= v_{j+1} + \beta_j^\top\alpha_j^{-1,\top}\alpha_j^{-1}\beta_j,\\
\zeta_j &= \Big(cholesky\Big(I_q + \gamma_j(v_j - \tilde v_{j+1})\gamma_j^\top\Big)\Big)^\top,\\
\tilde\beta_j &= \zeta_j^{-1,\top}\gamma_j(v_j - \tilde v_{j+1}),\\
\tilde v_j &= \tilde v_{j+1} + \tilde\beta_j^\top\tilde\beta_j,
\end{aligned}$$

and the variances are found along the forward computation of $\xi_j \in \mathbb{R}^{q\times q}$ *given by* $\xi_0 = 0_{q\times q}$*, and recursively for* $j = 1, \ldots, J$*,*

$$\xi_j = \big(I_q - \gamma_j^\top \zeta_j^{-1} \tilde{\beta}_j \big) \xi_{j-1} \big(I_q - \gamma_j^\top \zeta_j^{-1} \tilde{\beta}_j \big)^\top + \gamma_j^\top \zeta_j^{-1} \zeta_j^{-1,\top} \gamma_j.$$

Proof. The formula for the conditional mean follows by inserting the representations of the variance and covariance terms in

$$\mathrm{E}[V_n | X_n^{\tau_n}] = \mathrm{Cov}(V_n, X_n^{\tau_n}) \mathrm{Var}(X_n^{\tau_n})^{-1} (X_n^{\tau_n} - \Xi_n^{\tau_n} \theta).$$

Concerning the conditional variances we choose $\breve{\gamma} = \{\breve{\gamma}_j\}_{j=1}^J$ such that $\breve{\gamma}_j \in \mathbb{R}^{q\times q}$ has full rank and approximates γ_j. Let

$$\breve{V}_j \sim \mathcal{N}_q(0_q, \breve{\gamma}_j^\top \breve{\gamma}_j), \quad \text{for } j = 1, \ldots, J,$$

be independent random variables, set $\breve{Z}_0 = 0_q \in \mathbb{R}^q$, and let the random variables $\breve{Z} = \{\breve{Z}_j\}_{j=1}^J \in \mathbb{R}^{q\times J}$ and $\breve{X} = \{\breve{X}_j\}_{j=1}^J \in \oplus_{j=1}^J \mathbb{R}^D$ be given by

$$\breve{Z}_j = \breve{Z}_{j-1} + \breve{V}_j, \qquad\qquad \breve{X}_j = \Xi_{nj}^{\tau_n} \theta + \beta_j^{\theta,\tau_n} \breve{Z}_j + U_{nj}.$$

Then $\mathrm{Var}[X_n^{\tau_n}]$ and $\mathrm{Var}[\breve{X}]$ both have full rank, and since $\mathrm{Var}[X_n^{\tau_n}] \approx \mathrm{Var}[\breve{X}]$ we have $\mathrm{Var}[V_n | X_n^{\tau_n}] \approx \mathrm{Var}[\breve{V}|\breve{X}]$. The conditional variance $\mathrm{Var}[\breve{V}|\breve{X}]$ is more amenable to matrix manipulations since the matrices $\breve{\gamma}_j \in \mathbb{R}^{q\times q}$ are invertible. Using the Woodbury inversion formula, inserting the terms

$$\mathrm{Var}(\breve{V}) = \mathrm{L}_{\mathrm{mat}}(\breve{\gamma}^\top \breve{\gamma}, 0, 0), \qquad\qquad \mathrm{Var}[\breve{X}|\breve{V}] = \mathrm{Var}(\bar{U}) = \mathrm{L}_{\mathrm{mat}}(\alpha\alpha^\top, 0, 0),$$

and using the multiplication formulae for $\mathrm{L}_{\mathrm{mat}}$-matrices, we find

$$\begin{aligned}
\mathrm{Var}[\breve{V}|\breve{X}] &= \mathrm{Var}(\breve{V}) - \mathrm{Cov}(\breve{V}, \breve{X}) \mathrm{Var}(\breve{X})^{-1} \mathrm{Cov}(\breve{X}, \breve{V}) \\
&= \Big(\mathrm{Var}(\breve{V})^{-1} + \mathrm{Var}(\breve{V})^{-1} \mathrm{Cov}(\breve{V}, \breve{X}) \mathrm{Var}[\breve{X}|\breve{V}]^{-1} \mathrm{Cov}(\breve{X}, \breve{V}) \mathrm{Var}(\breve{V})^{-1} \Big)^{-1} \\
&= \Big(\mathrm{L}_{\mathrm{mat}}(\breve{\gamma}^{-1} \breve{\gamma}^{-1,\top}, 0, 0) + \mathrm{L}_{\mathrm{mat}} \left(\alpha^{-1}\beta, \alpha^{-1}\beta, \{\mathrm{I}_q\}_{j=1}^J \right)^\top \mathrm{L}_{\mathrm{mat}} \left(\alpha^{-1}\beta, \alpha^{-1}\beta, \{\mathrm{I}_q\}_{j=1}^J \right) \Big)^{-1} \\
&\approx \Big(\mathrm{L}_{\mathrm{mat}} \left(\zeta \breve{\gamma}^{-1,\top}, \tilde{\beta}, \{\mathrm{I}_q\}_{j=1}^J \right)^\top \mathrm{L}_{\mathrm{mat}} \left(\zeta \breve{\gamma}^{-1,\top}, \tilde{\beta}, \{\mathrm{I}_q\}_{j=1}^J \right) \Big)^{-1} \\
&= \mathrm{L}_{\mathrm{mat}} \left(\zeta \breve{\gamma}^{-1,\top}, \tilde{\beta}, \{\mathrm{I}_q\}_{j=1}^J \right)^{-1} \mathrm{L}_{\mathrm{mat}} \left(\zeta \breve{\gamma}^{-1,\top}, \tilde{\beta}, \{\mathrm{I}_q\}_{j=1}^J \right)^{-1,\top},
\end{aligned}$$

where $\zeta, \tilde{\beta} \in \oplus_{j=1}^J \mathbb{R}^{q\times q}$ are given by the backward equations stated in the theorem with γ replaced by $\breve{\gamma}$. Here we have used Proposition 1 and the approximation $\breve{\gamma}_j \approx \gamma_j$. The forward equations now follow by Proposition 3 and 4, and by using $\breve{\gamma}_j \approx \gamma_j$ again. Finally, since the approximations can be chosen to be arbitrarily good the theorem follows. □

4.4 Estimation via the Expectation-Maximization algorithm

To estimate the parameters $(\alpha, \gamma, \beta^{\text{phase}}, \beta^{\text{amp}}, \theta)$ given the observations $X_1^{\tau_n}, \dots, X_N^{\tau_n}$ we maximize the likelihood using the EM-algorithm, where the associated $V_n = \{V_{nj}\}_{j=1}^J \in \mathbb{R}^{q \times J}$ are treated as missing.

Recall that the likelihood computations rely on the Taylor approximations $X_n(t_j) \approx X_{nj}^{\tau_n}$ for $j = 1, \dots, J$, where

$$X_{nj}^{\tau_n} = \Phi(t_j + \tau_{nj})\Xi_n\theta + \dot{\Phi}(t_j + \tau_{nj})\big(\mu_{nj} + \beta_j^{\text{phase}} Z_{nj} - \tau_{nj}\big)\Xi_n\theta_X + \beta_j^{\text{amp}} Z_{nj} + U_{nj}^X$$

where τ_n has been chosen as the posterior maximum of Eq. 5; we roughly expect that $\tau_{nj} \approx \mathrm{E}[\beta_j^{\text{phase}} Z_{nj} | X_n^{\tau_n}]$.

To derive formulae for an EM-step we use the formulae for the conditional means and variances given the observed variables $X_n^{\tau_n}$ stated in Theorem 2. These formulae can be computed simultaneously in linear time using the recursion formulae stated in Proposition 2 with $p = J$.

The update formulae stated in theorem 3 below have intra-dependencies within the parameters. To resolve this we propose to use the *conditional expectation-maximization algorithm* [19]. Following that theorem, we update the parameters in turn, where the present values are used for the parameters that have not yet been updated.

Theorem 3. *Let $\mathcal{P}$ be the projection on the space of positive semi-definite D-dimensional matrices that matches the structural assumption on $\alpha\alpha^\top$, that is, whether we are doing PCA, PLS, FA, CCA or no restriction. Furthermore, suppose that the conditional means and variances are computed at the present parameter values.*

The EM-update for θ is given by

$$\begin{aligned}
\hat{\theta} = \Bigg(\sum_{n=1}^N \sum_{j=1}^J \Big(& \Xi_n^\top (\Phi(t_j + \tau_{nj}) + (\hat{\beta}_j^{phase} E[Z_{nj} | X_n^{\tau_n}] - \tau_{nj}) \dot{\Phi}(t_j + \tau_{nj}))^\top \\
& (\hat{\alpha}\hat{\alpha}^\top)^{-1} (\Phi(t_j + \tau_{nj}) + (\hat{\beta}_j^{phase} E[Z_{nj} | X_n^{\tau_n}] - \tau_{nj})) \dot{\Phi}(t_j + \tau_{nj})) \Xi_n \\
& + \Xi_n^\top \dot{\Phi}(t_j + \tau_{nj})^\top (\hat{\alpha}\hat{\alpha}^\top)^{-1} \dot{\Phi}(t_j + \tau_{nj}) \Xi_n \cdot \hat{\beta}_j^{phase} \mathrm{Var}[Z_{nj} | X_n^{\tau_n}] \hat{\beta}_j^{phase, \top} \Big) \Bigg)^{-1} \\
\cdot \Bigg(\sum_{n=1}^N \sum_{j=1}^J \Big(& \Xi_n^\top (\Phi(t_j + \tau_{nj}) + (\hat{\beta}_j^{phase} E[Z_{nj} | X_n^{\tau_n}] - \tau_{nj}) \dot{\Phi}(t_j + \tau_{nj}))^\top \\
& (\hat{\alpha}\hat{\alpha}^\top)^{-1} \big(X_{nj} - \hat{\beta}_j^{amp} E[Z_{nj} | X_n^{\tau_n}] \big) \\
& - \Xi_n^\top \dot{\Phi}(t_j + \tau_{nj})^\top (\hat{\alpha}\hat{\alpha}^\top)^{-1} \hat{\beta}_j^{amp} \mathrm{Var}[Z_{nj} | X_n^{\tau_n}] \hat{\beta}_j^{phase, \top} \Big) \Bigg).
\end{aligned}$$

The EM-update for β_j^{amp} *for* $j = 1, \ldots, J$ *is given by*

$$\hat{\beta}_j^{amp} = \left(\sum_{n=1}^{N} \left(\left(X_{nj} - \Phi(t_j + \tau_{nj}) \Xi_n \hat{\theta} + \dot{\Phi}(t_j + \tau_{nj}) \Xi_n \hat{\theta} (\tau_{nj} - \beta_j^{phase} E[Z_{nj} | X_n^{\tau_n}]) \right) \cdot E[Z_{nj} | X_n^{\tau_n}]^\top \right.$$
$$\left. + \dot{\Phi}(t_j + \tau_{nj}) \Xi_n \hat{\theta} \hat{\beta}_j^{phase} \mathrm{Var}[Z_{nj} | X_n^{\tau_n}] \right) \right)$$
$$\cdot \left(\sum_{n=1}^{N} \left(\mathrm{Var}[Z_{nj} | X_n^{\tau_n}] + E[Z_{nj} | X_n^{\tau_n}] E[Z_{nj} | X_n^{\tau_n}]^\top \right) \right)^{-1}.$$

The EM-update for β_j^{phase} *for* $j = 1, \ldots, J$ *is given by*

$$\beta_j^{phase} = \left(\sum_{n=1}^{N} \left(\hat{\theta}^\top \Xi_n^\top \dot{\Phi}(t_j + \tau_{nj})^\top (\hat{\alpha} \hat{\alpha}^\top)^{-1} \right. \right.$$
$$\cdot \left(\left(X_{nj} - \Phi(t_j + \tau_{nj}) \Xi_n \hat{\theta} + \dot{\Phi}(t_j + \tau_{nj}) \Xi_n \hat{\theta} \tau_{nj} - \beta_j^{amp} E[Z_{nj} | X_{nj}^{\tau_n}] \right) E[Z_{nj} | X_{nj}^{\tau_n}]^\top \right)$$
$$\left. \left. - \hat{\beta}_j^{amp} \mathrm{Var}[Z_{nj} | X_n^{\tau_n}] \right) \right)$$
$$\left(\sum_{n=1}^{N} \left(E[Z_{nj} | X_n^{\tau_n}] E[Z_{nj} | X_n^{\tau_n}]^\top + \mathrm{Var}[Z_{nj} | X_n^{\tau_n}] \right) \cdot \right.$$
$$\left. \left(\hat{\theta}^\top \Xi_n^\top \dot{\Phi}(t_j + \tau_{nj})^\top (\hat{\alpha} \hat{\alpha}^\top)^{-1} \dot{\Phi}(t_j + \tau_{nj}) \Xi_n \hat{\theta} \right) \right)^{-1}$$

The EM-update for α *is given by*

$$\hat{\alpha} \hat{\alpha}^\top = \mathcal{P} \left(\frac{1}{NJ} \sum_{n=1}^{N} \sum_{j=1}^{J} \left(\hat{A}_{nj} \mathrm{Var}[Z_{nj}^X | X_n^{\tau_n}] \hat{A}_{nj}^\top + \hat{R}_{nj} \hat{R}_{nj}^\top \right) \right),$$

where

$$\hat{R}_{nj} = X_{nj} - \Phi(t_j + \tau_{nj}) \Xi_n^X \hat{\theta} + \dot{\Phi}(t_j + \tau_{nj}) \Xi_n \hat{\theta}_X \tau_{nj} - \hat{A}_{nj} E[Z_{nj} | X_n^{\tau_n}]$$
$$\hat{A}_{nj} = \dot{\Phi}(t_j + \tau_{nj}) \Xi_n \hat{\theta} \hat{\beta}_j^{phase} + \hat{\beta}_j^{amp}.$$

The EM-update for γ_j *for* $j = 1, \ldots, J$ *is given by*

$$\hat{\gamma}_j^\top \hat{\gamma}_j = \frac{1}{N} \sum_{n=1}^{N} \left(\mathrm{Var}[V_{nj} | X_n^{\tau_n}] + E[V_{nj} | X_n^{\tau_n}] E[V_{nj} | X_n^{\tau_n}]^\top \right).$$

Proof. Twice the joint negative log likelihood given the approximation and the latent parameters $(X_n^{\tau_n}, V_n)_{n=1}^N$ equals

$$\begin{aligned} NJ \log\det(\alpha\alpha^\top) + N\sum_{j=1}^J \log\det\left(\gamma_j^\top\gamma_j\right) \\ + \sum_{n=1}^N\sum_{j=1}^J \operatorname{trace}\left[(\alpha\alpha^\top)^{-1} R_{nj} R_{nj}^\top\right] \\ + \sum_{n=1}^N\sum_{j=1}^J \operatorname{trace}\left[(\gamma_j^\top\gamma_j)^{-1} V_{nj} V_{nj}^\top\right], \end{aligned}$$

where we to shorten the notation have introduced

$$\begin{aligned} R_{nj} &= X_{nj} - \Phi(t_j+\tau_{nj})\Xi_n\theta + \tau_{nj}\dot\Phi(t_j+\tau_{nj})\Xi_n\theta - A_{nj}Z_{nj} \\ A_{nj} &= \dot\Phi(t_j+\tau_{nj})\Xi_n\theta\beta_j^{\text{phase}} + \beta_j^{\text{amp}} \end{aligned}$$

Thus, the conditional expectation of the joint negative log likelihood given the observations $X_1^{\tau_n}, \dots, X_n^{\tau_n}$ is given by

$$\begin{aligned} &NJ \log\det(\alpha\alpha^\top) + N\sum_{j=1}^J \log\det\left(\gamma_j^\top\gamma_j\right) \\ &+ \sum_{n=1}^N\sum_{j=1}^J \operatorname{trace}\left[(\alpha\alpha^\top)^{-1}\left(A_{nj}\operatorname{Var}[Z_{nj}|X_n^{\tau_n}]A_{nj}^\top + \mathrm{E}[R_{nj}|X_n^{\tau_n}]\mathrm{E}[R_{nj}|X_n^{\tau_n}]^\top\right)\right] \\ &+ \sum_{n=1}^N\sum_{j=1}^J \operatorname{trace}\left[(\gamma_j^\top\gamma_j)^{-1}\left(\operatorname{Var}[V_{nj}|X_n^{\tau_n}] + \mathrm{E}[V_{nj}|X_n^{\tau_n}]\mathrm{E}[V_{nj}|X_n^{\tau_n}]^\top\right)\right]. \end{aligned} \tag{7}$$

One EM-step may be found minimizing the conditional joint negative log likelihood given the observations. □

Formulae for updating parameters for α, and by that finding $\mathcal{P}$, are described in Appendix A.

Remark 5. *For many applications it is necessary that the functional basis $\Phi(t) \in \mathbb{R}^{D\times K}$ is sufficiently rich for the modelling to be successful. This implies that the parameter dimension p can become large, and a computational bottleneck can be the formation of the matrices $\Phi(t_j+\tau_{nj})$ and Ξ_n and the computation of their matrix product. However, if these matrices have a tensor structure, then the computations can be streamlined. Suppose that there exist row vectors $\phi_{nj}, \dot\phi_{nj} \in \mathbb{R}^{1\times K_0}$ and $\xi \in \mathbb{R}^{1\times p_0}$ such that*

$$\Phi(t_j+\tau_{nj}) = I_D \otimes \phi_{nj}, \qquad \dot\Phi(t_j+\tau_{nj}) = I_D \otimes \dot\phi_{nj}, \qquad \Xi = I_D \otimes I_{K_0} \otimes \xi_n,$$

which e.g. will be the case if full interactions are used between the coordinates of $X(t_j)$, the coefficients of the functional basis, and the coefficients from the experimental design. Then $K = D \cdot K_0$ and $p = D \cdot K_0 \cdot p_0$, and e.g.

$$\sum_{j=1}^{J} \Xi_n^\top \Phi(t_j + \tau_{nj})^\top \left(\alpha\alpha^\top\right)^{-1} \Phi(t_j + \tau_{nj}) \Xi_n = \left(\alpha\alpha^\top\right)^{-1} \otimes \sum_{j=1}^{J} \phi_{nj}^\top \phi_{nj} \otimes \xi_n^\top \xi_n.$$

Using expressions like this the EM-update for $\hat{\theta}$ may be rewritten as

$$\hat{\theta} = \left(I_D \otimes \sum_{n=1}^{N} \sum_{j=1}^{J} \phi_{nj}^\top \phi_{nj} \otimes \xi_n^\top \xi_n \right)^{-1}$$
$$\cdot \left(\sum_{n=1}^{N} \sum_{j=1}^{J} \left(\left(X_{nj} - \hat{\beta}_j^{amp} E[Z_{nj} | X_n^{\tau_n}] \right) \otimes \phi_{nj}^\top \right.\right.$$
$$\left.\left. - \hat{\beta}_j^{amp} \mathrm{Var}[Z_{nj} | X_n^{\tau_n}] \otimes \dot{\phi}_{nj}^\top \right) \otimes \xi_n^\top \right).$$

5 Application to lameness signals

In this section we fit an MCA model to $N = 89$ three dimensional acceleration signals from trotting horses with induced lameness [24]. The signals are displayed in Figure 1. There are $J = 101$ observation points per curve.

The 89 signals were collected from 8 different horses, and hence a random effect of horse is to be expected. We will not model this, but we remark that comparing the conditional means of the latent variables, i.e. $\mathrm{E}[Z_{nj}]$ would serve as a good indicator of possible random effects.

We applied the proposed model with $q = 3$ latent components, and unstructured covariance for $\alpha\alpha^\top$. We used a Fourier basis with $p_0 = 21$ basis functions for each dimension as the functional basis. Lameness was used as covariate; with five treatment groups this gave $p = 5$.

5.1 Preliminary results

Although likelihood values decreased in every iteration of our algorithm, the predictions associated with phase variation did not converge in due time, so these results are preliminary.

Amplitude covariance Pointwise values of amplitude variances as functions of time are shown in the lower panel of Figure 2 . Although the longitudinal signal has the sharpest peaks in data, the estimated variances are clearly overestimated.

A novel feature of the MCA model is cross-correlations; results are shown in in the upper panel of Figure 2. There is definitely some cross-covariance, but no clear pattern; the cross-correlations change signs many times through the domain.

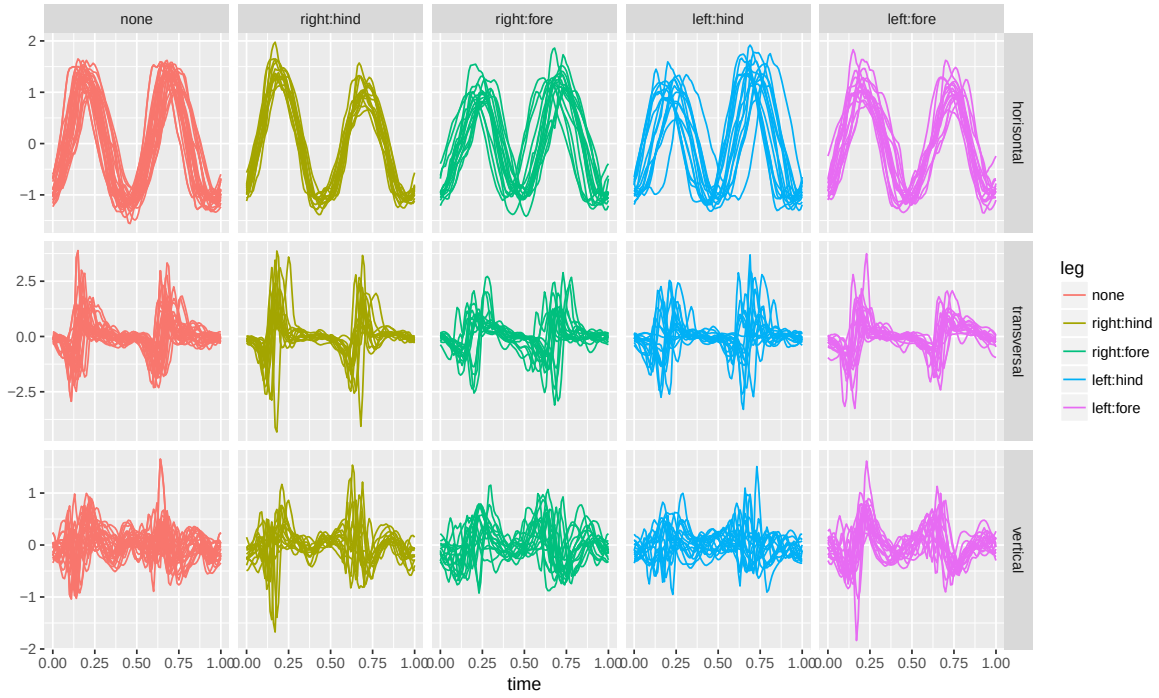


Figure 1: Acceleration signals from trotting horses with induced lameness.

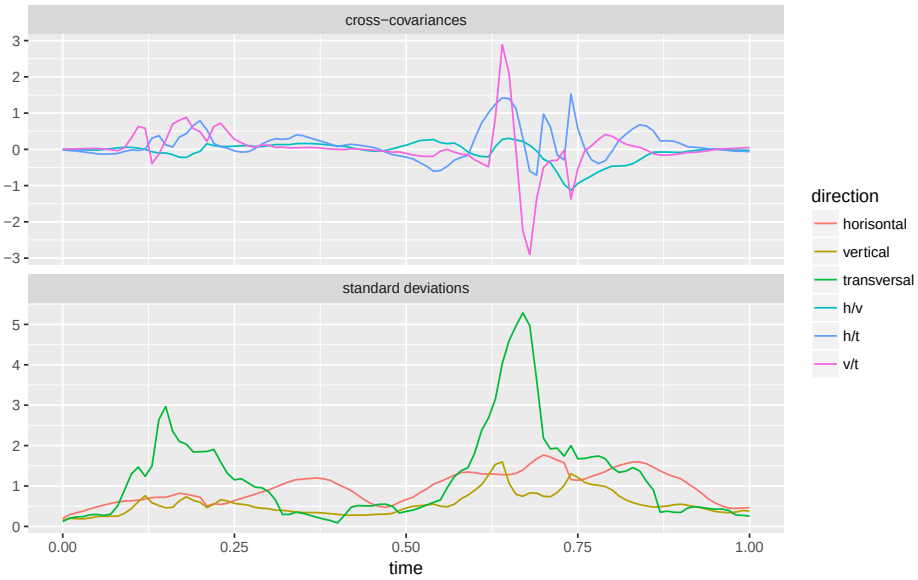


Figure 2: Pointwise amplitude variation

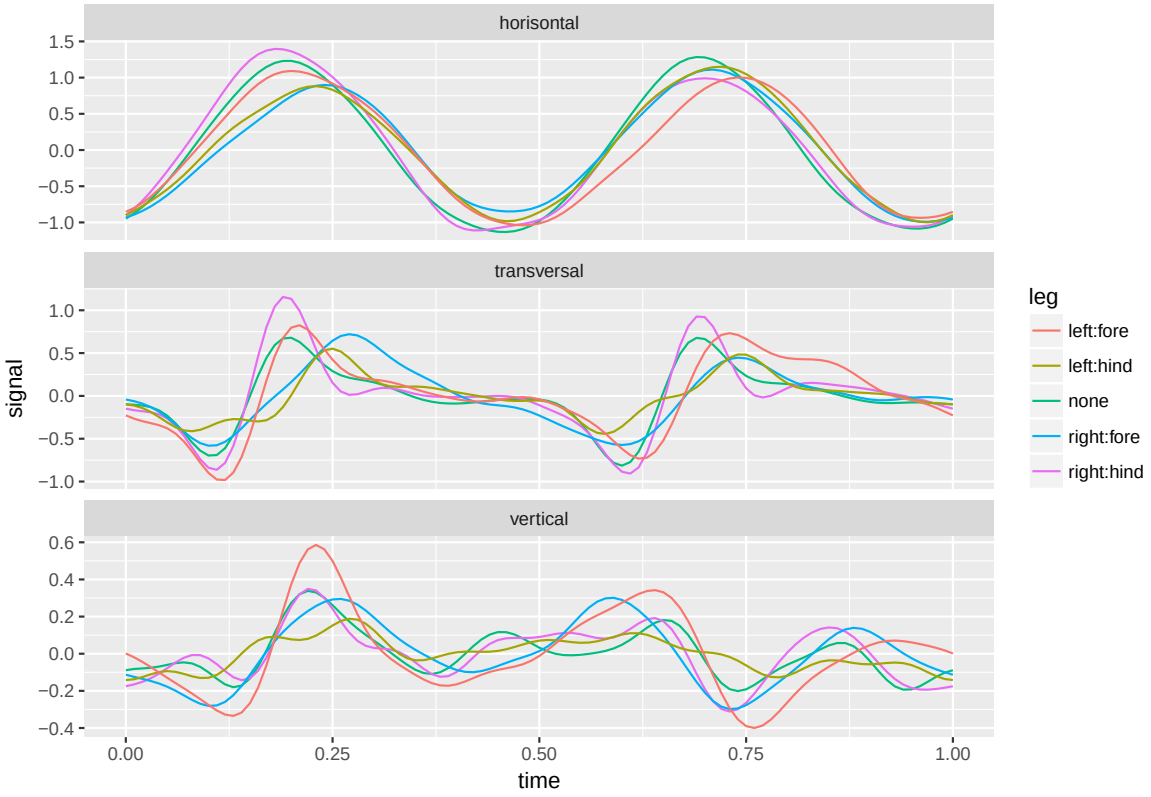


Figure 3: Estimated means for each lameness group

Mean structure Due to the lack of warping effects in our preliminary results, estimated population means are (and should be) close to the raw data means of the treatment groups; results are shown in Figure 3.

Phase variation Predicted warping trajectories were all very close to zero, the maximal deviation from the identity was 0.013. However, we still estimated phase variation at some locations of the domain, see Figure 4. It is located around a few peaks which roughly corresponds to the extremal values of the vertical signal; this in turn is associated with the take-off and flight phases of the trotting horses.

6 Discussion

The Markov component analysis has a great potential as one of the few models for functional data allowing both correlated multivariate responses and correlated phase and amplitude variation. However, the large number of parameters associated with the flexible MCA struc-

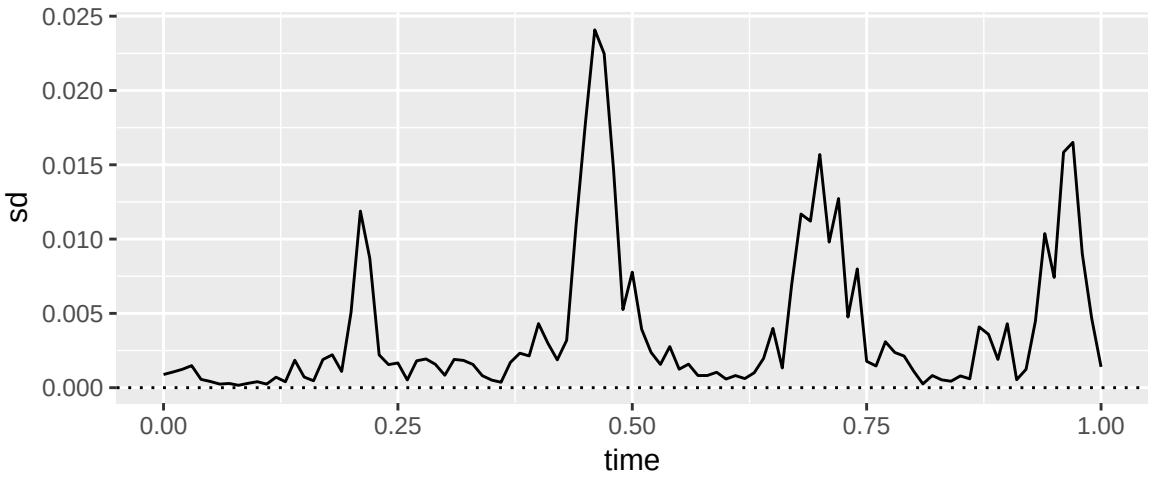


Figure 4: Estimated pointwise phase variation

ture is a challenge, and there is still work to be done in terms of the application of MCA. Although we believe our algorithm to be correct, fitting our model to data was harder than expected.

An interesting perspective is the application of a parametric structure on β similar to that of θ. This could potentially make the estimation more robust while still making use of the $\mathrm{L}_{\mathrm{mat}}$ framework of Section 3.1.

References

[1] K. Pearson, "On Lines and Planes of Closest Fit to Systems of Points in Space", *Philosophical Magazine*, vol. 2, no. 11, 559–572, 1901.

[2] H. Wold, "Estimation of principal components and related models by iterative least squares", In P.R. Krishnaiaah, *Multivariate Analysis*, Academic Press, 391–420, 1966.

[3] C. Spearman, "General Intelligence, Objectively Determined and Measured", *American Journal of Psychology*, vol. 15, 201–293, 1904.

[4] H. Hotelling, "Relations between two sets of variants", *Biometrika*, vol. 28, 321–377, 1936.

[5] J. Neveu, "Processes aléatoires gaussiens Séminaire de mathématiques supérieures", *Les presses de l'Université de Montréal*, 1968.

[6] J. Ramsay, "When the data are functions", *Psychometrika*, vol. 47, 379–396, 1982.

[7] S.E. Leurgans, R.A. Moyeed, B.W. Silverman, "Canonical correlation analysis when the data are curves", *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 55, 725–740, 1993.

[8] G. He, H.-G. Müller, J.-L. Wang, "Functional canonical analysis for square integrable stochastic processes", *Journal of Multivariate Analysis*, vol. 85, 54–77, 2003.

[9] J.-M. Chiou, H.-G. Müller, "Linear manifold modelling of multivariate functional data", *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 76, 605–626, 2014.

[10] T. Górecki, Ł. Smaga, "Multivariate analysis of variance for functional data", *Journal of Applied Statistics*, vol. 44, no. 12, 2172–2189, 2016.

[11] M.E. Tipping & C.M. Bishop, "Probabilistic Principal Component Analysis", *J. Royal Stat. Soc. C*, vol. 61, 611–622, 1999.

[12] F.R. Bach & M. Jordan, "A probabilistic interpretation of canonical correlation analysis", Technical Report 668, Department of Statistics, University of California, Berkeley, 2005.

[13] A. Berlinet & C. Thomas-Agnan, *"Reproducing Kernel Hilbert Spaces in Probability and Statistics"*, Springer Science + Business Media, New York, 2004.

[14] J. Durbin & S.J. Koopman, *Time Series Analysis by State Space Methods*, 2nd edition, Oxford University Press, New York, 2012.

[15] R.D. Gill & S.Johansen, "A survey of product-integration with a view toward application in survival analysis", *Ann. Statist.*, vol. 18, 1501–1555, 1990.

[16] B. Markussen, "Laplace approximation of transition densities posed as Brownian expectations", *Stoch. Proc. Appl.*, vol. 119, 208–231, 2009.

[17] B. Markussen, "Functional data analysis in an operator based mixed model framework", *Bernoulli*, vol. 19, 1–17, 2013.

[18] J. Goldsmith, J. Bobb, C.M. Crainiceanu, B. Caffo, D. Reich, "Penalized functional regression", *J. Comput. Graph. Stat.*, vol. 20, no. 4, 830–851, 2011.

[19] X.-L. Meng & D.B. Rubin, "Maximum likelihood estimation via the ECM algorithm: A general framework", *Biometrika*, vol. 80, no. 2, 267–278, 1993.

[20] N. L. Olsen, B. Markussen, L.L. Raket, "Simultaneous inference for misaligned multivariate functional data", *J. Royal Stat. Soc. C*, vol. 67, 1147–1176, 2018.

[21] J.C. Pinheiro & D. Bates, "Unconstrained parametrizations for variance-covariance matrices", *Statistics and Computing*, vol. 6, 289–296, 1996.

[22] P.Z. Hadjipantelis, J.A. Aston, H.-G. Müller, J. Moriarty, "Analysis of spike train data: A multivariate mixed effects model for phase and amplitude", *Electronic Journal of Statistics*, 8(2), 1797-1807, 2014.

[23] P.Z. Hadjipantelis, J.A. Aston, H.-G. Müller, J.P. Evans, "Unifying amplitude and phase analysis: A compositional data approach to functional multivariate mixede effects modeling of mandarin chinese", *Journal of the American Statistical Association*, 110(510), 545–559, 2015.

[24] M.H. Thomsen, A.T. Jensen, H. Sørensen, C. Lindegaard, P.H. Andersen, "Symmetry indices based on accelerometric data in trotting horses", *Journal of biomechanics*, 43(13), 2608-2612.

Appendix

A Updating coefficients for α

Here we consider different models for α and how to update parameters in an EM step; this in turn defines $\mathcal{P}$. Note that α only enters the expected likelihood value in formula (7) through:

$$NJ \log\det(\alpha\alpha^\top) + \sum_{n=1}^{N}\sum_{j=1}^{J} \operatorname{trace}\left[(\alpha\alpha^\top)^{-1}\left(A_{nj}\mathrm{Var}[Z_{nj}|X_n^{\tau_n}]A_{nj}^\top + \mathrm{E}[R_{nj}|X_n^{\tau_n}]\mathrm{E}[R_{nj}|X_n^{\tau_n}]^\top\right)\right] \quad (8)$$

If we define

$$S = \left(A_{nj}\mathrm{Var}[Z_{nj}|X_n^{\tau_n}]A_{nj}^\top + \mathrm{E}[R_{nj}|X_n^{\tau_n}]\mathrm{E}[R_{nj}|X_n^{\tau_n}]^\top\right) \quad (9)$$

we see that (S, NJ) is a sufficient statistic for $\alpha\alpha^\top$.

A.1 Unrestricted covariance

If we put no restriction on $\alpha\alpha^\top$, then (8) is minimised by $\alpha\alpha^\top = \frac{1}{NJ}S$.

A.2 pPCA

The (probabilistic) principal component analysis model assumes iid. covariance; $\alpha\alpha^\top = \sigma^2 I_D$. It is easily seen that (8) is minimised by $\sigma^2 = \frac{1}{NJD}\operatorname{trace}(S)$.

A.3 pCCA

We write $\beta_j^{\mathrm{amp}} = \begin{pmatrix}\beta_{j,1}^{\mathrm{amp}} \\ \beta_{j,2}^{\mathrm{amp}}\end{pmatrix}$, where $\beta_{j,1}^{\mathrm{amp}} \in \mathbb{R}^{d_1\times q}, \beta_{j,2}^{\mathrm{amp}} \in \mathbb{R}^{d_2\times q}$ with $d_1 + d_2 = D$. Then we have

$$X_{nj} = \begin{pmatrix}\beta_j^1 Z_{nj} + U_{nj}^1 \\ \beta_j^2 Z_{nj} + U_{nj}^2\end{pmatrix}, \quad U_{nj}^1 \sim N(0, \Psi_1), \qquad U_{nj}^2 \sim N(0, \Psi_2)$$

with U_{nj}^1 and U_{nj}^2 independent.

Thus $\alpha\alpha^\top = N(0, \begin{pmatrix}\Psi_1 & 0 \\ 0 & \Psi_2\end{pmatrix})$. Let $S = \begin{pmatrix}S_{11} & S_{12} \\ S_{21} & S_{22}\end{pmatrix}$, $S_{ij} \in R^{d_i\times d_j}$. Then (8) is minimised by:

$$\hat{\Psi}_k = \frac{1}{NJ}S_k, \quad k = 1, 2$$

A.4 FA

The assumption in the factor analysis model is: $\alpha\alpha^\top = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_D^2)$.
The minimiser of (8) is given by

$$\hat{\sigma}_k^2 = \frac{1}{NJ} \operatorname{diag}(S)_k, \quad k = 1, \ldots, D. \tag{10}$$

A.5 PLS

The assumption is the partial least squares model is $\alpha\alpha^\top = \begin{pmatrix} \sigma_1^2 I_{d_1} & 0 \\ 0 & \sigma_2^2 I_{d_2} \end{pmatrix}$, where $d_1 + d_2 = D$. Let $S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$, $S_{ij} \in R^{d_i \times d_j}$
Then the minimiser of (8) is given by

$$\hat{\sigma}_1^2 = \tfrac{1}{NJd_1} \operatorname{trace}(S_{11}), \quad \hat{\sigma}_2^2 = \tfrac{1}{NJd_2} \operatorname{trace}(S_{22}). \tag{11}$$

III
False discovery rates for functional data

Niels Lundtorp Olsen
Department of Mathematical Sciences
University of Copenhagen

Alessia Pini
Department of Statistical Sciences
Università Cattolica del Sacro Cuore

Simone Vantini
Department of Mathematics
Politecnico di Milano

Publication details
Preparing to submit.

False discovery rates for functional data

Niels Lundtorp Olsen, Alessia Pini, Simone Vantini

Abstract

Since Benjamini and Hochberg introduced false discovery rates (FDR) in their seminal paper (1995), this has become a very popular approach to the multiple comparisons problem. An increasingly popular topic within Functional Data Analysis is local inference, i.e., the continuous statistical testing of a null hypothesis along the domain. The principal issue in this topic is the infinite amount of tested hypotheses, which can be seen as an extreme case of the multiple comparisons problem.

In this paper we define and discuss the notion of false discovery rates in a very general functional data setting. Moreover, a continuous version of the Benjamini-Hochberg procedure is introduced along with a definition of adjusted p-value function. Some general conditions are stated, under which the functional Benjamini-Hochberg (fBH) procedure provides control of FDR. Two different simulation studies are presented; the first study has a one-dimensional domain and a comparison with the Fmax-method, and second study has a planar domain.

Finally, the proposed method is applied to satellite measurements of Earth temperature. In detail, we aim at identifying the regions of the planet where temperature has significantly increased in the last decades. After adjustment, large areas are still significant.

Keywords: functional data, local inference, multiple comparisons problem, Benjamini-Hochberg procedure, nonparametric inference

1 Hypothesis testing for functional data

Central to the field of statistics is statistical inference, in particular hypothesis testing. Functional data analysis (FDA) more often deals with "summary statistics" such as identifying mean curves/trajectories and principal modes of variation, including clustering and classification. Yet hypothesis testing is still a key part of FDA, where it is often done in conjunction with *functional regression*: What is the effect of a covariate on the response?, where covariates and response variables can be functions or scalars, depending on the setup.

In the common case of functional response and scalar covariates, functional regression is usually modelled through a linear model on the form $y_i = Ax_i + \text{noise}$, where y_i is a curve belonging to a suitable function space e.g. $L^2[0,1]$, x_i are covariate(s), and $A \in L$ is a linear operator which ought to be estimated. In the simple case of one covariate, the null hypothesis would be asking whether $A = 0$, and more generally the null hypothesis would be asking if $A \in U$ for a given subspace $U \subset L$.

This is hard question to ask for functional data and is not that frequently encountered. It is absent in some textbooks [13, 5] but has extensive treatment in [11]. There are various approaches to this issue, which importantly also depend on the scope of the test. We will distinguish between two kinds of tests, (i) *global tests*: does covariate x have 'influence' on curve θ in at least one part of the domain of θ, and (ii) *local tests* or *domain selections*: if covariate x has an influence, which part(s) of the domain of θ are significantly affected?

Global tests Global tests have been studied by various authors, for references see e.g. [11]. However, a crucial feature is that many of the available methods rely on (strong) parametric assumptions for data/estimators/covariances such as Gaussianity[1], which may be valid asymptotically but can be problematic for the usually small sample sizes and infinite dimensions that are characteristical for functional data analysis.

Non-parametric approaches such as permutation tests are popular alternatives to parametric tests. Curves are (randomly) permuted wrt. likelihood-independent transformations, and some suitable test statistic (e.g. deviation from mean) is evaluated for each permutation. However this is computationally expensive, and permutation tests are only asymptotically exact in the presence of several covariates.

Local tests Local tests have not been studied to the same extent as global tests. In the functional data analysis performing local inference carries several issues. The most important one is how to control the probability of committing a type 1 error globally over the whole domain. One recent framework for doing local testing on functional data is the *Interval-wise testing* by [12] extended to general linear models by [1]. These procedures perform non-parametric inference based on permutation schemes and provide (asymptotical) control of

[1] Note: This includes independence of PC scores (among other things)

the on each sub-interval of the domain. In detail, the probability of falsely selecting at least part of an interval where the null hypothesis is not violated is controlled. Another procedure is the Fmax-procedure [10, 17]. The Fmax-procedure is a method that provides strong control of the family-wise error rate, and like the IWT-framework it is based on permutation tests. The procedure is multivariate in nature, but it can also be applied to functional data.

An alternative approach to local testing in functional or spatial data is to use discrete features of the observed curves/fields such as local maxima or zero values. That is, instead of assessing a continuum of tests, one selects a finite number of data features for testing. Using discrete features has some challenge wrt. interpretation as one has to specify when two different observations can be considered instances of the same discrete feature, and there are no obvious definitions of domain selection. [6, 14] present some interesting methods with this, where they also proof control of FDR.

The remainder of the paper is organised as follows: Section 2 describes false discovery rates in the multivariate case and reviews related work. Section 3 presents the novel work of functional false discovery rates and functional Benjamini-Hochberg procedure. Two different simulation studies are presented in Section 4, and in Section 5 we apply the spherical IWT and false discovery rate adjustment procedure to a data set on climate change. Finally in Section 6 we highlight and discuss important points of this article. Proof of the main theorems are provided in the appendix.

2 False discovery rates for multivariate data

Background A central topic in statistics is multiple testing. Observed within virtually every area of statistics, it is fundamental in many statistical applications and multiple testing is recognised as an important statistical issue within many sciences.

Many ways to deal with multiple testing have been proposed with various advantages and disadvantages – one popular approach is to use the *False Discovery Rate* (FDR) [2]. The false discovery rate looks at the proportion of false rejections ("discoveries") among all among rejected hypotheses. The procedures controlling the FDR are generally more powerful than the ones controlling *family-wise error rate* (FWER), an alternative and more conservative approach to multiple testing, where the family-wise error is defined as one if any hypothesis is wrongly rejected (a false positive) and zero otherwise.

FDR is often applied in cases when a single or comparatively few false positives is not considered a serious issue, as long as their rate among all discoveries can be controlled. In [2] the Benjamini-Hochberg (BH) FDR-controlling procedure is introduced. In the succeeding literature, a number of other procedures for controlling FDR have been proposed (see [9] for a discussion). The paper [4] is important, as it introduces a modification of the BH procedure that controls FDR without specifying any dependency assumptions. More importantly, they

show that the original procedure introduced in [2] controls FDR under a weaker assumption than independence, namely *positive regression dependence* on the subset of true null hypotheses (PRDS).

Other closely related quantities for assessing the errors within the paradigm of multiple testing have been proposed such as the *weighted false discovery rate* (WDFR) [3], (see below) the *positive false discovery rate* [15], and the *local false discovery rate* [7].

However, in this paper we will only focus on the BH procedure and FDR which – due to its simple interpretation and definition – is still the most popular method for multiplicity correction. False Discovery Rates are only defined for finite numbers of hypotheses, and one must be careful when defining FDR on infinite sets of hypotheses.

False discovery rates Assume we are given a set of m *null hypotheses*, $G_1, \ldots, G_m$, each of which can either be true or false, and can either be accepted or rejected by a statistical test. Furthermore, let $w_1, \ldots, w_m$ be strictly positive weights with $\sum w_i = 1$, which we assume are a priori known. This will be used in the case of weighted false discovery rates. These weights can e.g. be interpreted as how important the different tests are, where the "usual" false discovery rate corresponds to the case of equal weights.

Definition 1 (False discovery rates, unweighted case)**.** The false discovery proportion is defined as:

$$Q = \frac{\#\{i : G_i \text{ is true but rejected}\}}{\#\{i : G_i \text{ is rejected}\}} \tag{1}$$

with $Q := 0$ whenever the deminator is zero. The false discovery rate is the expected value of this, $\mathrm{E}[Q]$.

Definition 2 (False discovery rates, weighted case)**.** The false discovery proportion in the weighted case is defined as:

$$Q = \frac{\sum_{i:G_i \text{ is true but rejected}} w_i}{\sum_{i:G_i \text{ is rejected}} w_i} \tag{2}$$

with $Q := 0$ whenever the denominator is zero. The weighted false discovery rate is the expected value of this, $\mathrm{E}[Q]$.

The Benjamini-Hochberg procedure Let $\{p_{(i)}\}_{i=1}^m$ be the p-values sorted in increasing order. Let $\{G_{(i)}\}_{i=1}^m$ be the corresponding ordering of the hypotheses and $\{w_{(i)}\}_{i=1}^m$ the corresponding ordering of weights. The very popular and easily applicable *Benjamini-Hochberg* (BH) *procedure* for multiple comparison adjustment [2, 3] is defined as follows:

Definition 3 (Benjamini-Hochberg procedure, unweighted case)**.** Define

$$k = \arg\max_i \left[p_{(i)} \leq \frac{i}{m}\alpha \right] \tag{3}$$

The Benjamini-Hochberg procedure is: *reject hypotheses* $G_{(1)}, \ldots, G_{(k)}$ *corresponding to the* k *smallest* p*-values and accept the rest.*

Definition 4 (Benjamini-Hochberg procedure, weighted case)**.** Define

$$k = \arg\max_i \left[p_{(i)} \leq \sum_{j:p_j \leq p(i)} w_j \alpha \right] \tag{4}$$

The weighted Benjamini-Hochberg procedure is: *reject hypotheses* $G_{(1)}, \ldots, G_{(k)}$ *corresponding to the* k *smallest* p*-values and accept the rest.*

Unlike most adjustment procedures introduced prior to this, the BH procedure is scalable: if data is duplicated such that one has twice the amount of hypotheses, the inference by using the BH procedure will still be the same.

Adjusted p-values A concept often used in context of multiple testing is *adjusted* (or corrected) *p-values*. Informally, the adjusted p-values for a multiple testing procedure are defined as corrections $\{\tilde{p}_i\}$ of the original p-values such that a null hypothesis G_i can be rejected at level α if $\tilde{p}_i \leq \alpha$.

The adjusted p-values for the unweighted Benjamini-Hochberg procedure are

$$\tilde{p}_{(i)} = \min(1, \tfrac{m}{i} p_{(i)}, \ldots, \tfrac{m}{m-1} p_{(m-1)}, p_{(m)}) \tag{5}$$

where $p_{(\cdot)}$ and $\tilde{p}_{(\cdot)}$ are the order statistics of p and $\tilde{p}$, respectively. By construction, the ordering is the same.

We can likewise define adjusted p-values for the weighted Benjamini-Hochberg procedure:

$$\tilde{p}_{(i)} = \min \left\{ 1, \frac{p_{(i)}}{\sum_{j:p_j \leq p_{(i)}} w_j}, \frac{p_{(i+1)}}{\sum_{j:p_j \leq p_{(i+1)}} w_j}, \ldots, \frac{p_{(m-1)}}{\sum_{j:p_j \leq p_{(m-1)}} w_j}, p_{(m)} \right\}, \quad i = 1, \ldots m \tag{6}$$

PRDS and control of the false discovery rate Benjamini and Hochberg showed in their seminal paper [2] that if the test statistics for different hypotheses are independent, then the false discovery rate is controlled by $\frac{m_0}{m}\alpha$ where m_0 is the total number of correct null hypotheses. The independence assumption was later relaxed by [4] to *Positive regression dependency on one* (PRDS).

Below we define the PRDS property and extend it to the infinite-dimensional case, which will be needed later.

Definition 5 (Positive regression dependency on one (PRDS))**.** Let '$\leq$' be the partial/usual ordering on $\mathbb{R}^l$. An *increasing set* $D \subseteq \mathbb{R}^l$ is a set satisfying $x \in D \wedge y \geq x \Rightarrow y \in D$.

A random variable $\mathbf{X}$ on $\mathbb{R}^l$ is said to be *PRDS on* I_0, where I_0 is a subset of $\{1, \dots, l\}$, if it for any increasing set D and $i \in I_0$ holds that

$$x \leq y \Rightarrow P(\mathbf{X} \in D | X_i = x) \leq P(\mathbf{X} \in D | X_i = y) \tag{7}$$

Let $\mathbf{Z}$ be an infinite-dimensional random variable, where instances of $\mathbf{Z}$ are functions $T \to \mathbb{R}$. We say that $\mathbf{Z}$ is PRDS on $U \subseteq T$ if all finite-dimensional distributions of $\mathbf{Z}$ are PRDS. That is, for all finite subsets $I = \{i_1, \dots, i_l\} \subseteq T$, it holds that $Z(i_1), \dots, Z(i_l)$ is PRDS on $I \cap U$.

We refer to [4] for a discussion on the PRDS property and how it relates to other types of dependency.

Theorem 6. Given a set of hypotheses $\{H_1, \dots, H_m\}$ and corresponding p-values $(p_1, \dots, p_m)$, let $I_0 = \{i_1, \dots, i_{m_0}\} \subseteq \{1, \dots, m\}$ be the index set corresponding to true null hypotheses $\{H_{i_1}, \dots, H_{i_{m_0}}\}$.

If the joint distribution of the p-values $(p_1, \dots, p_m)$ is PRDS on I_0, the BH procedure controls the FDR at level $\frac{m_0}{m}\alpha$ in the sense that

$$\mathrm{E}[Q] \leq \frac{m_0}{m}\alpha$$

where Q is the false discovery rate.

Proof. See [4, Theorem 1.2] □

3 False discovery rates for Functional Data

In this section we define the false discovery rate for functional data and propose a functional extension of the Benjamini-Hochberg procedure. These are the functional versions of the 'usual' false discovery rates and the BH procedure, analogous to the discrete cases.

Our definition of false discovery rate is very related to that of [16], although [16] uses a seemingly arbitrary lower bound for the measure of rejection region, not present in other papers on FDR, and avoids usage of p-values which are natural (albeit sometimes controversial) in context of multiple testing.

We prove control of the false discovery rate under regularity conditions and outline a simple algorithm for calculating the functional BH procedure; for simplicity and ease of presentation we only consider $T = (0, 1)^D$ in that theorem, and present it in two versions: an unweighted version and a weighted version, where the former is a special case of the latter. The unweighted case requires considerably less notation in formulating the theorem and proof. In section 3.3 we introduce the adjusted p-value function, an important tool for practical applications, and section 3.4 contains theoretical results about control of FDR.

General setting For the remainder of this section we assume that we have N functional samples, $y_1, \ldots, y_N : \mathbf{T} \to \mathbb{R}$, where $\mathbf{T} \subset R^d$ is an open and bounded subset of $\mathbb{R}^d$.

Suppose that we for each point $t \in \mathbf{T}$ have a null hypothesis H_t^0 together with an alternative hypothesis H_t^A, that we are interesting in testing. Furthermore, suppose that by pointwise application of some statistical test we obtain p-values for every t with the property that (H_t^0 true) $\implies p_t \sim U(0,1)$, where $U(0,1)$ is the uniform distribution on $(0,1)$. The p-values together make up a function $p : \mathbf{T} \to [0,1]$, the *p-value function*.

Let U be the set of the domain on which the null hypothesis is true, ie. $U = \{t \in \mathbf{T} : H_0^t$ is true$\}$. Let ν be a bounded measure on $\mathbf{T}$ that is absolutely continuous wrt. the Lebesgue measure, which we denote by μ. By absolute continuity we have $\nu = f \cdot \mu$ for some measurable function $f : \mathbf{T} \to [0, \infty)$. The function f can be interpreted as a *weight function* assigning more weight to some regions of $\mathbf{T}$ than others, and is equivalent to the weights used in Section 2.

3.1 Definition of functional false discovery rates

Given U and an instance of p, let $V = \{t : H_t^0$ is true and H_t^0 is rejected$\}$ be the region where the null hypothesis is wrongly rejected, and let $S = \{t : H_t^0$ is false and H_t^0 is rejected$\}$ be the region where the null hypothesis is correctly rejected. The set V corresponds to committing type I errors, and in a given research situation, it is desired that V is as small as possible and S is as large as possible.

Definition 7 (Functional false discovery rate)**.** Define the *functional false discovery rate* (FDR) as

$$\mathrm{E}[Q] = \mathrm{E}\left[\frac{\nu(V)}{\nu(V \cup S)} 1_{\nu(V \cup S) > 0}\right] \tag{8}$$

where Q is the *proportion of false discoveries*.

Remark 8 (False discovery rates for other manifolds)**.** In this paper we define false discovery rates for functional data defined on open subsets of $\mathbb{R}^d$. However, many smooth manifolds can be diffeomorphically mapped into open and bounded subsets of $\mathbb{R}^d$. The mapping gives naturally rise to a measure on this set, which can be used as measure for functional false discovery rates.

3.2 The functional Benjamini-Hochberg procedure: the adjusted threshold

Analogous to the multivariate case, we can define the Benjamini-Hochberg procedure for functional data:

Definition 9 (Functional Benjamini-Hochberg procedure)**.** Let $\alpha \in (0,1)$ be a desired significance level for the tests.The functional Benjamini-Hochberg (fBH) procedure is:

Reject hypotheses H_t^0 that satisfy

$$p(t) \leq \alpha^* \quad \text{where} \quad \alpha^* = \arg\max_r \frac{\nu(\{s : p(s) \leq r\})}{\nu(\mathbf{T})} \geq \alpha^{-1} r$$

We will refer to α^* as the *adjusted threshold* for the procedure, and the function $r \mapsto \nu(\{s : p(s) \leq r\})$ as the *cumulated p-value function*.

Two examples of the functional BH procedure are shown in Figure 1.

The main theoretical result of this article is that the fBH procedure can be approximated by the multivariate BH procedure, and that it controls the expected value of FDR by $\alpha\mu(U)$ under regularity conditions.

3.3 The functional Benjamini-Hochberg procedure: the adjusted p-value function

As an alternative to adjusting the threshold, we may adjust the p-value function itself.

This is analogous to adjusted p-values, which applies in the discrete case, and plays a similar role, ie. if $\tilde{p}(t) \leq \alpha$ this means that H_t^0 will be rejected when the (weighted) BH procedure is applied with threshold/significance level α.

The *adjusted p-value function* is defined as:

$$\tilde{p}(t) = \min_{s \geq p(t)} \left\{ 1, \frac{\nu(\mathbf{T})s}{\nu(r : p(r) \leq s)} \right\}, \quad t \in \mathbf{T} \tag{9}$$

Adjusted p-values allow us to quantify significance after adjustment and to simultaneously compute rejection areas for all values of α.

Proposition 10 (Control of FDR using adjusted p-value function)**.** Under the assumptions of Proposition 12 or Proposition 14, the adjusted p-value function controls the false discovery rate at level α in the sense that if we reject hypotheses on the set $\{t : \tilde{p}(t) \leq \alpha\}$, then the false discovery rate,

$$Q = \frac{\nu(\{\tilde{p}(t) \leq \alpha\} \cap V)}{\nu\{\tilde{p}(t) \leq \alpha\}} 1_{\nu\{\tilde{p}(t) \leq \alpha\} > 0}$$

satisfies $\mathrm{E}[Q] \leq \alpha\nu(U)$.

Proof. Let $\alpha \in (0, 1)$ be fixed, and let α^* be the associated cutoff level for the functional BH procedure following Definition 9. If we can show that $\tilde{p}(t) \leq q \Leftrightarrow p(t) \leq \alpha^*$, the control of FDR follows from Proposition 12 orProposition 14, respectively. This is true since

$$p(t) \leq \alpha^* \Leftrightarrow \exists s \geq p(t) : a(s) \geq rs \Leftrightarrow \exists s \geq p(t) : \frac{s}{a(s)} \leq \alpha$$

$$\tilde{p}(t) \leq \alpha \Leftrightarrow \exists s \geq p(t) : \frac{s}{a(s)} \leq \alpha$$

□

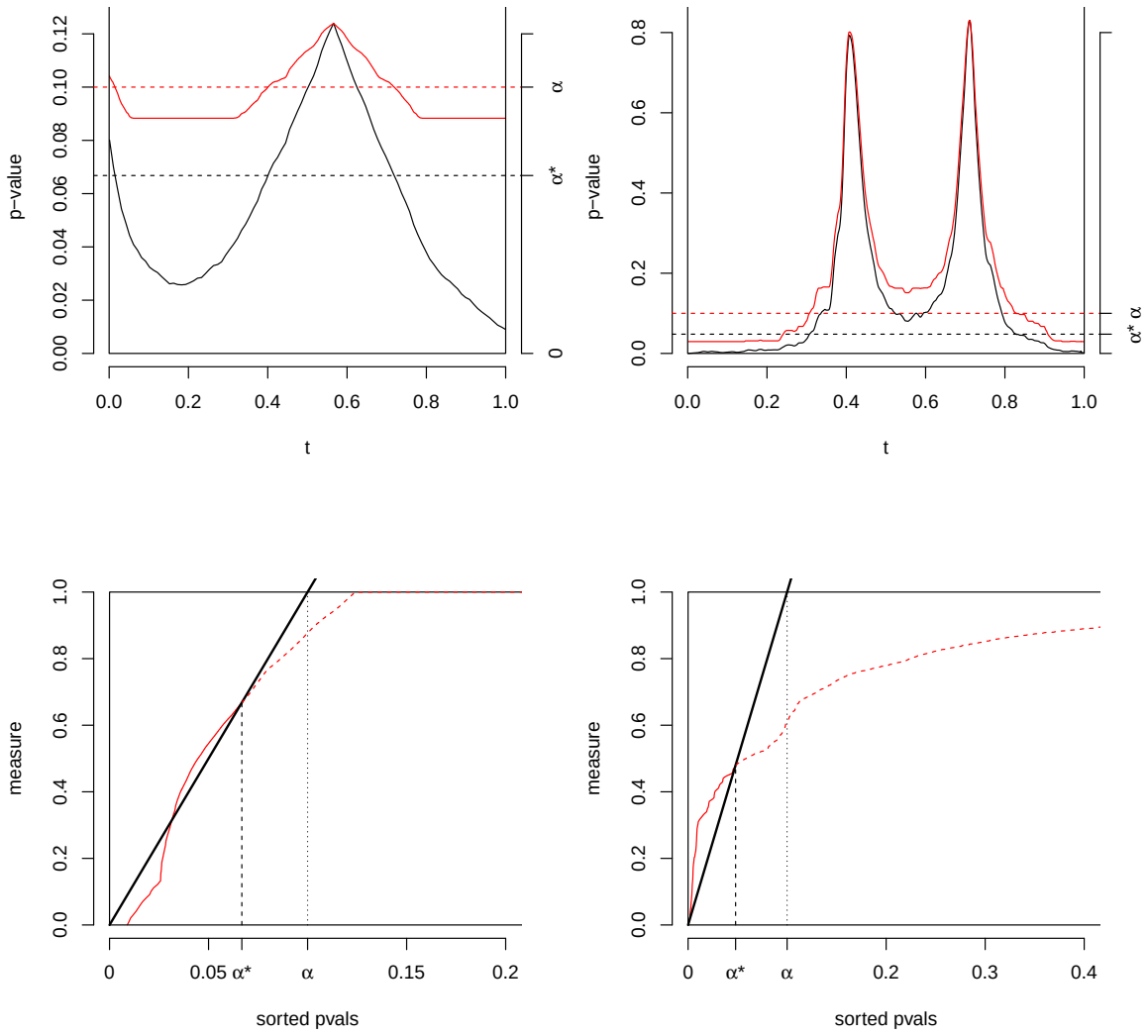


Figure 1: Two illustrations of the functional Benjamini-Hochberg procedure and adjusted p-value functions with $\alpha = 0.10$. Upper plots: black curves are original p-values; red curves are adjusted p-values. Lower plots: The red lines denote the cumulated p-value functions, and the thick lines have slope α^{-1}. Null hypotheses corresponding to the solid red lines are rejected, while those above are accepted

3.4 Control of false discovery rates for functional data

In this section, let $\mathbf{T} = (0,1)^d$. Below we prove that the functional Benjamini-Hochberg procedure controls FDR under PRDS and regularity assumptions. Let $\nu = f \cdot \mu$, and assume that f is a bounded and strictly positive density function $\mathbf{T} \to \mathbb{R}$. The case $f \equiv 1$ corresponds to equal weighting, ie. $\nu = \nu$.

Theorem 11. Let $\{S_k\}_{k=1}^{\infty}, S_1 \subset S_2 \subset \dots$ be a dense, uniform grid in $\mathbf{T}$ in sense that S_k weighted by f uniformly approximates all level sets of p and $p|_U$ with probability one:

$$P\left[\lim_{k\to\infty} \sup_r \frac{\sum_{i\in S_k \cap \{s:p(s)\le r\}} f(i)}{\#S_k} - \int_{\{s:p(s)\le t\}} f(x)\,\mathrm{d}x \to 0\right] = 1 \tag{10}$$

and

$$P\left[\lim_{k\to\infty} \sup_r \frac{\sum_{i\in S_k \cap \{s:p(s)\le r\}\cap U} f(i)}{\#S_k} - \int_{\{s:p(s)\le t\}\cap U} f(x)\,\mathrm{d}x \to 0\right] = 1 \tag{11}$$

Furthermore assume that p is PRDS wrt. the set of true null hypotheses with probability one, and that the assumptions about p-value function below hold true with probability one:

(a1) All level sets of p have zero measure,

$$\nu\{s : p(s) = t\} = 0 \quad \forall t \in \mathbf{T}$$

(a2) $\alpha^* \in (0, \alpha] \Rightarrow$: for any open neighbourhood O around α^* there exists $s_1, s_2 \in O$ s.t. $a(s_1) > \alpha^{-1} s_1, a(s_2) < \alpha^{-1} s_2$, where a is the cumulated p-value function (Definition 9).

(a3) $[\alpha^* = 0] \Rightarrow \min p(t) > 0$.

Then the functional BH procedure controls FDR at level $\alpha\nu(U)$, ie. $\mathrm{E}[Q] \le \alpha\nu(U)$, when applying the functional BH procedure at level α.

Note that the assumptions (16) and (17) are much simplified in the equal-weight case where $f \equiv 1$.

Proof. Following Proposition 14, which we prove below, and the notation of that proposition, it shows that $Q_k \to Q$ almost surely, and that $\limsup_{k\to\infty} \mathrm{E}[Q_k] \le \alpha\nu(U)$. As $0 \le Q_k \le 1$ for all k, it is now a simple application of the dominated convergence theorem to show that $\mathrm{E}[Q] \le \alpha\mu(U)$:

$$\mathrm{E}[Q] = \mathrm{E}[\lim_{k\to\infty} Q_k] = \lim_{k\to\infty} \mathrm{E}[Q_k] = \limsup_{k\to\infty} \mathrm{E}[Q_k] \le \alpha\nu(U)$$

□

As remarked in [4] the PRDS assumption is sufficient but not necessary, and (7) needs only to be true for certain sets defined in relation to the order statistics of p. The details are quite technical, and we refer to [4, Remark 4.2] and the general discussion of that paper.

Sufficient criteria for the one-dimensional case The assumptions (a1-a3) and equations (10) and (11) are consequences of the more simple criteria in the one-dimensional case, which must be true with probability one:

(d1) p is continuous.

(d2) There exists a maximal number of crossings, N_C, i.e.:

$$\#\{s \in [0, D] | p(s) = t\} \leq N_C \quad \forall t \in (0, 1]$$

(d3) U is a finite union of disjoint intervals.

These criteria will generally be true for smooth curves as we typically use to model functional data.

3.5 The functional Benjamini-Hochberg procedure: Finite approximation and computational cost

As we have a continuous amount of hypotheses and p-values, the functional BH procedure is in principle unattainable computationally. However, Propositions 12 and 14 show that the fBH procedure and functional false discovery rates are limits of the discrete BH procedure and FDR, which also outlines an algorithm for approximating the functional BH procedure. Proposition 12 and Corollary 13 can be seen as special cases of Proposition 14 and Corollary 15, respectively.

Unweighted case Here we assume equal weights, that si $\nu = \mu$ (the Lebesgue measure).

Proposition 12. Let $\{S_k\}_{k=1}^{\infty}, S_1 \subset S_2 \subset \dots$ be a dense, uniform grid in $\mathbf{T}$ in sense that S_k uniformly approximates all level sets of p and $p|_U$ with probability one:

$$P\left[\lim_{k\to\infty} \sup_r \frac{\#(S_k \cap \{s : p(s) \leq r\})}{\#S_k} - \mu\{s : p(s) \leq r\} \to 0\right] = 1 \tag{12}$$

and

$$P\left[\lim_{k\to\infty} \sup_r \frac{\#(S_k \cap \{s : p(s) \leq r\} \cap U)}{\#S_k} - \mu(\{s : p(s) \leq r\} \cap U) \to 0\right] = 1 \tag{13}$$

Furthermore assume that p is PRDS wrt. the set of true null hypotheses with probability one, and that the assumptions about p-value function below hold true with probability one:

(a1) All level sets of p have zero measure,

$$\mu\{s : p(s) = t\} = 0 \quad \forall t \in \mathbf{T}$$

(a2) $\alpha^* \in (0, \alpha] \Rightarrow$: for any open neighbourhood O around α^* there exists $s_1, s_2 \in O$ s.t. $a(s_1) > \alpha^{-1} s_1, a(s_2) < \alpha^{-1} s_2$, where a is the cumulated p-value function (Definition 9).

(a3) $[\alpha^* = 0] \Rightarrow \min p(t) > 0$.

Now define the k'th step false discovery proportion Q_k by applying the (usual) BH procedure at level α to p evaluated in S_k.

Mathematically, this can be defined by

$$Q_k = \frac{\#\{t \in S_k : p(t) \leq b_k\} \cap U}{\#\{t \in S_k : p(t) \leq b_k\}}, \quad b_k = \arg\max_r \frac{\#\{s \in S_k : p(s) \leq r\}}{\#S_k} \geq \alpha^{-1} r \tag{14}$$

Then Q_k behaves asymptotically as the functional false discovery proportion and asymptotically the false discovery rate $\mathrm{E}[Q_k]$ is controlled by $\alpha\mu(U)$:

$$\lim_{k\to\infty} Q_k = Q, \quad \limsup_{k\to\infty} \mathrm{E}[Q_k] \leq \alpha\mu(U) \tag{15}$$

where Q is defined as in (8).

Proof. Proposition 21 □

From the proof of the theorem we have the following important corollary which states that as the grid S_k becomes tighter and tighter, hypotheses are eventually rejected or accepted:

Corollary 13. For $t \in \cup_{m=1}^{\infty} S_m$ and $k \geq 1$, define $H_{t,k} = (t \in S_k) \wedge (p(t) \leq b_k)$ where b_k is is given by (14). That is, $H_{t,k}$ is true if the adjusted threshold at step k is larger than $p(t)$.

Assume $p(t) \neq \alpha$. Eventually, $H_{t,k}$ is either rejected or accepted.

Proof. Proposition 18 □

Weighted case The weighted case allows for more general measures than just the Lebesgue measure, but the notation is more tedious.

Let $\nu = f \cdot \mu$, and assume that f is a bounded and strictly positive density function $\mathbf{T} \to \mathbb{R}$.

Under almost similar assumptions to the unweighted case, we are able to control the false discovery rate at level α:

Proposition 14. Let $\{S_k\}_{k=1}^{\infty}, S_1 \subset S_2 \subset \ldots$ be a dense, uniform grid in $\mathbf{T}$ in sense that S_k weighted by f uniformly approximates all level sets of p and $p|_U$ with probability one:

$$P\left[\lim_{k\to\infty} \sup_r \frac{\sum_{i \in S_k \cap \{s : p(s) \leq r\}} f(i)}{\#S_k} - \int_{\{s : p(s) \leq t\}} f(x)\,\mathrm{d}x \to 0\right] = 1 \tag{16}$$

and

$$P\left[\lim_{k\to\infty}\sup_r \frac{\sum_{i\in S_k\cap\{s:p(s)\leq r\}\cap U} f(i)}{\#S_k} - \int_{\{s:p(s)\leq t\}\cap U} f(x)\,\mathrm{d}x \to 0\right] = 1 \tag{17}$$

Furthermore assume that p is PRDS wrt. the set of true null hypotheses with probability one, and that the assumptions about p-value function below hold true with probability one:

(a1) All level sets of p have zero measure,

$$\nu\{s : p(s) = t\} = 0 \quad \forall t \in \mathbf{T}$$

(a2) $\alpha^* \in (0, \alpha] \Rightarrow$: for any open neighbourhood O around α^* there exists $s_1, s_2 \in O$ s.t. $a(s_1) > \alpha^{-1}s_1, a(s_2) < \alpha^{-1}s_2$, where a is the cumulated p-value function (Definition 9).

(a3) $[\alpha^* = 0] \Rightarrow \min p(t) > 0$.

Now define the k'th step false discovery proportion Q_k by applying the (usual) BH procedure at level α to p evaluated in S_k, weighted by f evaluated in S_k.

Mathematically, this can be defined by

$$Q_k = \frac{\sum_{t\in S_k\cap U:p(t)\leq b_k} f(t)}{\sum_{t\in S_k:p(t)\leq b_k} f(t)}, \quad b_k = \arg\max_r \frac{\sum_{\{s\in S_k:p(s)\leq r\}} f(s)}{\sum_{\{s\in S_k\}} f(s)} \geq \alpha^{-1}r \tag{18}$$

Then Q_k behaves asymptotically as the functional false discovery proportion and asymptotically the false discovery rate $\mathrm{E}[Q_k]$ is controlled by $\alpha\nu(U)$:

$$\lim_{k\to\infty} Q_k = Q, \quad \limsup_{k\to\infty} \mathrm{E}[Q_k] \leq \alpha\nu(U) \tag{19}$$

where Q is defined as in (8).

Proof. See appendix □

Analogous to Corollary 13 we have the following corollary which states that as the grid becomes tighter and tighter, hypotheses are eventually rejected or accepted:

Corollary 15. For $t \in \cup_{m=1}^{\infty} S_m$ and $k \geq 1$, define $H_{t,k} = (t \in S_k) \wedge (p(t) \leq b_k)$ where b_k is given by (18). That is, H_t is true if the adjusted threshold at step k is larger than $p(t)$.

Assume $p(t) \neq \alpha$. Eventually, $H_{t,k}$ is either rejected or accepted.

Algorithm Propositions 12 and 14 outline an algorithm for approximating the functional BH procedure:

Apply the BH procedure to $\{p(t) : t \in S_k\}$ *with weights* $(f(t) : t \in S_k)$ *where the weights have been normalised to one.*

By corollaries 13 and 15, we also have that hypotheses will eventually be rejected or not, a very important property. If there are no weights, the normalising constant is $(\#S_k)^{-1}$.

Computational cost The fBH procedure adds $O(n \log n)$ computational cost to calculation of adjusted p-values, where $n = \#S_k$ is the number of approximation points,. This extra computational cost is due to the sorting of p-values, which has $O(n \log n)$ computational cost. As a default, calculation of pointwise p-values has complexity $O(n)$, thus the total computational cost is $O(n \log n)$.

However, sorting algorithms on modern computer are very fast, whereas calculating pointwise p-values can be comparatively slow, in particular if permutation tests are used, such as in Section 4.1. Thus the computational cost from fBH adjustment procedure is expected to be negligible in practice. In the simulation studies and the case study presented below, the calculation times for p-value adjustments were negligible.

4 Simulations

In this section two different simulation studies are performed, which differ both in scope and setting – in the first simulation study a functional-on-scalar regression is performed using permutation tests, while in the second simulation we test for mean equal to zero using one-sided t-tests.

The first simulation study has a more theoretical flavour and compares our proposed method to the Fmax method. The second study is intended to simulate a more realistic scenario with a number of disjoint peaks, and the performance of our method for various levels of α is analysed.

4.1 1D-simulation with comparison to Fmax-method

Description of simulation In this section we wanted to numerically assess the performance of our fBH procedure, and to compare it with the Fmax-procedure [10, 17]. The Fmax-procedure is a method that provides strong control of the family-wise error rate in a multivariate high-dimensional setting. It can be extended to functional data by applying it to the discrete point-wise evaluations of the functional data.

We simulated functional data according to the following functional-on-scalar linear model:

$$y_i(t) = \beta(t)x_i + \varepsilon_i(t) \quad i = 1, \ldots, n, t \in [0, 1]$$

where $n = 10$, $x_i = \frac{i-1}{n-1}$, and $\beta(t) = d \cdot f(t)$, with d ranging from 0 to 5. We modelled the function $f(t)$ with a cubic B-spline expansion with 40 basis functions and equally-spaced knots. The first h coefficients of the expansion were set to one, and the last $40 - h$ coefficients to zero. The resulting function assumed value 1 in the first part of the domain, 0 in the second part of the domain, with a smooth transition. We explored three values of the parameter h: $h \in \{10, 20, 30\}$.

The error functions $\varepsilon_i(t)$ were obtained simulating the coefficients of the same cubic B-spline expansion. The 40 coefficients were sampled independently from a standard normal distribution. The three panels in the first column of Figure 2 show an instance of the simulated functional data with $d = 5$, and $h = 10, 20, 30$, respectively. The functional data are colored in a gray scale that is proportional to the value of x_i. The study is much similar to the simulation study presented in [1], but here it is of interest to vary the domain where the null hypothesis is true.

The fBH and Fmax procedures are applied to test the following hypotheses:

$$H_t^0 : \beta(t) = 0; \quad H_t^1 : \beta(t) \neq 0.$$

The unadjusted p-value at point t was computed with a permutation test based on the *Freedman and Lane* method [8].

With $d \in \{0, 1, \ldots, 5\}$ and $h \in \{10, 20, 30\}$, we had 18 scenarios in total, but only 16 different ones, as the scenarios were identical for $d = 0$.

Simulation results We measured the performances of the two methods by evaluating the FWER, FDR, false positivity rate (i.e., the measure of the incorrectly rejected part of the domain over the measure of the domain where the null hypothesis is true) and sensitivity (i.e., the measure of the correctly rejected part of the domain relative to the total measure of the domain where the null hypothesis is false).We performed the tests at nominal level $\alpha = 0.05$.

Figure 2 reports the results of the simulation obtained averaging over 1000 instances. Each row of the figure report the results with a different value of h. Each panel on columns 2-5 reports one of the measures discussed before as a function of the parameter d for the unadjusted p-value (black), the fBH procedure (dark grey) and the Fmax procedure (light grey).

As expected by theory, the unadjusted p-value function does not control the FWER nor the FDR. It controls instead the false positive rate. The Fmax method controls the FWER, and by consequence, also the FDR. Finally, the fBH method controls the FWER only weakly

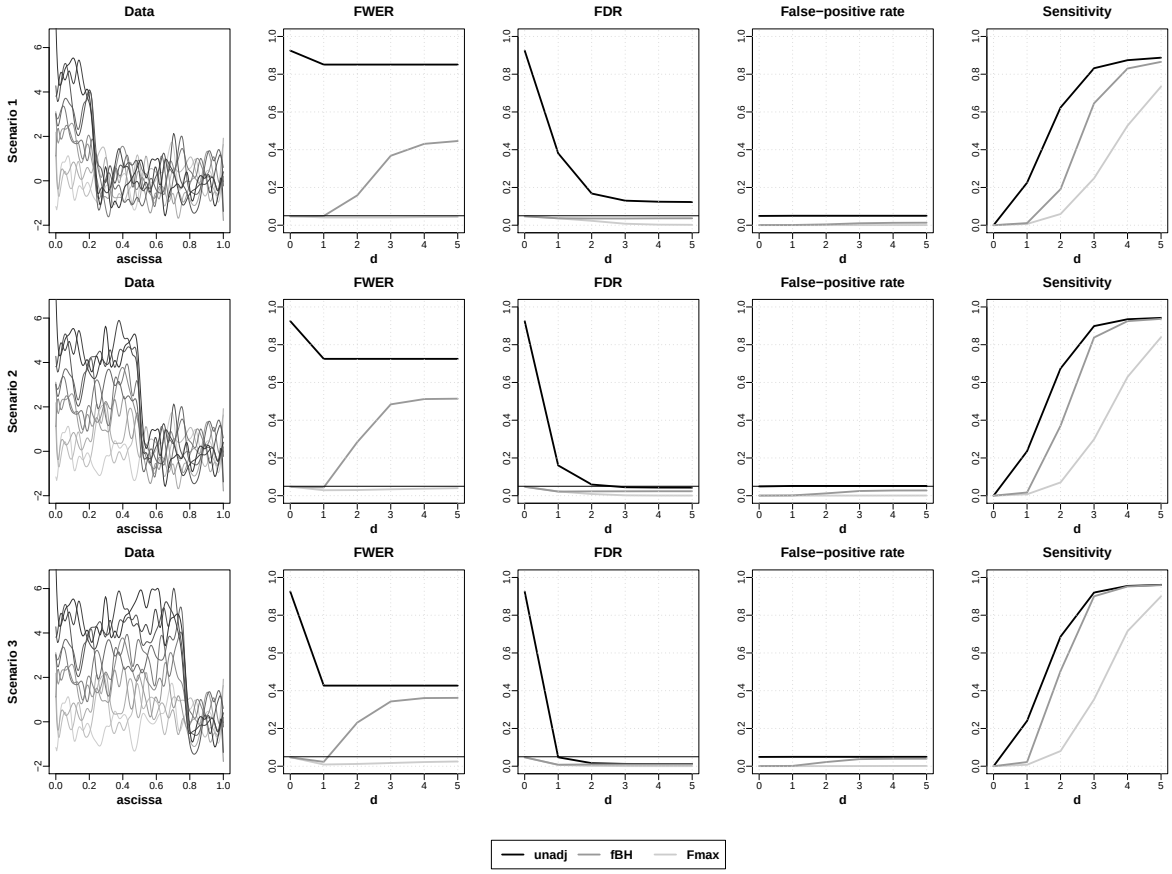


Figure 2: Simulation study: Familywise error rates, false discovery rates, false positive rate and sensitivity analysis of three methods for varying values of d. Scenario 1 corresponds to $h = 10$, Scenario 2 to $h = 20$, and Scenario 3 to $h = 30$. 'Data' shown in the left panels are examples of simulated data for $d = 1$.

(i.e., when $d = 0$ and by consequence H_t^0 is true for all t). It controls instead the FDR in all scenarios. Finally, we notice a trade-off between the type of control and the sensitivity of the methods. The Fmax, being provided with a stronger control, is also the less sensible to deviations from the null hypothesis. Instead, the fBH is provided with a less strong control, but it is more sensible.

4.2 2D-simulation

Description of simulation The base signal θ for this simulation consisted of 9 conical spikes with height $h = 1$ and diameter $d = 0.2$ arranged on the grid $\{0.25, 0.50, 0.75\}^2$ with the unit square $\mathbf{T} = [0, 1]^2$ as domain. Five spikes had positives values and the remaining four

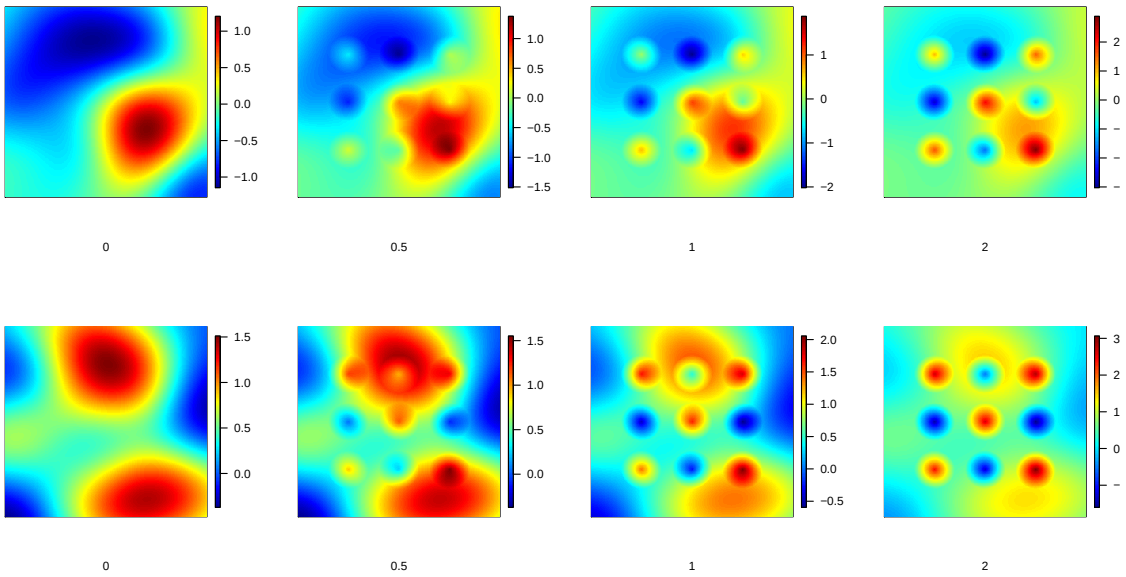


Figure 3: Two simulated fields with increasing signal strength

spikes had negative values. On top we added an error signal generated as a smooth Matérn field x_i with varying scale parameter. The observed signal is $y_i = \theta + x_i$ for $i = 1, \ldots, N$. See Figure 3 for an illustration.

The simulation was inspired by [6], who also studies false discovery rates in a setting of random fields, albeit with a quite different scope. We tested against the pointwise null hypothesis $H_0(t) : \theta(t) = 0$ using a pointwise, one-sided t-test over a fine lattice.

The Benjamini-Hochberg procedure was applied, and false positive rates (FPR), false discovery rates and sensitivity values were evaluated, along with false discovery rate for the unadjusted p-values for comparison. We simulated 2500 samples of the error process at grid 255×255. We assessed performance of the test in five different setups with varying strengths of the base signal and varying numbers of samples per t-test. Samples of the error process were recycled and used for all experimental setups, this is also the reason for the varying number of replications in table 1.

We remark that the observed data will show a "bend" at the edges of the spikes, making these easily detectable by other methods. However, the use of pointwise tests will ignore such features.

Simulation results We tested at various thresholds, $\alpha \in \{0.001, 0.01, 0.02, 0.03, 0.04, 0.05, 0.10\}$ for five different experimental setups described in Table 1; results are shown in Figure 4. The sensitivity values and false positive rates are defined as proportions of rejected/accepted

Setup no.	Signal size	# of samples per test	Replications	Sensitivity	FPR	FDR	FDR, unadjusted
1	2.0	20	125	0.498	0.00785	0.0252	0.113
2	2.0	10	250	0.226	0.00553	0.0257	0.135
3	2.0	40	62	0.654	0.00810	0.0242	0.117
4	1.0	20	125	0.116	0.00467	0.0249	0.155
5	0.5	20	125	0.0065	0.00295	0.0258	0.235

Table 1: Results from simulation at 5% threshold with 255x255 sample points. The varying number of replications is due to the upper limit of 2500 signals in total

hypotheses to the total number of false and correct hypotheses, respectively, after p-value adjustment. $\mu(U) = 71.7\%$ of the signal was zero, thus we would expect FDR to be controlled by 0.717α; for the 5% threshold of table 1, this is roughly 0.036.

Unsurprisingly, power and significance levels and FDR increase with α. Experiment 5 has almost no power, even at $\alpha = 0.10$, but still a comparatively large FDR, indicating that the Benjamini-Hochberg procedure is not too conservative in this setup. The false discovery rate is remarkably stable across experimental setups, and shows a linear tendency. As expected, FDR is well below α in all instances, and also below $\mu(U)\alpha$. In comparison, the FDR for the unadjusted p-values exceeds α in all setups except one, and varies considerably with the experimental setting.

5 Application: Analysis of Climate Data

Climate change is a huge issue, both politically and scientifically. The main issue are increasing temperatures with many adverse effects on weather and climate. Knowing that temperature has increased significantly on a global scale, we wanted to test where on Earth temperature has increased.

5.1 Data and model

Data consists of yearly averages of temperatures, starting in 1983 and ending in 2007, for each $1^\circ \times 1^\circ$ tile on Earth, using standard latitudes and longitudes. Temperatures are satellite measurements collected by NASA. These data were obtained from the NASA Langley Research Center Atmospheric Science Data Center Surface meteorological and Solar Energy (SSE) web portal supported by the NASA LaRC POWER Project. [2]

One crucial feature is that data was more densely sampled closer to the poles than close to Equator. Naturally, data also exhibits behaviour depending on local geography. Rather

[2] http://eosweb.larc.nasa.gov

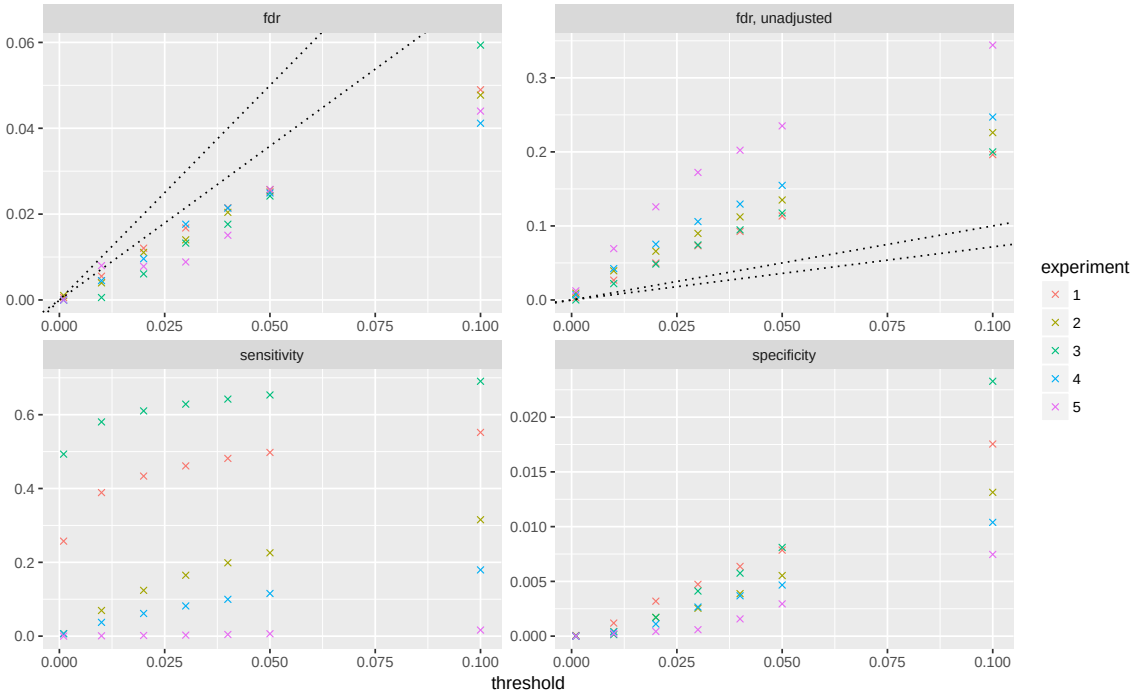


Figure 4: Average power, significance levels and FDR for different levels of α in five experimental setups as described in Table 1, along with FDR for the unadjusted p-values. The dotted lines have slopes 1 and 0.717, respectively.

Significance level	Unadjusted p value	fBH-adjusted p value
0.10	0.410	0.229
0.05	0.324	0.150
0.01	0.178	0.062
0.001	0.074	0.023

Table 2: Areas of significance of the correction methods at various significance levels as percentage of Earth total

than viewing data as truly areal, we considered data to sit on the midpoints of the tiles, e.g. $(73.5°S, 24.5°W)$ corresponds to the tile $(74°S, 73°S) \times (24°W, 25°W)$.

We applied the linear regression model $y_{st} = a_s + b_s \text{year}_t + \epsilon_{st}, s \in S^2, t \in \{1983, \ldots, 2007\}$, testing for *positive trend*, i.e. $H_s^0 : b_s = 0$ with alternative hypothesis $H_s^A : b_s > 0$. Figure 5 displays the temperature changes, values of the pointwise t-statistics and corresponding (unadjusted) p-values. Observe how much the test statistic varies across the globe, and how differences between land and ocean are more visible in the lower the plot compared to the upper plot.

To perform the BH procedure, we mapped the sphere into $\mathbf{T} = (-\pi, \pi) \times (-\pi/2, \pi/2)$ by (scaled) polar coordinates, ie. longitude and latitude. This mapping gives rise to a measure $\nu = f \cdot \mu$ on $\mathbf{T}$ where f is proportional to cos(latitude) cf. Remark 8. This measure gives uniform weights to all points on Earth, assuming Earth to be a perfect sphere. One-sided t-tests were used for obtaining unadjusted p-values. We used the same grid as the observations for approximating the BH procedure, in total $180 \times 360 = 64800$ grid points.

5.2 Results

The coverage areas at various significance levels are provided in Table 2.

Although more conservative by construction than unadjusted p-values, the fBH-adjusted p-values still retained large significant areas, indicating that the temperature increase observed in these areas is very unlikely to be a coincidence in but a small fraction of these areas. If we take look at the map, the North Atlantic Ocean and northern China stands out; it is evident/clear that these regions have experienced temperatures far above the normal in the latest years with the adverse weather effects this may cause.

6 Discussion

We successfully defined false discovery rates for functional data with for generic subdomains of $\mathbb{R}^k$, and by remark 8, this is easily extended to non-euclidean domains such as the sphere used in the data application. Furthermore, we devised a correction method for controlling FDR,

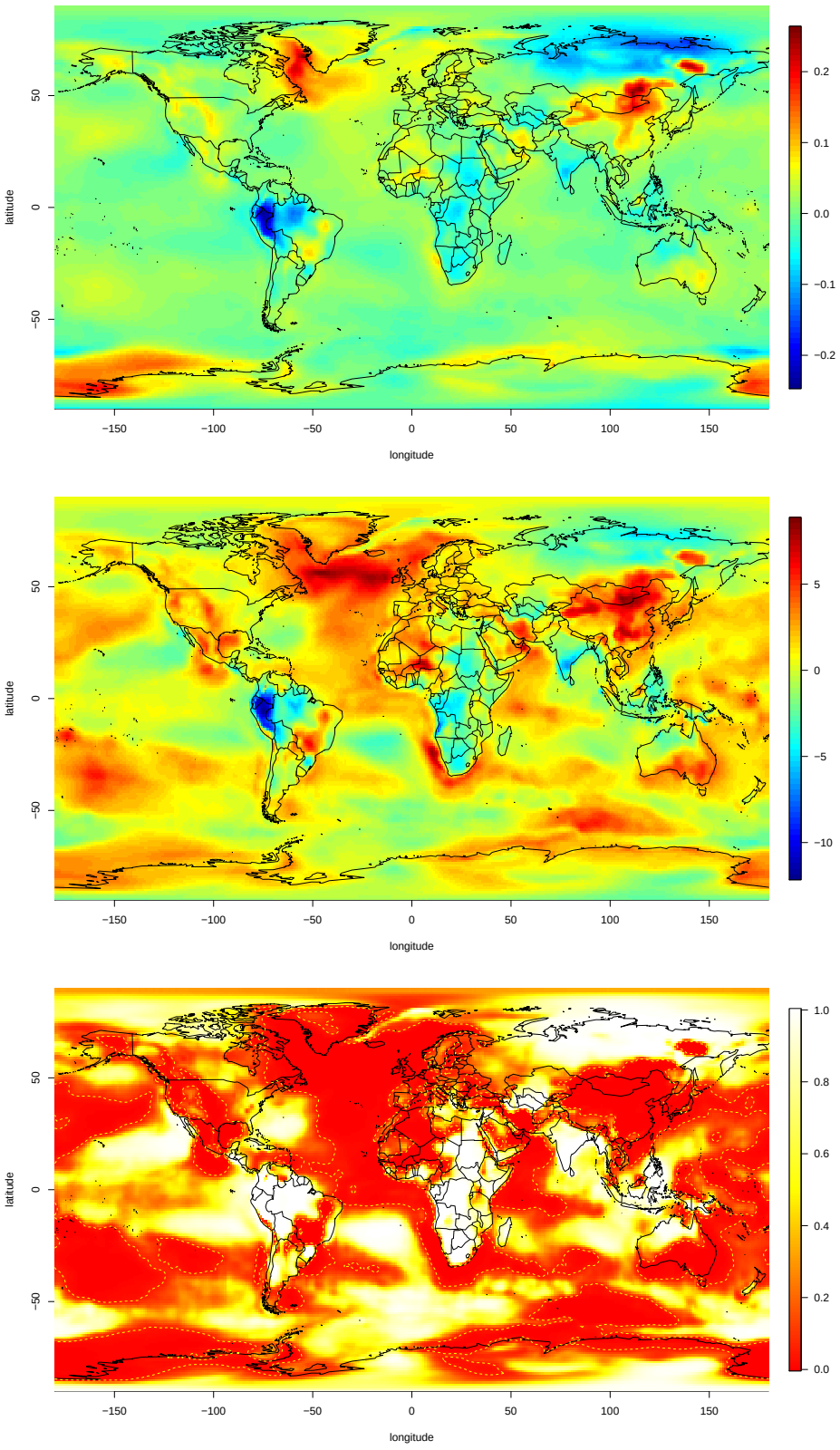


Figure 5: Above: Average yearly temperature change in centigrades, 1983-2007. Middle: Pointwise values of test statistic (t-distribution, 24 degress of freedom). Below: Unadjusted p-values with marking of the 5% threshold.

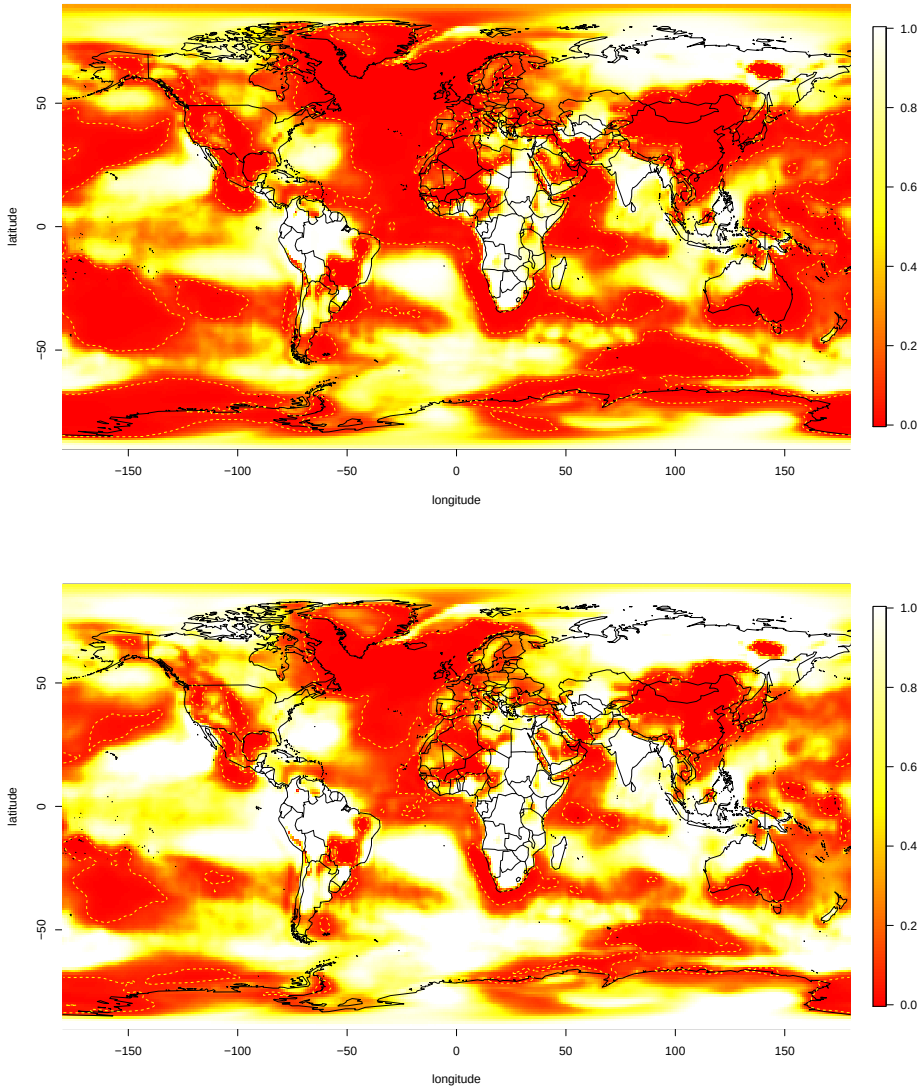


Figure 6: Upper plot: Unadjusted p-values. Lower plot: FDR-adjusted p-values. Dashed lines indicate 5% significance levels.

the *functional BH procedure*, along with an adjusted p-value function. The fBH procedure was successfully applied in two simulations and on a data set on climate change.

Two simulation studies allowed us to study the FBH procedure in a setting where the distributions and true null hypotheses were known. The false discovery rate was controlled by FBH procedure in all instances, unlike the unadjusted p-value functions. The signal-to-noise proved to be important: the sensitivity increased with signal strength, but the false discovery rate was remarkably constant. This is a desirable property, and we believe it holds true generally. The functional BH procedure is easily applicable, although it should be noted that as generally is the case in functional data analysis, it depends on how the functions are approximated/smoothed from data, and there are also some computational issues.

We demonstrated the applicability of the method and also gained insight into which regions of Earth that have seen temperature increases due to global warming in a recent time span. More advanced models and tests may further increase our understanding of local temperature changes in connection to warming, but we leave this as future work.

We would like to stress the minimal assumptions required of the fBH approach: the dependence structure of functional data such as the Earth climate data can be complex and difficult to model. However, our approach does not require specific modelling of the covariance structure of the data, we merely require a certain degree of positive association among p-values.

Due to its simple applicability, general setting and easy understanding, we expect functional FDR to have a great potential as a tool for local inference in functional data analysis. The functional BH procedure require only little computational power, and should at most be a minor issue in applications.

Functional FDR and the BH procedure suffer from some of the same issues as the discrete version. As noted in Section 2, many authors have proposed methods or quantities to deal with multiple testing. Given the success of formulating FDR and the BH procedure in a functional framework, it is likely that some of these other methods/quantities can be expanded to the functional case as well.

References

[1] Konrad Abramowicz, Charlotte K. Häger, Alessia Pini, Lina Schelin, Sara Sjöstedt de Luna, and Simone Vantini. Nonparametric inference for functional-on-scalar linear models applied to knee kinematic hop data after injury of the anterior cruciate ligament. *Scandinavian Journal of Statistics*, 2018.

[2] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[3] Yoav Benjamini and Yosef Hochberg. Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics*, 24(3):407–418, 1997.

[4] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, pages 1165–1188, 2001.

[5] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

[6] Dan Cheng, Armin Schwartzman, et al. Multiple testing of local maxima for detection of peaks in random fields. *Annals of Statistics*, 45(2):529–556, 2017.

[7] Bradley Efron, Robert Tibshirani, John D. Storey, and Virginia Tusher. Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96(456):1151–1160, 2001.

[8] D. Freedman and D. Lane. A nonstochastic interpretation of reported significance levels. *J. Bus. Econ. Stat.*, 1(4):292–298, 1983.

[9] Philipp Heesen, Arnold Janssen, et al. Inequalities for the false discovery rate (fdr) under dependence. *Electronic Journal of Statistics*, 9(1):679–716, 2015.

[10] Andrew P. Holmes, R.C. Blair, J.D.G. Watson, and I. Ford. Nonparametric analysis of statistic images from functional mapping experiments. *Journal of Cerebral Blood Flow & Metabolism*, 16(1):7–22, 1996.

[11] Lajos Horváth and Piotr Kokoszka. *Inference for functional data with applications*, volume 200. Springer Science & Business Media, 2012.

[12] Alessia Pini and Simone Vantini. Interval-wise testing for functional data. *Journal of Nonparametric Statistics*, 29(2):407–424, 2017.

[13] J.O. Ramsay and B.W. Silverman. *Functional Data Analysis*. Springer, second edition, 2005.

[14] Armin Schwartzman, Yulia Gavrilov, and Robert J Adler. Multiple testing of local maxima for detection of peaks in 1d. *Annals of Statistics*, 39(6):3290, 2011.

[15] John D. Storey. The positive false discovery rate: a bayesian interpretation and the q-value. *Annals of Statistics*, 31(6):2013–2035, 2003.

[16] Wenguang Sun, Brian J. Reich, T Tony Cai, Michele Guindani, and Armin Schwartzman. False discovery control in large-scale spatial multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(1):59–83, 2015.

[17] Anderson M. Winkler, Gerard R. Ridgway, Matthew A. Webster, Stephen M. Smith, and Thomas E. Nichols. Permutation inference for the general linear model. *Neuroimage*, 92:381–397, 2014.

Appendix

A Proof of Proposition 12 and Proposition 14

A.1 Assumptions

We begin by repeating the assumptions of Theorem 12 and Theorem 14.

Definition (PRDS)**.** Let '$\leq$' be the partial/usual ordering on $\mathbb{R}^l$. An *increasing set* $D \subseteq \mathbb{R}^l$ is a set satisfying $x \in D \wedge y \geq x \Rightarrow y \in D$.

A random variable $\mathbf{X}$ on $\mathbb{R}^l$ is said to be *PRDS on* I_0, where I_0 is a subset of $\{1, \ldots, l\}$, if it for any increasing set D and $i \in I_0$ holds that

$$x \leq y \Rightarrow P(\mathbf{X} \in D | X_i = x) \leq P(\mathbf{X} \in D | X_i = y) \tag{20}$$

Let $\mathbf{Z}$ be an infinite-dimensional random variable, where instances of $\mathbf{Z}$ are functions $T \to \mathbb{R}$. We say that $\mathbf{Z}$ is PRDS on $U \subseteq T$ if all finite-dimensional distributions of $\mathbf{Z}$ are PRDS. That is, for all finite subsets $I = \{i_1, \ldots, i_l\} \subseteq T$, it holds that $Z(i_1), \ldots, Z(i_l)$ is PRDS on $I \cap U$.

Let $\{S_k\}_{k=1}^{\infty}, S_1 \subset S_2 \subset \ldots$ be a dense, uniform grid in $\mathbf{T}$ in sense that S_k uniformly approximates all level sets of p and $p|_U$ with probability one.

For Theorem 12 that amounts to,

$$P\left[\lim_{k\to\infty} \sup_r \frac{\#(S_k \cap \{s : p(s) \leq r\})}{\#S_k} - \mu\{s : p(s) \leq r\} \to 0\right] = 1 \tag{21}$$

and

$$P\left[\lim_{k\to\infty} \sup_r \frac{\#(S_k \cap \{s : p(s) \leq r\} \cap U)}{\#S_k} - \mu(\{s : p(s) \leq r\} \cap U) \to 0\right] = 1 \tag{22}$$

whereas for Theorem 14, we need the density function f:

$$P\left[\lim_{k\to\infty} \sup_r \frac{\sum_{i \in S_k \cap \{s:p(s)\leq r\}} f(i)}{\#S_k} - \int_{\{s:p(s)\leq t\}} f(x)\,\mathrm{d}x \to 0\right] = 1 \tag{23}$$

and

$$P\left[\lim_{k\to\infty} \sup_r \frac{\sum_{i \in S_k \cap \{s:p(s)\leq r\} \cap U} f(i)}{\#S_k} - \int_{\{s:p(s)\leq t\}\cap U} f(x)\,\mathrm{d}x \to 0\right] = 1 \tag{24}$$

Furthermore, assume that p is PRDS wrt. the set of true null hypotheses with probability one, and that the assumptions about p-value function below hold true with probability one:

(a1) All level sets of p have zero measure,

$$\mu\{s : p(s) = t\} = 0 \quad \forall t \in \mathbf{T}$$

(a2) $\alpha^* \in (0, \alpha] \Rightarrow$: for any open neighbourhood O around α^* there exists $s_1, s_2 \in O$ s.t. $a(s_1) > \alpha^{-1} s_1, a(s_2) < \alpha^{-1} s_2$, where a is the cumulated p-value function (Definition 9).

(a3) $[\alpha^* = 0] \Rightarrow \min p(t) > 0$.

A.2 Proof details

For the ease of presentation, we will only consider Theorem 12. The proof of Proposition 14 is analogous but notationally tedious, as the counts are replaced by sums and the measures by integrals.

Let a_k be the cumulated p-value function for the k'th iteration of the BH procedure:

$$a_k(t) := N_k \#\{s \in S_k : p(s) \leq t\} \tag{25}$$

and define the k'th step false discovery proportion Q_k by applying the (usual) BH procedure at level q to p evaluated in S_k:

$$Q_k = \frac{\#\{t \in S_k : p(t) \leq b_k\} \cap U}{\#\{t \in S_k : p(t) \leq b_k\}}, \quad b_k = \arg\max_r \frac{\#\{s \in S_k : p(s) \leq r\}}{\#S_k} \geq \alpha^{-1} r \tag{26}$$

or equivalently $b_k = \arg\max_t a_k(t) \geq \alpha^{-1} r$.

Lemma 16. a_k converges to a uniformly as $k \to \infty$.

Proof: Follows from assumption (12) and definitions of a_k and a.

Lemma 17. b_k converges to α^* as $k \to \infty$

Proof. By Lemma 16, a_k converges uniformly to a. There are two cases: $\alpha^* = 0$ and $\alpha^* \in (0, \alpha]$.

Case 1, $\alpha^* = 0$: Let O be any open neighbourhood around zero. O^C (where the complement is wrt. [0,1]) is a closed set that satisfies $a(t) < \alpha^{-1} t$. By continuity of a there exists an $\epsilon > 0$ s.t. $a(t) < \alpha^{-1} t - \epsilon$ for all $t \in O^C$. As a_k converges uniformly to a, eventually for large enough k, $a_k(t) < \alpha^{-1} t$ for $t \in T \backslash O$, and thus $b_k \in O$ eventually. This was true for any O, and we conclude $b_k \to 0$.

Case 2, $\alpha^* \in (0, \alpha]$: By assumption, for any open neighbourhood $O \ni \alpha^*$, there exist $s_1, s_2 \in O$ s.t. $a(s_1) > \alpha^{-1} s_1$, $a(s_2) < \alpha^{-1} s_2$.

For $t > \alpha^*, t \notin O$, we have $\alpha^{-1} t - a(t) > \epsilon$ for some $\epsilon > 0$ by continuity of a. Hence by uniform convergence, it must hold that for k sufficiently large we have $a_k(t) < \alpha^{-1} t$ for $t > \alpha^*, t \notin O$. This was true for any O, and we conclude $\limsup b_k \leq \alpha^*$.

Conversely, we can show that $\liminf b_k \geq \alpha^*$, and thus $\lim b_k = \alpha^*$.

□

Define $A_k = \#\{t \in S_k : p(t) \leq b_k\}$, and define Q_k as the false dicovery proportion for the k'th iteration:

$$Q_k := \frac{\#(A_k \cap U)}{\#A_k} 1_{A_k \neq \emptyset} \tag{27}$$

Rejection areas Now we intend to prove that $H_{t,k}$ converges eventually. Note that $p(t)$ is independent of k, and that $H_{t,k} = (t \in S_k) \cap (p(t) \leq b_k)$, i.e. the event that the BH threshold at step k is larger than $p(t)$.

Proposition 18. For all t that satisfies $p(t) \neq \alpha^*$, $H_{t,k}$ converges eventually.

Proof. First note that if $t \notin S_k$ for all k, then $H_{t,k}$ is trivially zero for all k. So assume $t \in S_{k_0}$ for some k_0. As $k \to \infty$, $b_k \to \alpha^*$, and by assumption $p(t) \neq \alpha^*$. Eventually, as $k \to \infty$, $p(t)$ is either strictly larger or strictly smaller than b_k, proving the result. □

Convergence of Q_k Finally we need to show that $Q_k \to Q$. We show this, by proving convergence of the nominator and denominator, and arguing that $Q = 0$ implies that Q_k is 0 eventually.

Define $H^0 = \{t | p(t) > \alpha^*\}$, ie. the acceptance region, and $H^1 = T \backslash H^0$, the rejection region. Note that $\mu(H^1) = a(\alpha^*) = \alpha^{-1}\alpha^*$.

Note that $H_1 = V \cup S = \{t : p(t) \leq \alpha^*\}$ and $H_1 \cap U = V$.

Proposition 19. $N_k \# A_k \to \mu(H^1)$ and $N_k \#(A_k \cap U) \to \mu(H^1 \cap U)$.

Proof. For k, define $J_k = \{t : p(t) \leq b_k\}$. Note that $A_k = J_k \cap S_k$. Observe that by assumption about uniform convergence on levels sets (equation (12)):

$$N_k \#(J_k \cap S_k) - \mu(J_k) \to 0 \text{ for } k \to \infty.$$

Next observe that due to (1) continuity of a, (2) $b_k \to \alpha^*$ and (3) the fact that we are considering sets on the form $\{t : p(t) \leq x\}$, we are able to conclude that

$$\mu(J_k \triangle H^1) \to 0 \text{ for } k \to \infty. \tag{28}$$

and we conclude $N_k \# A_k \to \mu(H^1)$.

For the second part, observe that (equation (13))

$$N_k \#(U \cap J_k \cap S_k) - \mu(U \cap J_k) \to 0 \text{ for } k \to \infty.$$

We just argued that $\mu(J_k \triangle H^1) \to 0$. It remains true when "conditioning" on a measurable set, in this case U:

$$\mu((J_k \cap U) \triangle (H^1 \cap U)) \to 0 \text{ for } k \to \infty. \tag{29}$$

and we conclude $N_k \#(A_k \cap U) \to \mu(H^1)$. □

For $\alpha^* = 0$ we have the following stronger result:

Lemma 20. If $\alpha^* = 0$, then $\#A_k = 0$ eventually.

From this lemma it follows that $N_k \#A_k = 0$ (and thus Q_k as well) eventually.

Proof. Since $\alpha^* = 0$, $a(t) < \alpha^{-1} t$ for all $t > 0$. By assumption, $\min p(t) > 0$, and thus $a_k(s) = 0$ for $s < \min p(t)$ and all k.

By continuity of a, it follows that there exists $\epsilon > 0$ s.t. $\alpha^{-1} t - a(t) > \epsilon$ on the interval $[\min p(t), 1]$, and by uniform convergence of a_k we get that for large enough k, $a_k(t) < \alpha^{-1} t$ for all $t \geq \min p(t)$.

Combining with $a_k(t) = 0$ for $t < \min p(t)$, we get that eventually $a_k(t) < \alpha^{-1} t$ for every $t > 0$ and thus $b_k = 0$. From this (remember $\min p(t) > 0$) we conclude that all hypotheses are rejected eventually, ie. $\#A_k = 0$. □

Theorem 21. Q_k converges to Q almost surely, and $\limsup_{k \to \infty} \mathrm{E}[Q_k] \leq \alpha \mu(U)$.

Proof. By Lemma 20, Q_k converges to Q when $\alpha^* = 0$, and by Proposition 19 Q_k converges to Q when $\alpha^* > 0$ since $\mu(H^1) = \alpha^* / \alpha > 0$.

Applying Benjamini and Yekuteli's original proposition, Theorem 6, (now we use the PRDS assumption), we have $\mathrm{E}[Q_k] \leq \alpha N_k \#(S_k \cap U)$ for all k. By setting $r = 1$ it follows from (13) that $\lim_{k \to \infty} N_k \#(S_k \cap U) = \mu(U)$, and hence $\limsup_{k \to \infty} \mathrm{E}[Q_k] \leq \alpha \mu(U)$.

□

IV

Statistical modelling of conidial discharge of entomophthoralean fungi using a newly discovered *Pandora* species

Niels Lundtorp Olsen
Department of Mathematical Sciences
University of Copenhagen

Pascal Herren
Department of Plant and Environmental Sciences
University of Copenhagen

Bo Markussen
Department of Mathematical Sciences
University of Copenhagen

Annette Bruun Jensen
Department of Plant and Environmental Sciences
University of Copenhagen

Jørgen Eilenberg
Department of Plant and Environmental Sciences
University of Copenhagen

Publication details
Ready for submission.

Statistical modelling of conidial discharge of entomophthoralean fungi using a newly discovered *Pandora* species

Niels Olsen, Pascal Herren,
Bo Markussen, Annette Bruun Jensen, Jørgen Eilenberg

Abstract

Entomophthoralean fungi are characterized by discharging conidia that infect insects. The temporal pattern of conidial discharge is a crucial for or understanding of the epizootic development and biological control potential.

Mycelia from a newly discovered *Pandora* species were incubated at various temperatures in darkness, and conidial discharge was measured at regular intervals. The aim is to study the effects of temperature on the discharge and to characterize the variation in the associated temporal pattern with focus on peak location and shape.

We use a novel modification of a statistical model, that simultaneously estimate warping and amplitude effects, into a setting of generalized linear models. This model is used to analyse data, and to test hypotheses of peak location and discharge.

Our findings show that high temperature leads to an early and fast decreasing peak, whereas there are no significant differences in total number of discharged conidia.

1 Introduction

There are many examples of shooting mechanisms in living organisms. Among purposes for shooting mechanisms is reproduction, for example in fungi [1], and there are many different names for the various mechanisms. [1] list a range of shooting mechanisms in fungi allowing spores (for example conidia) to be discharged. One category "fluid pressure catapult" seems designed to allow fungi to convert elastic energy into kinetic energy, ensuring that spores are discharged at sufficient speeds. The large conidia (mostly between 15 and 40 microns in length) in fungus order Entomophthorales demands high energy to be discharged. They are discharged by a rapid pressure-driven eversion of the septum between the spore and the conidium [2]. Fungi from Entomophthorales are insect or mite pathogens and their infection success depends, among other things, on the attachment of the discharged conidium after landing on host cuticle [3]. The conidia of entomophthoralean fungi are discharged with fluid from the conidiophore, which further assist the conidium to stick to host cuticle after landing [2, 4].

The temporal pattern of conidial discharge from infected and dead hosts have been studied for several species of Entomophthorales belonging to for example the genera *Entomophthora*, *Entomophaga*, *Pandora* and *Zoophthora* [4, 5, 6, 7, 8]. The studies show the same overall pattern: after a lag phase of a few hours after the death of the host, conidial discharge is initiated. Depending on host species, fungus species, and temperature, the peak in discharge intensity will be reached within one or two days, thereafter the intensity drops although conidia may still be produced and discharged several days after death of the host. In principle the same pattern appears when conidia are discharged from in vitro cultures. Here the starting point will be, when a mycelium mat is transferred to a plate, from where conidial discharge will be initiated.

Due to the sticky conidia, it is a methodological challenge to study patterns (at a quantitative level) over time of conidial discharge in Entomophthorales, and people have used various methods to collect and count discharged conidia. In [9] different methodologies applying to Entomophthoralean fungi are reviewed, and a common trait is that the setup should as much as possible reflect the natural condition, where insects are killed and thereafter initiate discharge of conidia. Different laboratory setups have been used for obtaining discharged conidia counted on glass slides referring to specific time intervals and/or different distances [4, 6, 10]. The data treatment in studies on conidial discharge is mostly rather simple and include for example calculations of mean and standard deviation for replicates, pair wise comparisons or analysis of variance, and a description in words about peak of intensity and length of period with conidial discharge. While these methods are valid and may offer a fair background for conclusions, they nevertheless do not harbor and by that make use of the total information in the study.

Entomophthoralean conidia (fig. 1, right shows a conidia of *Pandora* sp. from *Cacopsylla* sp.) are the infective units of entomophthoralean fungi, and for the majority of species they

are actively discharged [11]. The conidial discharge of different species of Entomophthorales is affected by temperature [12, 6, 13, 14]. In *P. neoaphidis*, the total number of discharged conidia is significantly lower at extreme temperatures below 10 and above 25°C [15, 12] and highest conidial discharge from mycelium mats is at temperatures between 10°C and 20°C [13]. The duration of conidial discharge from cadavers of *Pandora nouryi*, another species infecting aphids, increased with decreasing temperature, and lasted up to 120 hours at 8°C [14]. Once the conidia of entomophthoralean fungi are discharged from the conidiophores they have a short longevity[16, 17].

Pandora fungi as potential biocontrol agents Psyllids from the genus *Cacopsylla* (Hemiptera: Psyllidae) cause economic damage to pear trees (*Pyrus spp.*) and apple trees *(Malus spp.)* in Europe. Some of these species not only cause direct damage by sucking on the phloem and secreting honeydew [18], but they also transmit *Candidatus* (Ca.) Phytoplasma pyri to pear and Ca. Phytoplasma mali to apple trees; diseases known as *pear decline* and *apple proliferation* (AP) respectively [19]. In Germany and Italy losses due to AP can be very high. [20] calculated annual losses at that time to be up to 125 million Euros.

The most common treatment of these insects is to spray eggs and nymphs with synthetic insecticides such as spinosyns or tetronic acids of up to six times per year [21]. Such extensive use of chemical insecticides leads to resistance development, and effective modes of action to control *Cacopsylla spp.* become increasingly unavailable due to resistance development and stricter regulations that must be addressed when new chemical pesticides are being registered [22]. Therefore, alternative control methods are needed.

In 2016, a potential fungal biocontrol agent of *Cacopsylla spp.* was found on psyllids from a pear orchard in Vedbæk near Copenhagen, Denmark. It was shown to have a high virulence against several *Cacopsylla spp.* and can be grown in vitro [23]. This new species is from the genus *Pandora* (Entomophthorales: Entomophthoraceae) and may appear to be an undescribed species [23]. Most entomopathogenic fungi within the order Entomophthorales have a very narrow host range and will therefore not affect non-target insects [24]. At the same time they can cause natural epizootics in many arthropod species, including pest species [25] and therefore possess attractive characters for being developed as biological control agents.

Statistical modelling of temporal evolution in biological systems using functional data For biological processes evolving over time, people often like to think in terms of idealized systems with a clear time-dependent profile. However, it often the case that different instances/replications of such processes show some variation in timing. Within statistics, this variation is commonly referred to as *temporal variation* or *phase variation*.

The usual interpretation of temporal variation is *biological time*; that is, the clock of the underlying biological system is out of synchronization with the idealised system (this may be for various reasons), but it is the same underlying processes that are taking place. A common

example is puberty for boys: healthy boys enter the pubertal stage (which has some common characteristics for all boys) at some point, but when that happens varies a lot betweens individuals.

Inferring the effect of biological time requires replications of the same experiment, and when the underlying structure is a continuous process, such data is naturally handled within the framework of functional data analysis. We believe that the effects of temporal variation are often forgotten in the biological sciences. This is sometimes a problem, as ignoring temporal variation can lead to weak or even misleading conclusions.

There are various approaches to modelling functional data with temporal variation (misaligned functional data). We intend to follow the methodology of [26], which we will refer to as the *pavpop* model (*Phase and Amplitude Variation of POPulation means*). This methodology has been used in different applications with great success [27, 28]. The main idea of [26] is simultaneous modelling of amplitude and temporal variation, where temporal variation is modelled as a spline interpolation of a latent Gaussian variable that represents temporal deviation from the idealised system. For a review of methods for handling misaligned functional data, we refer to [26, 27].

Whereas classification is often part of papers on misaligned functional data, inference in form of hypothesis testing has got little attention in misaligned functional data. In general, inference in functional data is not easy and requires either strong parametric assumptions, which can be wrong, or the use of non-parametric tests, which can be computationally difficult.

Purpose and content of this study From a biological perspective, the duration and intensity of conidial discharge of the newly discovered *Pandora* sp. are important factors for transmission of this fungus in apple and pear orchards. By understanding how temperature affects the intensity of conidial discharge, we aim at better understanding the development of epizootics, which is crucial in the development process of this fungus as a biological control agent. Furthermore we aim at getting a better understanding of the temporal evolution of condidial discharge in entomophthoralean fungi by applying dedicated methods from functional data analysis.

From a statistical perspective, all previous applications using the methodology [26] have so far dealt with continuous data, where the amplitude variation is assumed to Gaussian. The novelty of this work is the extension and application of the pavpop model to discrete data, which are generated from an unobserved biological system with temporal evolution. Furthermore, we also consider inferential questions, which is new to this methodology as well. In a more broad context, this can be seen as combining the pavpop model with generalized linear models. In this application we use a negative binomial response model; the supplementary material describes various other response models.

In this study, discharge of conidia from mycelium mats (Figure 1, left) was studied in different temperatures over time. We hypothesize that high temperature leads to an early and

fast decreasing peak when looking at (the intensity of) conidial discharge, whereas a low temperature leads to a late and more slowly decreasing peak, and that a low temperature leads to a higher total production of conidia in the first 120 hours, as compared with higher temperatures.

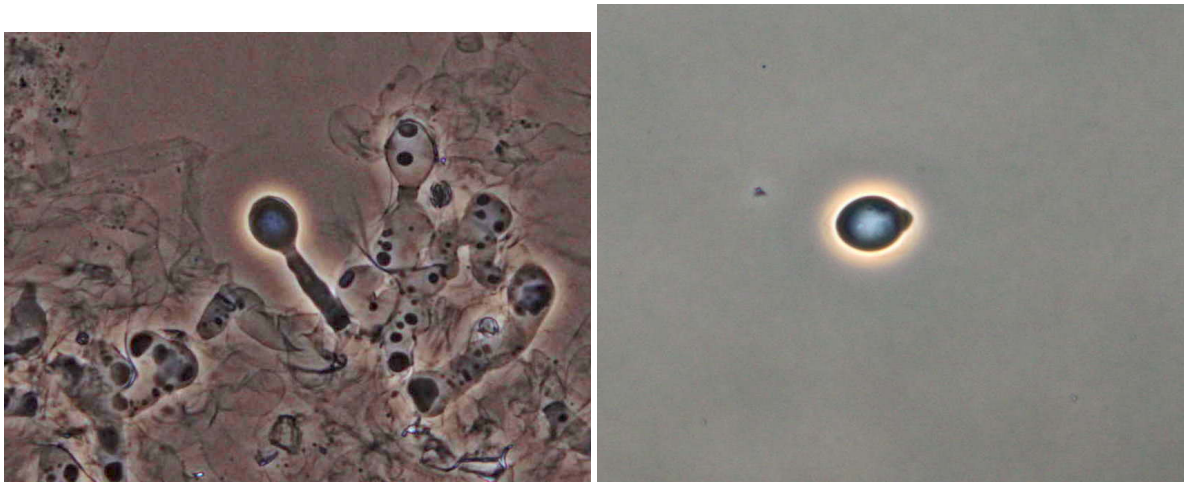


Figure 1: The mycelium of Pandora sp. from *Cacopsylla spp.* with a primary conidium on top of a conidiophore (left) and a discharged primary conidium (right).

2 Methods

2.1 Experiment and data collection

Mycelium production An isolate of *Pandora* sp., designated KVL 16-44, isolated from an infected *Cacopsylla* sp. was grown on Sabouraud Dextrose Agar (SDA) supplemented with egg yolk and milk. To produce fresh material mycelium mats were transferred to petri dishes (55 mm diameter) and incubated at 18.5 °C in dark conditions for 20 days.

Conidia production Filter papers of 18 x 18 mm, moistened with 0.75 ml of autoclaved water, were placed in the middle of lids of petri dishes (34 mm diameter). Four squares of 5 x 5 mm mycelium mats were cut from the same mycelium mat 20 mm away from the centre of the petri dish and put upside down in the edges of the moist filter papers. All lids were put on the counter parts of the empty petri dishes and they were kept in three different temperatures (12, 18.5 and 25 °C) in complete darkness at 100% RH. The mycelium mats were facing downwards. Five replicates per temperature were 15 in total.

Conidia discharge over time To measure the conidial discharge a small stripe of Parafilm with a cover slide (18 x 18 mm) on top was placed inside the lower part of the petri dish and they were placed in the incubators for 30 min. The cover slides were placed underneath the four mycelium mats. The cover slides were then removed and the conidia laying on the slide were stained with lactic acid (95%). This procedure was repeated every eight hours for 120 h, in total 16 time points with the first time point right after the transfer of the mycelium mat to the filter paper. Moreover, the lower parts of the petri dishes were cleaned with ethanol (70%) and demineralised water every eight hours to ensure that primary conidia did not discharge secondary conidia on the cover slides. The conidia were counted in each of the 4 corners of the cover slide. In total, we got four observations per time-point, replicate and temperature (4 counts * 5 replicates * 16 time-points * 3 temperatures = 960 observations). Conidia on slides were counted with the aid of a light microscope (OLYMPUS) at x 400 magnification on the whole field of view.

2.2 Statistical modelling

We consider a set of N unobserved or latent *mean curves*, $u_1, \ldots, u_N : [0,1] \to \mathbb{R}$. from $J = 3$ treatment groups. The mean curves are assumed to be generated according to the following model

$$u_n(t) = \theta_{f(n)}(v_n(t)) + x_n(t), \quad n = 1, \ldots, N \tag{1}$$

where f maps curves into treatment groups. That is, to each subject corresponds a fixed effect (θ_j), which is perturbed in time (v_n) and amplitude x_n, both assumed to be random.

To each curve corresponds a set of discrete observation $(t_{n1}, y_{n1}), \ldots, (t_{nm_n}, y_{nm_n}) \in [0,1] \times \mathcal{Y}$ where $(t_{n1}, \ldots, t_{nm_n})$ are m_n pre-specified time points and $\mathcal{Y} \subseteq \mathbb{R}$ is the sample space for the observations.

We assume that the observations conditionally on the latent mean curves are independently generated from an exponential family with probability density function

$$p(y|\eta) = b(y) \exp(\eta \cdot y - A(\eta, y)), \quad \eta \in \mathbb{R}, y \in \mathcal{Y} \tag{2}$$

where η is the value of the latent mean curve at a given time, and y is the canonical statistic for the observations. A and b are functions defining the exponential family. We assume that $A(\cdot, y) \in C^2(\mathbb{R})$ for all y with the property that $A''_\eta(\eta, y) > 0$ for all η, and we assume that all hyperparameters describing A and b are known and fixed beforehand. More details on response models can be found in the supplementary material.

The amplitude variation x_n is assumed to be a zero-mean Gaussian process. Fixed effects are modeled using an appropriate spline basis, and phase variation is modelled by random warping functions characterised by Gaussian variables $w_n \in \mathbb{R}^{m_w}$ where $w_n = 0$ corresponds to the identity function on $[0,1]$. More details on fixed effects and phase variation can be found in the supplementary material.

Estimation in this model is presented in the supplementary material. Estimation is a major challenge, as direct inference is not feasible due to the large number of latent variables. Furthermore, unlike [27], the response is not Gaussian, which require some additional considerations. We propose to use a twofold Laplace approximation for doing approximate maximum likelihood estimation; details on the Laplace approximation are found in the supplementary material.

2.3 Data analysis

As described in the methods section, data consist of 960 observations (4 counts * 5 replicates * 16 time-points * 3 temperatures) in $\mathbb{N}_0$. The largest count was 211, and a large fraction of the counts was zero.

Samples no. 7 and no. 13 "collapsed" during the experiment after 48 and 40 hours, respectively, and measurements after collapse were excluded in the data analysis.

Response model A popular choice for modelling count data (from biological experiments) are Poisson models. This is backed by a strong theoretical reasoning; using our data as an example, one would expect that while the fungi are placed in the incubators, they would independently discharge conidia at random and at a constant rate. This is a typical example of a Poisson process, for which a statistician would use a Poisson model.

However, a unique feature of the data set was the four samples taken from each batch used for conidia count, which can reasonably be assumed to be i.i.d. conditionally on the latent curve u. This allowed us to estimate the dispersion directly, and to assess if Poisson model was in reasonable agreement with the observed data. The dispersion parameter is a quantity that relates the variance to the mean. The dispersion is 1 for Poisson models and larger than 1 for negative binomial models.

By comparing sample means and sample variances across the 240 measurements, we validated if a Poisson assumption was reasonable or not. As indicated in Figure 2, this was clearly not case. Data were clearly overdispersed; a dispersion of one corresponds to the dotted line. Therefore we fitted an unstructured negative binomial regression model with common rate r to the data to investigate goodness-of-fit of using that model. This was validated; the estimated rate was $r_0 = 4.658$, the dashed line in Figure 2 indicates the corresponding mean/variance relation. This value was fixed and used in the subsequent analysis.

Having estimated the dispersion, the counts at individual measurements were added for the subsequent analysis as the sum of counts is a sufficient statistic for our model. The sum of iid. negative binomial random variables is again negatively binomially distributed; the rate parameter r is multiplied by the number of counts; thus we got $r = 4 \cdot r_0 = 18.63$. The summed counts are displayed in Figure 3.

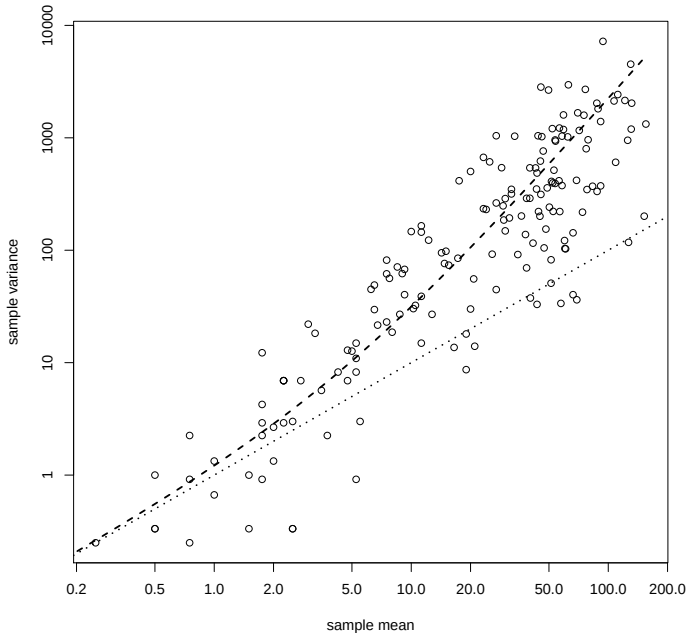


Figure 2: Sample variance as function of sample means across measurements. Dashed line is fit using a NB(4.66)-model; dotted line is fit using a Poisson model.

Model for mean curves Time was rescaled to the unit interval such that $t = 0$ corresponded to 0 hours and $t = 1$ corresponded to 120 hours. Warping functions were modelled as increasing cubic (Hyman filtered) splines with $m_w = 7$ equidistant internal anchor points with extrapolation at the right boundary point. The latent variables w_n were modeled using a Matérn covariance function with smoothness parameter $\alpha = 3/2$ and unknown range and scale parameters. This corresponds to discrete observations of an integrated Ornstein-Uhlenbeck process. This gave a flexible, yet smooth, class of possible warping functions which also take into account that the internal clocks of individual fungi could be different at the end of the experiment.

Population means $\theta_{\text{cold}}, \theta_{\text{medium}}, \theta_{\text{warm}}$ were modeled using natural cubic splines with 11 basis functions and equidistant knots in the interval $[0, 1]$. Natural cubic splines are more regular near boundary points than b-splines which reduce the effect of warping on estimation of spline coefficients.

Amplitude covariance x was modelled using a Matérn covariance function with unknown range, smoothness and scale parameters; see supplementary material for details.

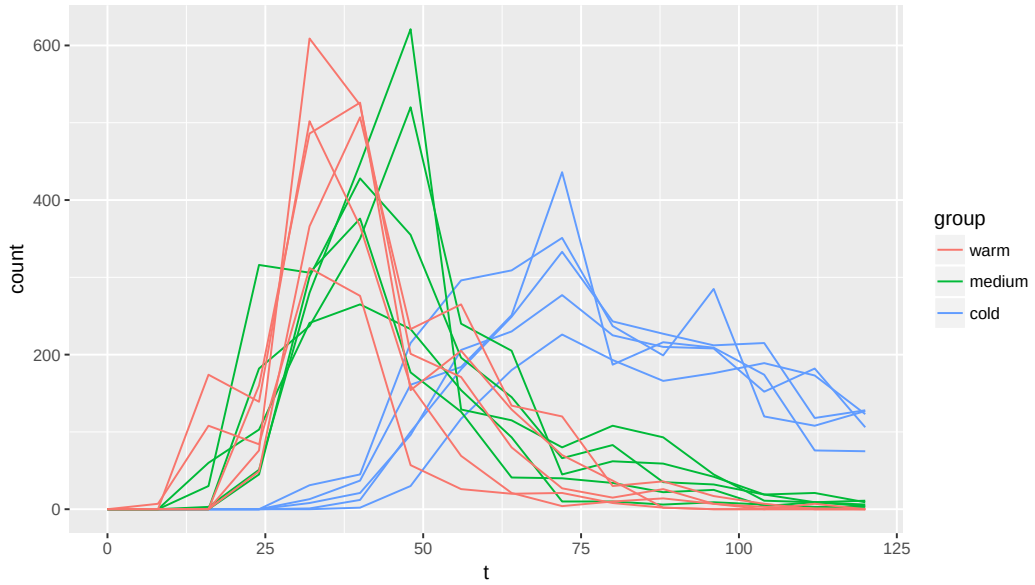


Figure 3: Summed counts of conidia for the individual fungi as functions of time, color-coded according to temperature.

Hypotheses We define 'peak location' as the time with maximal condidial discharge, and 'peak decrease' as the average decrease in discharge between 'peak location' and end of the experiment:

$$\text{peak location}_j = \arg\max \theta_j, \quad \text{peak decrease}_j = \frac{\max(\theta_j) - \theta_j(120h)}{120h - \text{peak location}_j}$$

Note that this on log-scale, so peak decrease should be interpreted as a relative decrease of conidial discharge.

One can qualitatively assess the hypotheses without strict definitions, but in order to do statistical inference, a mathematical definition is needed. We remark that here we consider population means; temporal variation may also affect peak location for individual fungi.

3 Results

Predicted mean trajectories for u, evaluated at observed time points, along with population means are displayed in Figure 4. We observe a slightly odd behaviour around $t = 0$. This is an artifice; most observations around $t = 0$ are zero. When the predicted values of u are exp-transformed, these are mapped into almost-zero values. The three population means are clearly separated, which is in concordance with our hypothesis and fit well into what we

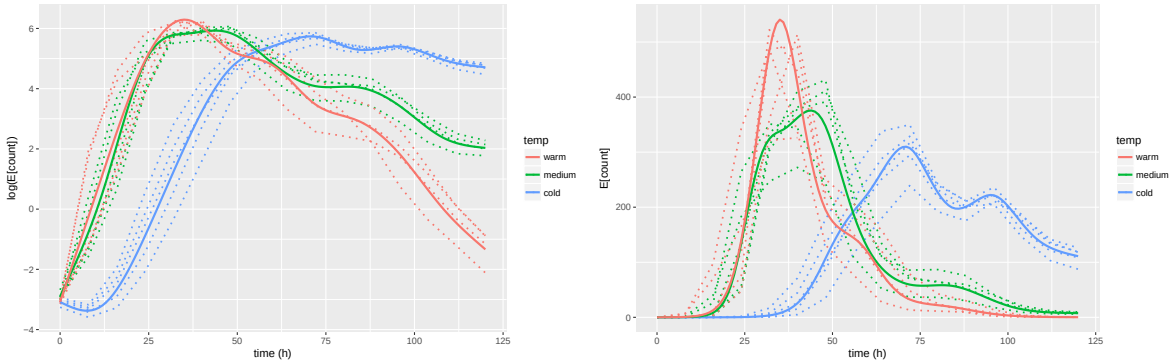


Figure 4: Predicted trajectories for u (dashed lines). Left is on model scale, right is exp-transformed (same scale as observations). Thick lines indicate estimated population means.

expected: θ_{warm} peaks first and has the highest peak; θ_{middle} is in-between and θ_{cold} peaks latest and has a smaller and more slowly decreasing peak.

Predicted warping functions are displayed in Figure 5. The scale parameter for the warp covariance was estimated to be 0.026; this corresponds to a standard deviation of around 3.1 hours on temporal displacement, or a 95% prediction interval of roughly 6 hours.

The results in Figure 5 are closely connected with those in Figure 4: a vertical change in Figure 5 corresponds to a horisontal change in Figure 4. One may interpret the trajectories in figures 4 and 5 as *smoothing* of the data: Figure 3 shows the raw data counts; Figure 4 displays the smoothed curves, which are our guesses/predictions of the intensity of conidial discharge (the underlying biological quantity of interest) for individual fungi; and finally Figure 5 displays the corresponding predictions of the biological times, the temporal deviations which for an individual fungi essentially determine the intensity of conidial discharge.

The trajectory for an indivual fungus is of little interest by itself as that fungus is confined to this experiment. However, when the trajectories are viewed together, they illustrate the variation on population level allowing us to assess variation between individual fungi from the same treatment group, and also to compare this to fungi from other treatment groups.

Discharge of conidia above certain levels For practical applications it is relevant to know when the intensity of conidia discharges reach a given level and for how long this happens. Although one conidium is enough to infect an insect [29], the chance of a conidium landing on an insect is small. Therefore we chose a range from low to very high discharge of conidia. The lowest threshold was 0.5 and and the largest threshold was 5.5 with a step size of 0.5. One corresponds to an increase in conidia discharge of $\approx 65\%$. Using the results of the analysis, we simulated trajectories of u from the model. For a given trajectory and threshold, we measured the first time this threshold was reached, and for how long u remained above this level.

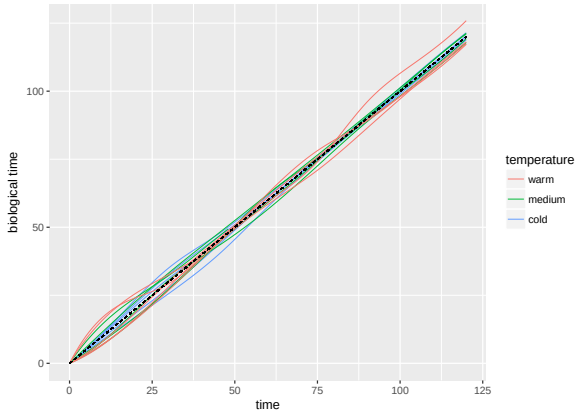


Figure 5: Predicted warp functions. Black line indicates the identity, ie. no temporal deviation.

The results are seen in Figure 6. There are generally some large variations, but according to the results, fungi at low temperatures are consistently slower at reaching the threshold. It should be noted duration is only counted until end of experiment (120 h) so the actual duration values for cold fungi could be larger when viewed over a longer time span.

Total conidia discharge The total number of discharged conidia by individual fungi is displayed in Table 1. Looking at the numbers, there is a decrease in total conidia count towards higher temperatures, even when discarding samples 7 and 13.

cold	medium	warm
1575	2003	1742
2019	902*	1764
1921	1510	787*
2019	1991	1470
2323	1720	1769

Table 1: Sums of discharged conidia. * indicate thats these fungi collapsed during the experiment.

However, a one-way anova test gave a p-value of 0.075 (excluding the collapesed fungi), and pairwise Wilcoxon tests and a Kruskal-Wallis test gave even larger p-values. So while it is evident that temperature has an effect on conidia discharge as a function of time, we are not able to detect a significant effect of temperature on the total amount of conidia discharged within the first 120 h.

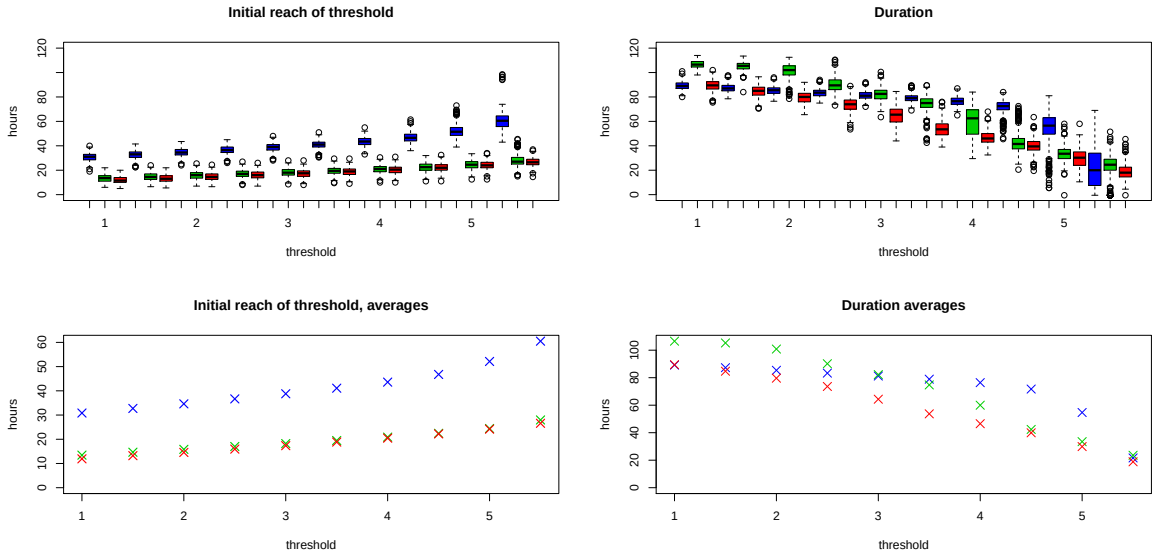


Figure 6: Left: First time conidial discharge intensity reaches given threshold, for different thresholds. Some trajectories didn't reach given thresholds and have been omitted from the corresponding boxplots. Right: Duration that conidial discharge intensity is above given threshold. Blue: 12.0 °C, Green: 18.5 °C, Red: 25 °C.

3.1 Inference for population means

Following the approach outlined in the supplementary material, we estimated the information matrices for the spline coefficients, $I_{cold}, I_{medium}, I_{warm}$. The information matrices themselves are of little interest, but following Berstein-von Mises theorem the inverses evaluated at time points can be used as quantifiers for the uncertainty and standard errors, see Figure 7. We have much more uncertainty for small values of θ. This is as expected; small values of θ corresponds to few conidia counts and thus only little data to estimate from. The pointwise standard errors for θ in regions with large counts are around 0.20-0.25 or 20-25% when exp-transformed.

3.2 Peak location and decrease

Using the standard error estimates from the previous section, we made inference on the location and decrease of peaks. This was done by simulating from the approximate distributions of the estimators. 1000 simulations were used, results are in Table 2. As we expected, θ_{cold} peaked late, around 70h after start, while the fungi stored at higher temperatures peaked much earlier. We observed a large and skewed 95% confidence interval for peak location of θ_{medium}, even containing the similar confidence interval for θ_{warm}. Regarding the second element, peak decrease, we saw a roughly linear relationship between temperature and decrease.

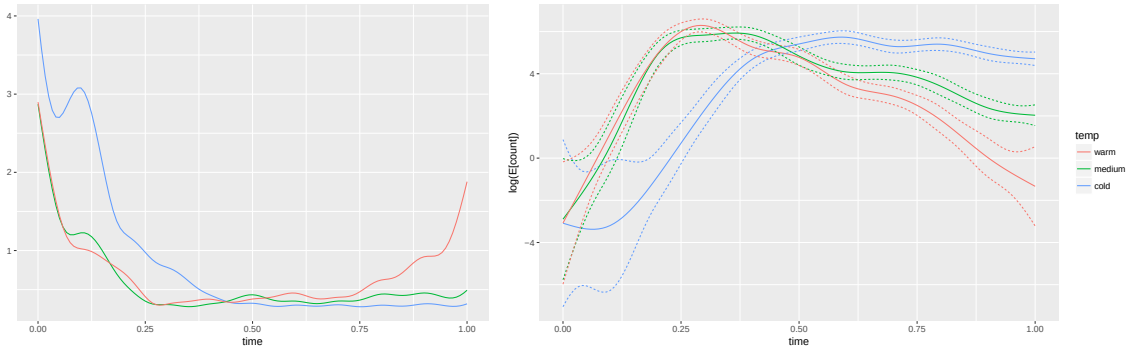


Figure 7: Left: Pointwise evaluations of $1.96 \cdot I^{-1}$, where I is the information matrix. Right: Corresponding pointwise confidence intervals.

The confidence interval for θ_{warm} is broader than the other confidence intervals; this is due to the lack of data for small values of θ, cf. Figure 7. However, all confidence intervals are clearly separated at a 95% level, and we can firmly conclude that lower temperatures leads to a more slowly decreasing peak, with the consequence of increasing the duration of high conidial discharge.

	2.5%	Estimate	97.5%		2.5%	Estimate	97.5%
cold	66.0	70.7	73.7	cold	1.29	2.10	2.99
medium	32.7	43.8	46.6	medium	4.14	5.07	5.98
warm	33.7	35.1	36.1	warm	6.78	8.94	11.42

Table 2: Approximate 95% confidence intervals for peak location (left) and peak decrease (right). Units are hours after start of experiment and %/h, respectively.

Credibility of hypotheses By comparing the approximate distributions of the estimators, we could assess the credibility of of our hypotheses. This was done by pairwise comparison of estimators using $q = P(f(\hat{X}) < f(\hat{Y}))$, where $f(\hat{X})$ and $f(\hat{Y})$ are the posterior distributions of parameter functions, e.g. $f(X) = peak(\theta_{cold})$ and $f(Y) = peak(\theta_{middle})$.

$f(X) = f(Y)$ implies $q = 0.5$, so small or large values of q are evidence against the hypothesis $f(X) = f(Y)$. Results are shown in Table 3. Apart from peak(θ_{middle}) = peak(θ_{warm}), all q-values are very close to one. As a result, our analysis very strongly supports that higher temperatures lead to faster decreasing peaks, and that a low temperature gives a late peak in comparison to middle and high temperatures.

hypothesis	q
$\text{peak}(\theta_{cold}) = \text{peak}(\theta_{middle})$	1.00
$\text{peak}(\theta_{cold}) = \text{peak}(\theta_{warm})$	1.00
$\text{peak}(\theta_{middle}) = \text{peak}(\theta_{warm})$	0.108
$\text{slope}(\theta_{cold}) = \text{slope}(\theta_{middle})$	0.9999
$\text{slope}(\theta_{cold}) = \text{slope}(\theta_{warm})$	1.00
$\text{slope}(\theta_{middle}) = \text{slope}(\theta_{warm})$	0.9993

Table 3: Pairwise comparisons of hypotheses with credibility values.

3.3 Robustness of statistical analysis

Leave-one-out-analysis To assess the uncertainty and robustness of the estimates, a leave-one-out analysis was performed: One observation (or in our case, one curve) is removed from the data set, and the model is fitted to the reduced data set. This is done for all N observations in turn, and the results are compared in the end. These results should preferably not differ by much; this is called *robustness*; lack of robustness is an indication of overfitting, that is too many features or variables are included in the model. Robustness is related to *generalised cross-validation*; see e.g. [30] for a reference.

As our model is highly non-linear and consists of several layers, each with different parameters, it was of interest to study the robustness. As seen in Table 4 we got a fairly large spread on the amplitude covariance parameters. However, this can be explained by the many kinds of variation in data; it is more relevant that the mean curves are very robust, as the population means are main interest of this study. The explanation behind the large spreads observed in the beginning is that large negative values are mapped into almost-zero values.

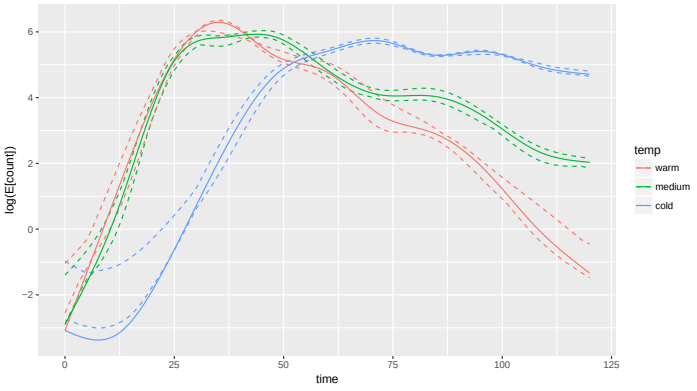


Figure 8: Pointwise estimates, and upper and lower bounds for leave-one-out analysis.

Parameter	nb-dispersion	$\text{range}_{\text{amp}}$	$\text{smoothness}_{\text{amp}}$	$\text{scale}_{\text{amp}}$	$\text{range}_{\text{warp}}$	$\text{scale}_{\text{warp}}$
Lower bound	4.40	0.314	2.30	0.0066	0.072	0.023
Estimate	4.66	0.458	7.21	0.072	0.083	0.026
Upper bound	5.21	0.523	10.0	0.084	0.691	0.034

Table 4: Parameter estimates and leave-one-out results. Note: An upper bound of 10 for the Matern-smoothness was used in the analysis.

4 Discussion

With the applied statistical methods, we were able to characterize the temporal patterns of conidial discharge to a much better degree than previous studies, and we characterized the variation between individual fungi of the same population. With a 95variation is too little for changing the overall shapes, but still large enough to be important for the analysis and to shift the peaks for individual fungi significantly.

Good statistical methods are essential when analysing biological systems with a temporal pattern, and allow researchers to get a better interpretation of data. Advanced statistical methods are not always better than simple ones, but the applied methods should be able to capture all essential variations in data.

Examples of statistical analyses (some using the pavpop model) of other biological systems, where a model of the temporal variation was essential for the data analysis and interpretation of results, include electroforectic spectra of cheese [31], growth of boys [27] and hand movements [28, 27].

In this study we demonstrated the flexibility of the pavpop model by successfully fitting to a complete different kind of data: namely discrete data with a lot of zeros, where a Gaussian approximation would be unreasonable. With this success, there is reason to believe that this framework would work well in applications with other commonly used response models, for example binary response models (logistic regression).

Having several counts per measurement allowed us to look into response models. The Poisson model was invalidated, so we applied a negative binomial model instead. This is also relevant for similar/future studies: a negative binomial distribution gives rise to larger standard errors on estimates than a corresponding Poisson model, which increases the risk of making type I errors.

There were some uncertainties in the estimation, but given the comparatively small amount of data, this is adequate. The robustness analysis can be used to asses which parameters are identifiable. Although some of the variance parameters were not well identified, the dispersion parameter, average temporal deviation and population means were found to be quite robust.

We were not able to detect significant differences in total number of discharged conidia in this study. However, the fungi stored at 12 °C were still discharging many conidia at $t = 120h$,

so there is good reason to believe that there are significant differences when using a longer time span; the authors have data that supports this, too. In a study conducted on mycelial mats of *Pandora neoaphidis* over 168 h this could be observed: At 25 °C the mycelium mats produced less primary conidia compared to mycelium mats incubated at 10, 15 or 20 °C [13]. Aphid cadavers infected by *P. neoaphidis* discharged similar numbers of primary conidia at temperatures between 5 and 25 °C in the first 24 hours [12].

On the other hand, we detected significant differences in peak location and shape: high temperature leads to an early peak but fast decreasing intensity of conidial discharge compared to low temperature. Other authors also found an earlier peak and faster decreasing intensity of conidial discharge at 25 °C compared to lower temperatures in other species of fungi, but the position of the peak and decrease of conidial discharge intensity was not statistically analysed [10, 14].

Our findings agree with those of [14]; lower temperature leads to longer durations of conidial discharge. When the host population is large, the chance of a conidium landing on a host is larger and there is no advantage of prolonging the conidial discharge [14]. The effects of temperature on temporal pattern of conidial discharge are important in practical applications and for the potential of this species as a biocontrol agent. The most important factor is the duration of intense conidial discharge, thus we believe the biocontrol potential to be largest at cold temperatures; the effect is illustrated in Figure 7. To get a better understanding of the environmental tolerance of a fungus regarding conidial discharge, experiments including fluctuating temperature, different relative humidity and light levels need to be conducted. Furthermore, the conidial discharge from insect cadavers can be measured to get a better understanding of the development of epizootics in the field. The presented statistical framework will likely be of great benefit for future data analysis of any experiments in which conidial discharge is measured over time.

References

[1] Sakes A, van der Wiel M, Henselmans PW, van Leeuwen JL, Dodou D, Breedveld P. Shooting mechanisms in nature: a systematic review. PloS one. 2016;11(7):e0158277.

[2] Money NP. Spore production, discharge, and dispersal. In: Waktinson SC, Boddy L, Money NP, editors. The Fungi (Third Edition). Academic Press; 2016. p. 67–97.

[3] Boomsma JJ, Jensen AB, Meyling NV, Eilenberg J. Evolutionary interaction networks of insect pathogenic fungi. Annual Review of Entomology. 2014;59:467–485.

[4] Eilenberg J. The culture of *Entomophthora muscae* (C) Fres. in carrot flies (*Psila rosae* F.) and the effect of temperature on the pathology of the fungus. Entomophaga. 1987;32(4):425–435.

[5] Aoki J. Pattern of Conidial Discharge of an Entomophthora species ("grylli" type)(Entomophthorales: Entomophthoracea) from Infected Cadavers of *Mamestra brassicae L.* (Lepidoptera: Noctuidae). Applied Entomology and Zoology. 1981;16(3):216–224.

[6] Hemmati F, Mccartney HA, Clark SJ, Deadman ML, et al. Conidial discharge in the aphid pathogen *Erynia neoaphidis*. Mycological Research. 2001;105(6):715–722.

[7] Hajek AE, Davis CI, Eastburn CC, Vermeylen FM. Deposition and germination of conidia of the entomopathogen *Entomophaga maimaiga* infecting larvae of gypsy moth, *Lymantria dispar*. Journal of Invertebrate pathology. 2002;79(1):37–43.

[8] Wraight S, Galaini-Wraight S, Carruthers R, Roberts DW. *Zoophthora radicans* (Zygomycetes: Entomophthorales) conidia production from naturally infected *Empoasca kraemeri* and dry-formulated mycelium under laboratory and field conditions. Biological Control. 2003;28(1):60–77.

[9] Hajek AE, Papierok B, Eilenberg J. Methods for study of the Entomophthorales. In: Lacey LA, editor. Manual of Techniques in Invertebrate Pathology (Second Edition). Elsevier; 2012. p. 285–316.

[10] Kalsbeek V, Pell JK, Steenberg T. Sporulation by *Entomophthora schizophorae* (Zygomycetes: Entomophthorales) from housefly cadavers and the persistence of primary conidia at constant temperatures and relative humidities. Journal of Invertebrate Pathology. 2001;77(3):149–157.

[11] Shah PA, Pell JK. Entomopathogenic fungi as biological control agents. Applied Microbiology and Biotechnology. 2003;61(5-6):413–423.

[12] Yu Z, Nordin G, Brown G, Jackson D. Studies on *Pandora neoaphidis* (Entomophthorales: Entomophthoraceae) infectious to the red morph of tobacco aphid (Homoptera: Aphididae). Environmental entomology. 1995;24(4):962–966.

[13] Shah PA, Aebi M, Tuor U. Effects of constant and fluctuating temperatures on sporulation and infection by the aphid-pathogenic fungus *Pandora neoaphidis*. Entomologia experimentalis et applicata. 2002;103(3):257–266.

[14] Li W, Xu WA, Sheng CF, Wang HT, Xuan WJ. Factors affecting sporulation and germination of *Pandora nouryi* (Entomophthorales: Entomophthoraecae), a pathogen of *Myzus persicae* (Homoptera: Aphididae). Biocontrol Science and Technology. 2006;16(6):647–652.

[15] Glare TR, Milner RJ. Ecology of entomopathogenic fungi. In: Handbook of Applied Mycology. Marcel Dekker; 1991.

[16] Hajek AE, Meyling NV. Fungi. In: Hajek AE, Shapiro-Ilan DI, editors. Ecology of Invertebrate Diseases. John Wiley & Sons; 2018.

[17] Furlong MJ, Pell JK. The influence of environmental factors on the persistence of *Zoophthora radicans* Conidia. Journal of Invertebrate Pathology. 1997;69(3):223–233.

[18] Alford DV. Pests of fruit crops: a colour handbook. CRC press; 2016.

[19] Seemüller E, Schneider B. '*Candidatus* Phytoplasma mali', '*Candidatus* Phytoplasma pyri'and '*Candidatus* Phytoplasma prunorum', the causal agents of apple proliferation, pear decline and European stone fruit yellows, respectively. International Journal of Systematic and Evolutionary Microbiology. 2004;54(4):1217–1226.

[20] Strauss E. Phytoplasma Research Begins to Bloom. Science. 2009;325(5939):388–390. doi:10.1126/science.325_388.

[21] Naef A, Kuske S, Holliger E, Kuster T, et al. Pflanzenschutzempfehlungen für den Erwerbsobstbau 2018/2019. vol. 210. Agroscope; 2018.

[22] Jarausch B, Jarausch W. Psyllid vectors and their control. Phytoplasmas: Genomes, Plant Hosts and Vectors. 2010; p. 250–271.

[23] Jensen AH. A new insect pathogenic fungus from Entomophthorales with potential for psyllid control; 2017.

[24] Hajek AE, Shapiro-Ilan DI. Ecology of Invertebrate Diseases. John Wiley & Sons; 2018.

[25] Campos-Herrera R, Lacey LA, Hajek AE. Methods for studying the ecology of invertebrate diseases and pathogens. Ecology of Invertebrate Diseases. 2017; p. 19–47.

[26] Raket LL, Sommer S, Markussen B. A nonlinear mixed-effects model for simultaneous smoothing and registration of functional data. Pattern Recognition Letters. 2014;38:1–7.

[27] Olsen NL, Markussen B, Raket LL. Simultaneous inference for misaligned multivariate functional data. Journal of the Royal Statistical Society: Series C (Applied Statistics);67(5):1147–1176. doi:10.1111/rssc.12276.

[28] Raket LL, Grimme B, Schöner G, Igel C, Markussen B. Separating timing, movement conditions and individual differences in the analysis of human movement. PLoS Computational Biology. 2016;12(9):e1005092.

[29] Yeo H, Pell J, Walter M, Boyd-Wilson K, Snelling C, Suckling D. Susceptibility of diamondback moth (*Plutella xylostella* (L.)) larvae to the entomopathogenic fungus, *Zoophthora radicans* (Brefeld) Batko. In: Proceedings of the New Zealand Plant Protection Conference. New Zealand Plant Protection Society; 1998; 2001. p. 47–50.

[30] Friedman J, Hastie T, Tibshirani R. The Elements of Statistical Learning. 2nd ed. Springer; 2009.

[31] Rønn BB. Nonparametric maximum likelihood estimation for shifted curves. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2001;63(2):243–259.

Supplementary material

Statistical estimation

Direct inference in the statistical model (1) and (2) is not feasible due to the large number of latent variables. Furthermore, unlike the setup in [27], the response is not Gaussian, which further complicate estimation. One solution would be to use MCMC methods, which are generally applicable. However, we propose to use a double Laplace approximation for doing approximate maximum likelihood estimation.

This Laplace approximation actually consists of a linearisation around the warp variables w_n followed by a Laplace approximation on the discretised mean curves $\boldsymbol{u}$; $\boldsymbol{u}_n = \{u_n(t_{nk})\}_{k=1}^{m_n}$ for $n = 1 \ldots, N$. When these approximations are done at the maximum posterior values of $(w_n, \boldsymbol{u}_n)$, this is equivalent to a Laplace approximation jointly on $(w_n, \boldsymbol{u}_n)$.

The main difference from the estimation procedure of [27] is the non-trivial addition second layer of latent variables $\boldsymbol{u}$.

Posterior likelihood To perform Laplace approximation, we need the mode of the joint density of observations and latent variables; this can be found by maximising the posterior likelihood for the latent variables.

The joint posterior negative log-likelihood for sample n is proportional to

$$L = \big[\sum_{k=1}^{m_n} A(u_{nk}) - u_{nk}y_{nk}\big] + \frac{1}{2}(\boldsymbol{\gamma}_{w_n} - \boldsymbol{u}_n)^\top S_n^{-1}(\boldsymbol{\gamma}_{w_n} - \boldsymbol{u}_n) + \frac{1}{2}w_n^\top C^{-1} w_n \tag{3}$$

where $\boldsymbol{\gamma}_{w_n}$ denote the vector $\{\theta_{f(n)}(v(t_{nk}, w_n))\}_{k=1}^{m_n}$. Spline coefficients for the fixed effects are indirectly present in the posterior likelihood through $\boldsymbol{\gamma}_{w_n}$; more details follow below. It should be noted that under relatively mild assumptions, minimizing (3) for a fixed w is a convex optimization problem.

Likelihood approximation To approximate the likelihood, we firstly linearise around w^0 to approximate $p(u)$ with a Gaussian distribution and secondly we make a Laplace approximation of the joint linearised likelihood.

The linearization around w^0 to approximate the likelihood for density the mean curves, $p(u_n)$, is described in detail in [26, 27]. The result of doing this is a Gaussian approximation of the latent u, ie. $\boldsymbol{u}_n \overset{D}{\approx} \tilde{\boldsymbol{u}}_n$ where $\tilde{\boldsymbol{u}}_n \sim N(r_n, V_n)$. r_n and V_n are obtained from the Taylor approximation of u around the posterior maximum w_n^0; for details we refer to [26, 27].

In general, the Laplace approximation of an integral on the form $\int_{\mathbb{R}^d} e^{f(x)}\, \mathrm{d}x$ around the mode x_0 of f is given by

$$(2\pi)^{d/2}|-f''(x_0)|^{-1/2}e^{f(x_0)} \tag{4}$$

where $|-f''(x_0)|$ is the determinant of the Hessian of $-f$, evaluated in x_0. This approximation is exact if f is a second-order polynomial, and generally the approximation is directly related to the second-order Taylor approximation of f at x_0.

Up to some constants, which do not depend on the parameters, the likelihood for a single curve in the linearized model is given by the following integral, which we want to approximate:

$$L_n^{\text{lin}} \propto \int_{\mathbb{R}^{m_n}} |V_n|^{-1/2} \exp\left(-\tfrac{1}{2}(\boldsymbol{u}_n - r_n)^\top V_n^{-1}(\boldsymbol{u}_n - r_n) + \sum_{k=1}^{m_n} y_{nk}\boldsymbol{u}_{nk}^0 - A(u_{nk}) \right) \mathrm{d}\boldsymbol{u}_\mathrm{n} \quad (5)$$

Assuming $\boldsymbol{u}_n^0$ to be the maximum of the posterior likelihood (3), one can show that the negative logarithm of the Laplace approximation around $(\boldsymbol{u}_n^0, w_n^0)$ is given by:

$$1/2 \log|\tilde{\Sigma}_n| + \sum_{k=1}^{m_n}(A(u_{nk}^0) - y_{nk}\boldsymbol{u}_{nk}^0) + p(\boldsymbol{u}_n^0)$$

where $\tilde{\Sigma}_n = V_n^{-1} + 2diag(A''(\boldsymbol{u}_n^0))$ and $p(\cdot)$ is the negative log-density for a $N(r_n^0, V_n)$-distribution. By assumption, $A''(u_{nk}^0) > 0$, so $|\tilde{\Sigma}_n| > |V_n|^{-1}$.

The total log-likelihood for all observations is then approximated by

$$\sum_{n=1}^{N}\left[\log|\tilde{\Sigma}_n| + \log|V_n| + (\boldsymbol{u}_n^0 - r_n)^\top V_n^{-1}(\boldsymbol{u}_n^0 - r_n) + 2\sum_{k=1}^{m_n}(A(u_{nk}^0) - y_{nk}\boldsymbol{u}_{nk}^0)\right] \quad (6)$$

Inference We propose to use alternating steps of (a) estimating spline coefficients for the fixed effects and predicting the most likely warps and mean curves by minimizing the posterior log-likelihood (3) and (b) estimating variance parameters from minimizing the approximated log-likelihood (6).

Fixed effects and phase variation

Fixed effects are modeled using a spline basis that is assumed to be continuously differentiable, e.g. a Fourier basis or B-spline bases. A typical choice for non-periodic data would be B-splines; we have used natural cubic splines in the data application. Fixed effects are estimated using the posterior likelihood (3). For a fixed value of w_n, $\boldsymbol{\gamma}_{w_n}$ is a linear function of spline coefficients, and thus the optimal value can be found using standard linear algebra tools.

Phase variation is modelled by random warping functions $v_n = v(\cdot, w_n) : [0,1] \to D$, parametrized by independent zero-mean Gaussian variables $w_n \in \mathbb{R}^{m_w}$. $v : [0,1] \times R^{m_w} \to D$ is a suitable spline interpolation function, such that $v(\cdot, 0)$ is the identity on $[0,1]$.

The latent trajectories v_n are modelled as deviations from the identity function at pre-specified time points $(t_k)_{k=1}^{m_w}$, subject to a Hyman filtered, cubic spline interpolation for insuring monotonicity, $v_n(t_k) \approx t_k + w_{nk}$. A more detailed discussion of modelling phase variation using increasing spline functions can be found in [27].

Uncertainty for fixed effects As our model is highly non-linear, we cannot expect to find closed-form expressions for the uncertainty in parameters. Furthermore, the latent variables complicates assessment of uncertainty as these are uncertain themselves.

A standard quantifier for assessing uncertainty in statistical models is the *information matrix*, which under regularity conditions can be approximated by the second-order derivative of the log-likelihood at MLE. However, directly using (6) would underestimate the information, as (6) depend on the optimal value of the posterior likelihood (3), which itself is a function of parameters.

Let c_j denote the spline coefficients which determine the population mean θ_j for treatment group j. c_j is determined from the posterior likelihood $L = L(c, u, w)$, formula 3. As u and w are latent, it would be wrong to use the second derivative of L for the information matrix; instead we use the second derivative of $f(c) = L(c, u(c), w(c))$, where u and w are viewed as functions that map c into the max-posteriors of u and w given c.

This will more correctly ensure that the uncertainty of u and w is taken into account when estimating the information matrix. Furthermore, positive definiteness of L'' will imply positive definiteness of f''.

Response models

In the application presented in this paper we assume that $(y|u)$ follows a negative binomial distribution. There are various choices of response models, a list of important ones are stated below. Note that not all exponential families fits naturally with our methodology; $y|u$ must be well-defined for all $u \in \mathbb{R}$.

Binary response: For binary responses, the sample space is $\mathcal{Y} = \{0, 1\}$. If we define $p := P(Y = 1|\eta)$, and set $A(\eta) = \log(1 + e^\eta)$, we get that $\eta = \log(p) - \log(1 - p)$, the canonical link function for regression models with binomial response.

Poisson model: $Y \in \mathbb{N}_0$ where $A(\eta) = e^\eta$, the canonical link function. The conditional mean satisfies $E[Y|\eta] = e^\eta$.

Negative binomial model: Negative binomial distributions are often viewed as overdispersed versions of Poisson models. Let the rate parameter $r > 0$ be given such that $V[Y|\eta] = E[Y|\eta] + E[Y|\eta]^2/r$; the limit $r \to \infty$ corresponds to the Poisson model.

We get $A(\eta, y) = (r + y)\log(1 + \frac{e^\eta}{r})$ and $A''(\eta; y) = (y + r)\frac{re^\eta}{(r+e^\eta)^2}$. Unlike the Poisson and binomial models, the link function A depends on y, but it is easily seen that $A(\eta, y)$ approximates e^η in the limit $r \to \infty$.

Normal distribution with known variance σ^2: For normal distributions we have $Y \in \mathbb{R}$, and by setting $\tilde{Y} = Y/\sigma^2$. Then $A(\eta) = \eta^2/2\sigma^2$, $E[\tilde{Y}|\eta] = \eta/\sigma^2$, and $E[Y|\eta] = \eta$. This is the most basic response model, and the used in [26]. [26] use a different formulation and also treats σ^2 as an unknown parameter. The Laplace approximation becomes exact when using normal distributions, simplifying estimation to become the approach used in [27].

Matern covariance function

The Matérn covariance function is commonly used in functional data analysis and spatial statistics. It is given by

$$f_{\sigma,\alpha,\kappa}(s,t) = \sigma^2 \frac{2^{1-\alpha}}{\Gamma(\alpha)} \left(\frac{\alpha|s-t|}{\kappa} \right)^{\alpha} K_{\alpha} \left(\frac{\alpha|s-t|}{\kappa} \right), \quad s,t \in \mathbb{R} \tag{7}$$

where K_α is the modified Bessel function of the second kind. Here σ is the scale parameter, α is the smoothness parameter and κ is the range parameter.

3 Conclusion and outlook

In this thesis some new dedicated methods for functional data have been presented. The applicability of the methods was demonstrated on various data sets. Five very different data sets both in terms of scientific areas and sizes were analysed, and we gained insight into the statistical variation within these data.

We made a comparison with other methods on the hand movement data; one notable conclusion is that warping functions or phase distance is an important feature and not just nuisance. The methods which incorporate phase distance had markedly better classification results compared to other methods.

Another subject was that of multivariate functional data, where we in Paper I introduced the *dynamical correlation structure* for modelling cross-correlation and applied it on two data sets. The MCA model also allows for varying cross-correlation, but with a different, high-parametric approach. It would be of interest for future studies to compare this with the dynamical correlation structure.

The methodology in Paper IV is strongly related to that in paper I, but looking into a different direction, namely discrete data consisting of conidia counts. We successfully applied the proposed model to a data with a negative binomial response, and there is reason to believe that this study can be helpful in mycology research. From a statistical point of view, we demonstrated the methodology of Raket et al. (2014) in a generalized linear context.

The scope of paper III and the addendum to it was quite different, namely local inference. Here we went into a non-parametric setting, where the estimation of parameters or alignment of data was of little interest. Although the functional BH and the spherical IWT procedures rely on parametric tests, the dependence structure is of less relevance.

Our results showed that Earth temperatures have increased significantly in many regions. This is no surprise, but the important thing is that we made functional multiplicity corrections to account for the multiple comparison problem. We chose to focus on increasing temperatures; one could also focus on decreasing temperatures. Despite of global warming, there are regions where temperatures have significantly decreased, although this area is much smaller, and perhaps less important, than regions with increasing temperatures.

There are many possible variations on the IWT and fBH procedures, which would be interesting to explore in future research. An extensive comparison of local inference methods is also an interesting perspective.

There are several future perspectives as presented in the discussion parts of the papers, although they point in various directions.

Many perspectives lie in modelling of warping functions as discussed in Section 1.2 and paper I. One perspective is warping functions with continuous space and more general regularity criteria. Other perspectives are correlations between amplitude and phase variation, and the combination of warps with differential equations, as discussed in Section 1.2.3. Although not discussed in this thesis and not relevant in Paper III, warping may be an issue in local inference. The interesting issue would not so much be if there is an effect of warping but *how much* effect there is? – too little alignment would not allow to detect significant differences, whereas too much alignment could amplify the registered signals too much; where variation is not due to warping. To sum up, I believe there is much open research in alignment of functional data, in particular its interplay with other areas of functional data analysis.

Although functional data analysis is a well-established field, many open research problems still remain. In this thesis we have looked into new methods for functional data with many interesting perspectives.