Analysis of Functional Data with Focus on Multinomial Regression and Multilevel Data

PhD thesis by

SEYED NOUROLLAH MOUSAVI

Department of Mathematical Sciences
University of Copenhagen
Denmark

PhD School of Science - Faculty of Science - University of Copenhagen

Seyed Nourollah Mousavi
Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
DK-2100 København Ø
Denmark

nourollah@math.ku.dk
n-mousavi@araku.ac.ir

PhD thesis submitted to the PhD School of Science, Faculty of Science, University of Copenhagen, Denmark, in October 2015.

Academic advisor: Helle Sørensen
University of Copenhagen, Denmark

Assessment Committee: Bo Markussen (chair)
University of Copenhagen, Denmark

Simone Vantini
University of Politecnico di Milano, Italy

Lina Schelin
UmeåUniversity, Sweden

ISBN 978-87-7078-940-0

Preface

This dissertation is submitted in partial fulfillment of the requirements for the Ph.D. degree at the Faculty of Science, University of Copenhagen, Denmark. The work was carried out at the Department of Mathematical Sciences, University of Copenhagen, from July 2011 to October 2015.

The thesis is broadly concerned with statistics, with contributions to functional data analysis. Motivating applications are primarily to be found in bioscience fields, although the results can be used for all functional data from other fields.

My interest in Functional Data Analysis was initiated when I started my PhD program at University of Copenhagen, by my supervisor Helle Sørensen. First and foremost, I would like to give the special thanks to Helle, for her enthusiasm and unreserved support on both my academic development and personal life over the last four years. Many of the ideas in the thesis originally came from my discussions with her. Also a great thanks to Ana-Maria Staicu for introducing me to the world of multilevel functional data, for a really nice time during my five-month stay abroad at North Carolina State University and for her patience during numerous Skype conversation.

To everyone at the Department of Mathematical Sciences, thank you for creating a scientifically exciting environment. Special thanks are due to Anders Tolver and Martha Muller for collaboration and discussion.

Special recognition goes out to my family for their support, encouragement and patience during my study. This has canceled thousands of kilometers of distance. My last and deepest gratitude goes to my beloved wife Sima, who continuously provided me with care and support, gave me new perspectives on life and my lovely daughter Saba, who made the nine last months of our life tough but very exciting, sweet, and promising.

Seyed Nourollah Mousavi
Copenhagen, October 2015

Summary

Functional data analysis (FDA) is a fast growing area in statistical research with increasingly diverse range of application from economics, medicine, agriculture, chemometrics, etc. Functional regression is an area of FDA which has received the most attention both in aspects of application and methodological development. Our main concerns are two types of functional regression, namely, functional predictor regression (scalar-on-function) and function-on-function regression. In particular, in the first paper included in this thesis, we introduce multinomial functional regression model to analyze functional data with a categorical response (more than two classes) and a functional predictor. To this end, a combination of discrete wavelet transform and LASSO penalization is considered. This model is applied to two datasets, one regarding lameness detection for horse and another regarding speech recognition.

In the second paper, we consider functional logistic regression via wavelet and LASSO which is a specific case of multinomial functional regression with two classes for the response and compare the efficiency (from classification point of view) of this model with two other models, namely, functional penalized regression and function regression using functional principle components. The comparison is based on simulation study and data application.

In the third paper, we study a constrained version of function-on-function regression, in which both response and predictor are defined at same domain and the prediction of the response at time t only depends on th concurrently observed predictor. We introduce a version of this model for multilevel functional data of the type subject-unit, with the unit-level data being functional observations.

Finally, in the fourth paper we show how registration can be applied to functional data by considering a simple biomechanical constraint and then this approach is applied to a functional dataset from a juggling experiment.

Dansk resumé

Funktionel dataanalyse (FDA) er et hurtigt voksende område af statistik med flere og flere anvendelsesområder indenfor økonomi, medicin, biologi, kemometri mm. Funktionel regression er et et område indenfor FDA som har fået en del opmærsomhed, både hvad angår metodeudvikling og anvendelser. Vi vil interessere os for to slags regression. Den forklarende variabel er i begge tilfælde funktioner, mens responsen enten er diskret eller funktionel. I afhandlingens første artikel introducuerer vi en regressionsmodel for kategorisk respons (med tre eller flere mulige værdier) og en funktionel prædiktor. Vi kombinerer wavelets og LASSO-regularisering til at estimere i modellen. Metoden anvendes på to datasæt; et der vedrører detektion af halthed hos heste og et der vedrører talegenkendelse.

I den anden artikel betragter vi funktionel logistisk regression, dvs. den situation hvor responsvariablen er binær. Vi bruger igen estimationsmetoden med wavelets og LASSO og sammenligner denne metode med to eksisterende metoder med henblik på prædiktionsevne.

I den tredje artikel betragter vi regression hvor både prædiktor og respons er funktioner. Vi antager at alle funktioner er defineret på samme domæne og at fordelingen af responsen til tid t kun afhænger af prædiktorfunktionen gennem værdien på same tidspunkt. Dette kaldes "the concurrent model". Vi introducerer en version af modellen for hierarkiske funktionelle data hvor der er flere funktionelle observationer per individ.

Endelige betragter vi i den fjerde atikel et registreringsproblem med naturlige biomekaniske restriktioner og forslår en metode der tager højde for dette. Metoden anvendes på et datasæt vedrørende jonglering.

Contents

1 Introduction

It is in human nature to be curious about the natural phenomena and try to understand them. To describe a phenomenon, a mathematical model is needed. To this end, some observations on a phenomenon must be quantified. As most measurements are contaminated with noise or measurement error, in order to take into account some amount of uncertainties, we need to investigate on a statistical model.

1.1 Moving from classic data to functional data

Most statistical analyses involve one or more observation taken from a number of individuals in order to make an inference about the general population. Some times we wish to explain the observed quantity y as a response or dependent variable by a number of other quantities, $x_1, x_2, \cdots, x_p$ as covariates or independent variables. Perhaps the simplest model which is used to explain this relationship is the linear model:

$$y = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \epsilon \tag{1.1}$$

where $\beta_0, \beta_1, \cdots, \beta_p$ are coefficients and ϵ is an error term that accounts for uncertainties. We refer 1.1 as linear model. The goal of regression model is to predict y from x, assessment of the effect of or relationship between explanatory variables on the response, and a general description of the data structure. Several approaches have been developed to estimate the parameters and also statistical inference about the parameters and the error term. The simplest approach is based on minimizing the sum of residual square and is commonly known as ordinary least squares (OLS) method. Also the methods are extended to remedy the multicollinearity between the data and overcome with high dimensionality issue which is common nowadays for classic data for instance, partial least squares, penalization methods among others.

In an increasing number of fields because of advancements in technology and computation, these observations are curves (functions) or images, i.e. an observed

intensity is available at each point on a line segment, a part of a plane, or a volume. Domain for these data is usually time, but it can be anything: distance, space, ...

To get the feel of functional data, Figure 1.1 displays three different functional datasets from different fields. A 3D-accelerometer is attached to the back of the horse and acceleration is measured in three directions while the horse is trotting (Halling Thomsen *et al.* , 2010). The top left panel shows a sample of functions from lameness of horse where the x-axis and y-axis represent time and vertical acceleration, respectively The top right panel displays a sample of 10 functional observations related to colon carcinogenesis dataset for the rats with the fish diet who received the butyrate pellets at the indicated time (24 hours) post-AOM injection (Sgambato *et al.* , 2000). The concentration of p27, a cell cycle inhibitor, is measured for each cell within crypt. In this dataset the measurement of p27 is considered as a function of location of the cell within the crypt. The bottom panel represents a dataset in speech recognition which are log-periodograms corresponding to recording continuous speech of 50 male speakers and are available in the `ElemStatLearn` package (Halvorsen, 2012). The dataset consists of five phonemes: ”sh” as in ”she”, ”dcl” as in ”dark:, ”iy” as the vowel in ”she:, ”aa” as the vowel in ”dark”, and ”ao” as the first vowel in ”water”. Here log-periodogram a is function of frequency. We will use these three datasets to perform the proposed approaches in the thesis in data application sections.

The measurements in these datasets are observed only at discrete time points, nonetheless the measurements could, in theory, be measured at any time points, distance and frequency during the period of the study. Therefore, it would be natural to consider these data as function which is defined in continuous argument. Several functional data set have been studied in Ramsay & Silverman (2002). Note, however, it should not be misleading that functional data consider all very high dimensional dataset, for instance, DNA microarray usually must be treated as discrete data.

As it can be seen in three sub-plots of Figure 1.1, functional data is often complicated, complex with a large number of related quantities which can not easily be described by mathematical formula. In addition, the variation between replications might be hard to explain. The idea to make an easier to think about the data is to view each replication as a single observation. The question arise here is that which dataset would be treat as functional data. Generally speaking, one could say that there are some necessities for functional data: first, the data must believably derive from an underlying smooth process. Second, there are enough data to extract the essential feature of the underlying process. Third, there are some repetitions in order to study the interest variations. Finally, in functional data there is no need equally-spaced or perfect measurement.

Functional Data Analysis (FDA) is a field of statistics and probability which deal with these kinds of data, coined by Ramsay & Dalzell (1991). FDA is a general

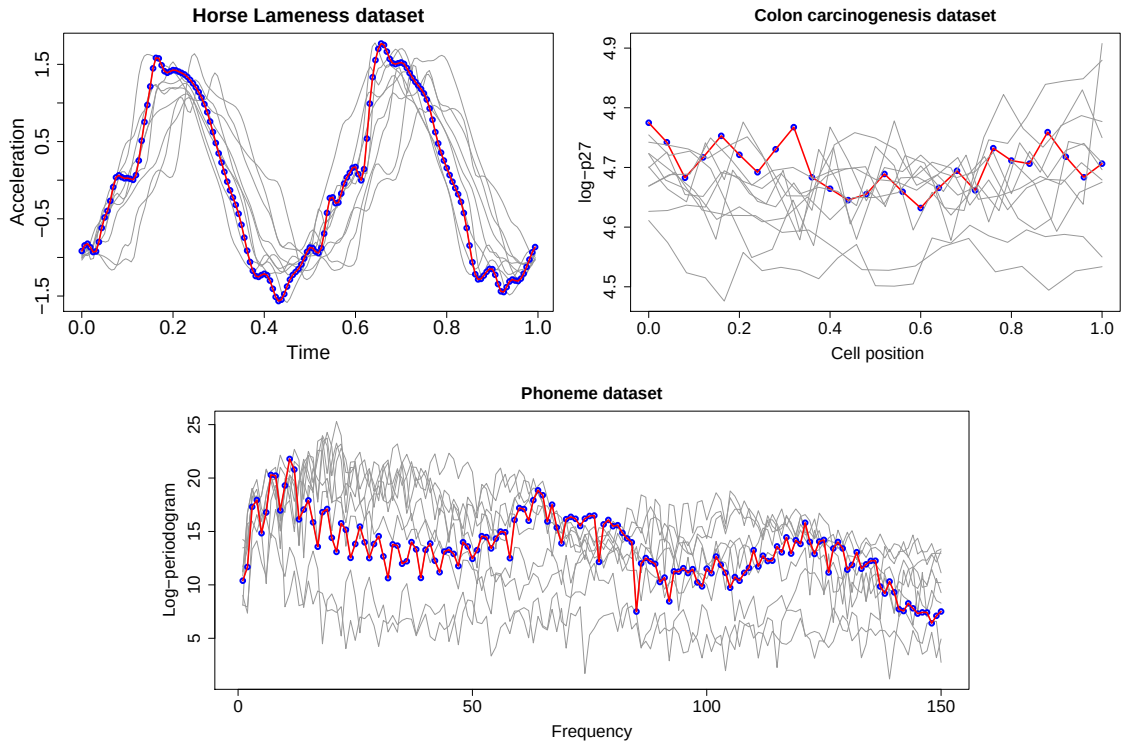

Figure 1.1: Examples of functional data from different disciplines. A random sample of 10 random function is shown for lameness of horse dataset (the top left panel), colon carcinogenesis dataset (the top right panel), and phoneme dataset (the bottom panel). Highlighting one example of the curves for each dataset (the red solid) along the observed data of the functions (the blue circle points).

way of thinking to seek for the basic unit of information in the entire observed function rather than a string of numbers. In other words, in FDA, we consider a set of functions in comparison with the classic multivariate statistics that works with a matrix of observations. The aims of FDA are more or less the same as for other branch of statistics which can be listed below (among others):

- to represent and transform the data in an appropriate way for further analysis.
- to display the data with aim of highlighting various characteristics.
- to investigate the main sources of variation and pattern among the data.
- to explain variations in the response variable by hiring the information of the covariate variables.

Most statistical analyses are based one or more observations in a sample, with the aim of making inferences about the general population from which the sample is drawn. To this end, we combine information either *across* the sampled units or *within* sampled units. One unique characteristics of FDA is the need to combine information both across and within functions, which Ramsay and Silverman called *replication* and *regularization*. In FDA we consider each function as the sampled unit, replication involves combing information across functions in order to make inferences about the general population from which the sample is drawn. Regularization includes borrowing strength across observations within a function exploiting the expected underlying structural relationships within a function in order to improve efficiency and interpretability.

Functional regression is an area of FDA which is maturing both in application and methodological development. Functional regression is an association between the response and predictor in functional data and is split into three types: (1) *functional predictor regression* (scalar-on-function), (2) *functional response regression* (function-on-scalar) and (3) *function-on-function regression*. Two constraint versions of the latter model have been studied in the literature: (i) *concurrent model* in which the outcome and the predictor are assumed to be defined on the same domain and in addition the prediction of the response at t only depends on the observed covariate at time t. (ii) *historical functional linear model* in which the response at current time t relates to the covariate function observed on time window with length Δ prior to t. In this thesis our main attention is on functional predictor regression in situations when the response is categorical (more than two classes) and the concurrent model. The concurrent model is developed for multilevel functional data of the type subject-unit with the unit-level data being functional.

1.2 Objective of the thesis

The main goal of this thesis is to develop association models for functional data. In particular, we aim at

- Developing a classification method based on functional predictor regression.
- Accommodating the concurrent model to multilevel functional data of the type subject-unit with the unit level data being functional.
- Comparing different strategies for classification, both in case of two and more than two groups.
- Applying the methods to relevant functional data and thereby contribute to research in other fields.

- Registration subject to a biomechanical constraint with application to a functional dataset from a juggling experiment.

1.3 Thesis structure

In Chapter 2 the necessary definitions for FDA are given, and some basic tools for pre-processing of functional data are described . In particular, the registration procedure is represented and applied to the dataset of horse lameness. In addition, the necessary knowledge for bootstrap for classic data is given, we show how bootstrap is applied in functional data setting. Chapter 3 presents an summary on Multilevel functional data as well as correlated multilevel functional data is explained. In Chapter 4 some potential and interesting perspectives are discussed, after which the bibliography with the references used in chapters 1-4 can be found. Finally, the papers containing our contributions are collected, each one equipped with its own bibliography.

2

Tools for functional data

In this chapter some necessary definitions and basic tools for functional data including basis expansion, registration, bootstrap are presented. Registration is applied to lameness of horse dataset in order to separate phase and amplitude variation.

2.1 Notation and mathematical definition of functional data

After getting a feeling for functional data, it is turn to define functional data from mathematical point of view. Before defining functional data we need to have a precise definition of random functional variables.

Definition 1. *For a probability space*$(\Omega, \mathcal{F}, \mathbb{P})$ *, a random variable* $\mathbf{X} : \Omega \to E$ *is a functional random variable iff every* $X \in E$ *is a function* $X : M \to F$ *for some non-degenerate continuum* M*. Continuum* M *is non-degenerate iff* $|M| > 1$*.*

We refer X as real functional random variable iff every $X \in E$ is a real function, i.e. a function mapping $T \to \mathbb{R}$, where $T = [a, b], a, b \in \mathbb{R}$, and $a < b$.

In practice, most often the problem of functional data deal with a set of all continuous n-differentiable functions on a real domain. A functional dataset is a generated dataset by a functional random variable. So, now we are ready to make a precise definition of functional dataset.

Definition 2. *A set* $\{X_1,\ X_2,\ \cdots,\ X_n\}$ *is a functional dataset iff* $X_i \sim \mathbf{X},\ \forall\ i$*, for some functional random variable* $\mathbf{X}$*, and* $X_i \perp\!\!\!\perp X_j$ *for all* $i \neq j$*. Each element of functional dataset is referred in the literature as functional datum.*

Therefore, the main idea of functional data analysis is to make an inference and prediction on functional random variable X from the given functional dataset $\{X_1,\ X_2,\ \cdots,\ X_n\}$. It is surprising that we do not aim to observe X_i which is an uncountable, infinite dimensional functional random variable. In practice, we only

observe a finite sample of observation. Suppose that the finite observation functional data

$$\left\{\left\{(Y_{1,1},t_{1,1}),\cdots,(Y_{1,m_1},t_{1,m_1})\right\},\cdots,\left\{(Y_{n,1},t_{n,1}),\cdots,(Y_{n,m_n},t_{n,m_n})\right\}\right\}$$

is given where $t_{i,j} \in T,\ i = 1,2,\cdots,n\ ,\ j = 1,2,\cdots,m_i$. One could interpret that $\{Y_{i1},Y_{i2},\cdots,Y_{im_i},i = 1,2,\cdots,n\}$ are observation of a single variable taken repeatedly on the ith subject of a size n sample at m_i time points. Suppose that $\{X_i(t),t \in T\}$ is the sample signal for the ith subject, then $Y_{ik} = X_i(t_{ik}),k = 1,2,\cdots,m_i$ when the measurements are observed without noise; otherwise $Y_{ik} = X_i(t_{ik}) + \epsilon_{ik}$ where ϵ_{ik} is a measurement error associated with recording Y_{ik}.

The first step is to construct functional data (an infinite object) from the finite observed sample. There are two common approaches to construct functional dataset in context of functional data: basis functions and kernel smoothing. In the following, we will give a short summary of using basis functions. As in many other analyses in statistics and modeling in mathematics, we need to have some assumptions that can be satisfied by the functional random variable. In functional data usually but not always, the standard assumption is that the functional random variable and its first k derivatives are continuous. Here k, the order of the derivative depends on the problem at hand. In other words, we assume that the underlying process is smooth and with this assumption we induce that the adjacent observations should be linked together.

2.2 Basis representation

Motivated from multiple linear regression model which provides an expansion of the response as a linear combination of the associated covariates or the fact that many functions can be approximated by a linear combination of a set of appropriately chosen basis function, we can use basis functions to provide an accurate approximation of the functional datum.

Definition 3. *Suppose $L^2 = L^2(T)$ the space of all squared integrable function defined on T. The inner product defined on L^2 is $< f,g >= \int_T f(x)g(x)\,dx$. A system of basis functions $\{\phi_k(x)\}_{k\geq 1}$ is called orthonormal if $\|\phi_k\|^2 = \int_T |\phi_k(t)|^2\,dt = 1$ and $< \phi_k,\phi_{k'} >= \delta_{kk'}$, where $\delta_{kk'}$ is the Kronecker delta, i.e. $\delta_{kk'}$ is 1 for $k = k'$ and 0 otherwise.*

Basis functions should be chosen to represent the characteristics of functional datum. For instance, Fourier bases are a good choice for periodic functions. Other

common basis functions used in the literature (not all orthogonal) are truncated power series, B-spline, wavelet, monomial, functional principle component. Once the basis functions are chosen, we express a functional observation X_i as

$$X_i(t) = \sum_{k=1}^{\infty} c_{ik}\phi_k(t) \approx \sum_{k=1}^{K_x} c_{ik}\phi_k(t)$$

where the $\{\phi_k\}_{k=1}^{K_x}$, are the basis functions. This approach has a several advantages: first, instead of storing all the data points, one stores the coefficients of the expansion, namely, the c_{ik}. Second, an initial dimension reduction and third, some smoothing. is done on the data.

Note that choosing the number of basis functions K_x is important and critical for all subsequent computations. Small numbers of basis functions mean little flexibility and larger numbers of basis functions results in flexibility, but may overfit. Generally the value of K_x is chosen so that the plotted functional object resemble original data with some smoothing that eliminates the most obvious noise (Ramsay & Silverman, 2005; Febrero-Bande & Oviedo de la Fuente, 2012).

The natural question that arises at this point is how large we should choose the number of basis functions. On the one hand we need to choose K_x large enough in order to catch the most feature in the data, but on the other hand we would like to estimate the smooth functional datum. To this end, we need to make a trade-off between the lack of data fit and the variability of the curve. A measure of the roughness of the fitted function can be defined. For instance, one way to characterize the roughness of a curve is by the size of its curvature, i.e. $PEN_2(x) = \int_T [D^2(X(t))]^2\, dt$. By considering the roughness penalty on the fitted curve, the objective function is thus

$$PENSSE_\lambda(X) = \sum_{j=1}^{m}(Y_j - X(t_j))^2 + \lambda \int_T \{D^2X(t)\}^2\, dt$$

where the smoothing parameter control the trade off between the lack of data fit, as measured by the the first therm and the variability of the function, as measured by the the second term in the objective function. The smoothing parameter can be obtained using a data-driven methods such as cross validation (CV), generalized cross validation (GCV), etc.

2.3 Registration

In a functional dataset, there are two source of variations: phase variation and amplitude variation. Variation in the magnitude or size of functional data is referred

to as amplitude variation which measures the differences in the y-axis while variation in the time scale is often referred to as phase variation which measures the differences in the x-axis. Most techniques in FDA are designed to handle amplitude variation. Depending upon the particular problem at hand, phase variation might be a nuisance or not of primary interest, i.e.in a number of spectral datasets, and so it should be removed from the data before further analysis. While in some situations, phase variation is the main focus and amplitude variation might be a nuisance or is not of primary interest. Although there are some cases when both variations are important. Separating amplitude variation from phase variation is still a challenging problem in FDA.

The procedure of removing phase variation has been investigated under different names in different disciplines, namely, curve registration, curve alignment, and time warping in statistics, biology, and engineering respectively. Curve registration in functional data is a procedure of transforming the time argument such that the curves are more aligned. To this end, we need to estimate time-warping function such as $h_i(t)$ such that $X_i^*(t) = X_i(h_i(t))$ are more aligned. On one hand warping function reflects the variation on the x-axis and on the other hand produces horizontally aligned curves in order to reduce the variability. In registration literature, the most well known methods are shift, landmark, and continuous registration. Time warping-function for the shift registration is a $h_i(t) = t+\delta_i$, where the shift parameter δ_i align the curves, while in the other methods the time warping function may be possibly non-linear. In some application, we might use shift registration and one of the other registration methods together.

2.3.1 Registration on horse lameness dataset

In this section, registration process has been applied on dataset of horse lameness. This dataset will be used and described in detail in Paper I. In clinical lameness examination of horses, a lameness score is assigned based on visual inspection of the locomotion pattern. Not only is detection of lameness and identification of the lame limb a difficult task even for experienced veterinarians; there is also large variation between multiple evaluation undertaken by the same veterinarian. Therefore, objective measurements would be helpful as supplement to have visual examination.

The data consists of 85 signals of vertical acceleration. Each signal is composed of 8 cycles of a bi-phased signal. We denote the observations

$$\begin{aligned} &\{Y_i\ ,\ X_{ij}(t)\}, \quad i = 1, 2, \cdots, 85\ \ ,\ \ j = 1, 2, \cdots, 8\ ,\ t \in [0, 8] \\ &\qquad\qquad Y_i \in \{NO, LF, LH, RH, RF\},\ X_{ij} : [0, 1] \to \mathbb{R} \end{aligned}$$

where the lameness groups NO, LF, LH, RH, RF are corresponding to normal or healthy condition and lameness on left-fore, left-hind, right-hind, and right-hand leg,

respectively. Top panel of Figure 2.1 displays a signal from a horse with lameness on the right-fore limb after smoothing with a 500 B-spline basis. Based on video tape, the first peak relates to stance phase on the RH/LH diagonal. The zero point crossing after LF/RH diagonal were identified and represented by the vertical solid blue lines in the figure. Considering zero crossing points, seven complete cycles were selected shown in bottom left panel. Using continuous registration, these seven subsignals were aligned and displayed in bottom middle panel. The average of aligned subsignals was shift aligned to minus cosine curve. More detail about these processes can be found in Appendix A in Paper I.

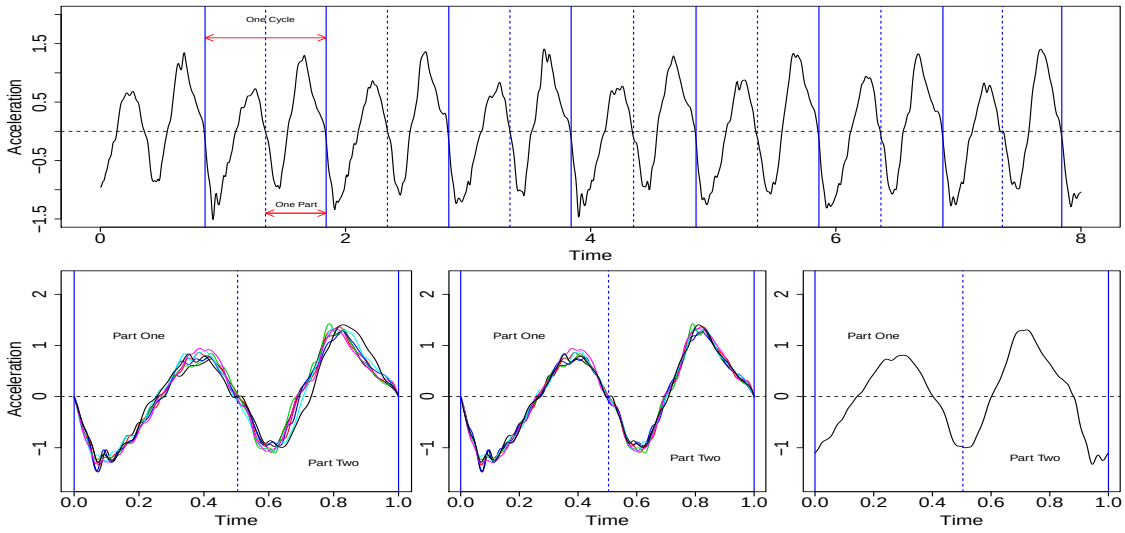

Figure 2.1: Converting the raw data to functional including smoothing and preparing for standard methods in functional data analysis including registration and averaging. Top panel represents a roughly smooth signal for a horse with lameness on the right-fore limb(black) and zero crossing point after phase stance LF/RH diagonal (solid blue line). Bottom left panel shows seven complete cycles before registration and displayed after registration(middle bottom panel). Right bottom panel represents the average of seven aligned subsignal after a shift aligned to a minus cosine function.

The size of the signals differ between part 1 and part 2 which this variation refers to amplitude variation while the location of features differ between part 1 and part 2 which this variation corresponds to phase variation. It is clear that in the horse dataset there are both types of variations in the signals and also both types are related to lameness. Here we are going to use a symmetry score based on phase variation for some classification purpose. In Paper I multinomial functional regression with wavelets and LASSO penalization has been used to analyze and detect lameness where the amplitude variations play the main rule.

It is well-known that the two parts of a cycle for a healthy horse (without lameness) must be symmetric and lameness will disturb this symmetry. Sørensen *et al.* (2012) have defined three different scores to quantify this asymmetry. Here we use a symmetry score, W, based of phase variations which is related to the location of the peaks in two parts in order to use in classification. Figure 2.2 represents the mean of seven subsignals for a horse with lameness on RF limb after registration and minus cosine shift alignment. The green curve displays the mean on $[0, 1/2]$ while the red one is the mean on $[1/2, 1]$. The continuous registration was used to align the part two to part one. The aligned curve for part two is shown in blue one. In order to quantify differences in phase between part one and part two, the symmetry score, W, measures phase displacement of the two part of the signals and was defined as

$$W = \hat{h}(t^*) - t^*, \qquad \text{where} \qquad t^* = \text{argmax}_{t\in[0,1/2]}|\hat{h}(t) - t.| \tag{2.1}$$

Here $h(t)$ is the warping function that transforms acceleration time t. It is clear that $\hat{h}(t) - t$ measures the delay of the part 2 which this value could be positive or negative. So, W represents the largest delay for each signal. For instance, the largest delay for the shown signal of a horse with lameness on RF limb is -0.06. The negative sign is due to early peak on the second part in compare with the first part. This W-score for the healthy horse must be close to or ideally zero because of the symmetry of the two parts.

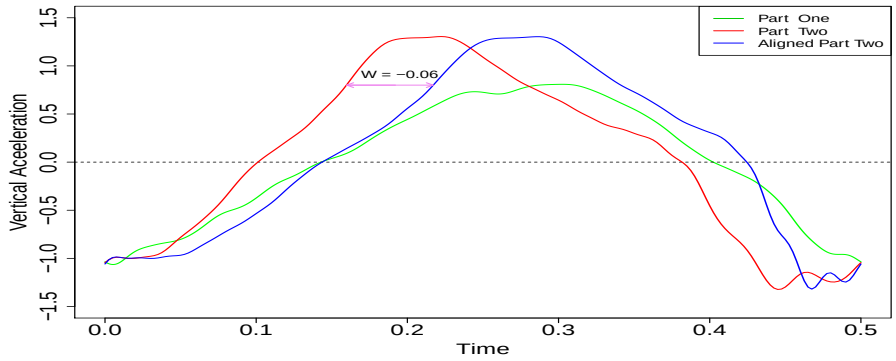

Figure 2.2: Aligned part 2 with part 1 for a horse with lameness on RF limb.

The comparison of the W-scores for different groups is better carried out by inspecting the boxplot of the W-scores which is shown in left panel of figure 2.3. As it was expected, the mean of W-scores for the healthy horse is very close to zero while for the other groups these scores are far from zero. Also from the boxplot can be seen that W-scores for the RF/LH diagonal are almost all negative while for the another diagonal these scores are always positive. This is not surprising as the peak of the signals for the horses with lameness on RF/LH would be on the second part

as we would expect the horse to put less pressure on the ground with this diagonal in compared with the other diagonal and therefore the time for the peak of the part one is smaller than the time for the peak of the part two.

Finally, the differences between aligned parts are considered and the average over signals from each group is shown in right panel of Figure 2.3. The conclusion based on this figure is like the boxplot of W-scores. As we expected the differences for the horses with healthy conditions, is close to zero across the time domain $[0, 1/2]$ while for the other groups these differences are far from zero. The maximum differences have happened on the time of the peak of the part one. The results of these section have been presented as poster in workshop 'Statistics of Time Warping and Phase Variation' at Mathematical Biosciences Institute (MBI) in November 2012 in Columbus, Ohio.

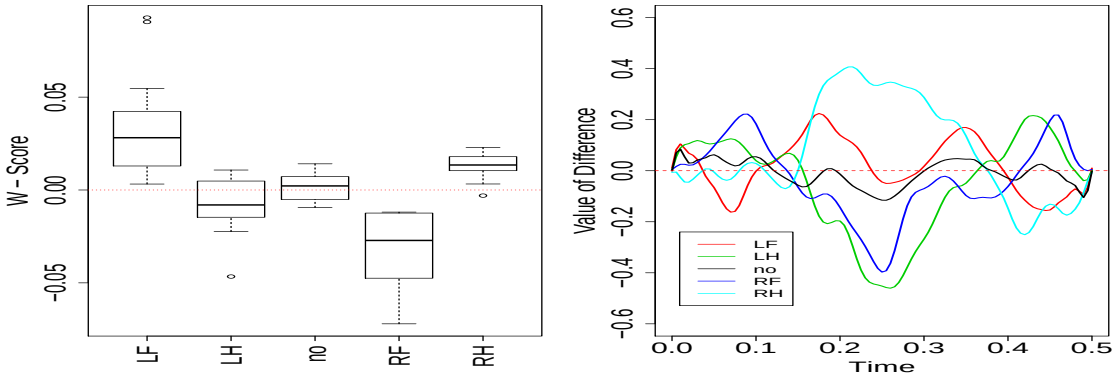

Figure 2.3: Boxplot of W-scores for 85 signals(right panel) and the differences between aligned parts which averaged over signals from each group.

2.3.2 Constrained registration

Registration and phase variation are still challenging and open problems in functional data specially from software implementation side. These issues were the focus of workshop of 'Statistics of Time Warping and Phase Variation' at Mathematical Biosciences Institute (MBI) in November 2012 in Columbus, Ohio. In the workshop four challenging real datasets including juggling dataset were considered in order to apply the existing methods of time-warping registration and do comparisons. In some applications, there are some constraints such as biomechanical constraints in datasets which might be destroyed by applying time-warping registration. It would be ideal if registration process can take into account all knowledge of the data generating system in such a way that the all information will be preserved in the synchronized curves. Considering a simple biomechanical constraint, namely, the

fixed length between the finger tip and the joint of the juggler in juggling dataset is the motivation for Paper IV. The juggling data consists of x, y, z position of the juggler's index finger when he was juggling three balls. More detail on the data can be found in the paper.

2.4 Bootstrap

The term of bootstrapping was introduced by Efron (1979). Bootstrap is a non-parametric approach to statistical inference that substitutes computation for more traditional distributional assumptions and asymptotic results. At first, Efron was motivated to use the bootstrapping for two most important problems in applied statistics, the determination of an estimator for a particular parameter of interest and the determination of confidence interval for the parameter. But because of the bootstrap's generality, it has been developed to much wider classes of problems including regression models, forecasting and time series analysis, survival analysis, clustering analysis, etc. Also it has developed for various disciplines including psychology, geology, econometrics, biology, engineering, chemistry, etc (see Chernick, 2007).

From practical point of view, bootstrap approach need to generate bootstrap sample or re-samples. The general procedure of bootstrapping can be performed as follows. Suppose that we would like to make an inference about a function of θ, $T(\theta)$, where θ is an unknown parameter of the distribution.

- Compute the estimate of $T(\theta)$ using the original data, namely, $T(\hat{\theta})$.
- Generate a bootstrap sample by generating a sample with replacement from the empirical distribution.
- Compute $T(\theta^*)$ the value of $T(\hat{\theta})$ obtained by using the bootstrap sample instead of original data where θ^* is the estimate of θ based on the bootstrap sample.
- Repeat step 2 and 3 K_B times.
- Make an inference on $\{T_b(\theta^*), b = 1, 2, \cdots, K_B\}$.

There is a great attempt to apply the bootstrap in a wide number of topics such as regression setting. The least square approach is the first approach and most often works very well under certain assumption on the residuals. However, it is well known that when the residuals are distributed Gaussian or approximately Gaussian,

the least square is able to do the best job to construct the confidence interval and hypothesis test for the parameters in the regression model. But it is problematic when the assumptions are violated. For example, when the residuals have heavy-tailed distribution or even when there are few outliers. In this situations the bootstrap can help and we would like to apply the bootstrap approaches to make an inference about the parameters in the model.

In connection with regression setting there are two basic approaches to bootstrap depending on the problem at hand.

1. Observations resampling: first bootstrap the vector of the data including response and covariates and fit the model to the resample data. Directly sampling the observations would treat the covariates as random rather than fixed.

2. Residual resampling: in this approach first fit the model to the data and compute the residuals. Next bootstrap the residuals and generate the response as $y_b^* = X\hat{b} + e_b^*$ where X is the original covariates, y_b^* is the bootstrap sample and e_b^* is resample residual. Using $\{y_b^*, X\}$ to fit a model and estimate the parameters. In this approach the covariates X is used as fixed.

In the first scheme, the resampled design matrix does not equal the original design matrix. If the number of observations(cases) n, is moderately large, it is no problem but for small n and also when there are few observations with large effect on the design matrix, using wrong fitted model leads to take into account appropriate measure of uncertainty. This means that in this case, observation resampling is robust. The second approach is efficient when the correct model is used. So, for the residual resampling, careful model checking at first is necessary.

2.4.1 Bootstrap for functional data

In the previous section we discussed using bootstrap approach for the classic data. Now lets turn into the bootstrap procedure as resampling methodology for functional data(Cuevas *et al.* , 2006; Febrero-Bande & Oviedo de la Fuente, 2012). Suppose that $\boldsymbol{x}(t) = \{x_1(t), x_2(t), \cdots, x_n(t)\}, t \in T$ are observed functional data generated by underlying stochastic process $X \in L^2(T)$ with finite second moments. Suppose that we are going to make a confidence ball for a functional statistic $T(t) = T(X(t))$. The performance of the bootstrap confidence bands for functional data are constructed as follows for the given original data $\boldsymbol{x}(t)$:

1. Estimate the functional statistic using original data $T(t) = T(x_1(t), x_2(t), \cdots, x_n(t))$.

2. Generate K_B bootstrap sample from the original functional data called $\boldsymbol{x}_b^*(t) = \{x_1^*(t), x_2^*(t), \cdots, x_n^*(t)\}$ for the b resample. Note that in resampling for functional data, the curves are chosen with replacement.

3. Compute $T_b^*(t) = T(x_1^*(t), x_2^*(t), \cdots, x_n^*(t))$ as an estimator of the sample functional statistic from the b resample.

4. Compute $d_b = d(T(t), T_b^*(t)), b = 1, 2, \cdots, K_B$, where the metric $d(\cdot, \cdot)$ is associated with a norm, and define d_α the quantile $(1-\alpha)$ of the distance between the bootstrap resample and the sample estimate.

5. Construct the bootstrap confidence ball of level $(1-\alpha)$ as $CB_{(1-\alpha)} = T_b^*(t) \in E$ such that $d_b \leq d_\alpha$. In other words remove those curves $T_b^*(t)$ which the relevant distance d_b is larger than d_α.

In the classical bootstrap resampling methodology for univariate data, there are two alternatives bootstrap procedure called smoothed bootstrap and parametric bootstrap. As in bootstrap resampling the data are drawn with replacement method, it might be that some of the data replicated. The smoothed bootstrap is sometimes used in order to avoid the appearance of repeated measures in the artificial sample. The basic idea is replacing the standard bootstrap sample by the $x_b^0 = x_b^* + z_b$, where x_b^* is drawn from given data and z_b is independent from x_b^* and normally distributed $z_b \sim N(0, h)$.

In order to use smoothed bootstrap fo functional data setting is sufficient to replace $\boldsymbol{x}_b^*(t)$ by $\boldsymbol{x}_b^0(t)$ in step 2 where $\boldsymbol{x}_b^0(t) = \boldsymbol{x}_b^*(t) + z_b$ and $z_b \sim N_N(\mathbf{0}, s\Sigma)$. Here N_N and s refer to multivariate normal distribution and smoothing parameter respectively. In addition, the index N in multivariate normal distribution represents the number of observations per curve. Based on the experience, it turns out that in functional data setup, the smoothed bootstrap version works better(Febrero-Bande *et al.* , 2010).

In Paper II. we used the bootstrap procedure for functional logistic regression in order to inspect the uncertainty of the estimators.

3

Multilevel functional data

There are several studies in public health or some other disciplines when the observation are collected as functional data form (curves or images) on large number of subjects at multiple visits, levels or units (Morris *et al.* , 2001; Morris & Carroll, 2006; Schrack *et al.* , 2013; Crainiceanu *et al.* , 2009). On the one hand it is expedient to borrow the methods from standard multilevel modeling, and on the other hand we need to use the methods in functional data analysis in order to preserve the functional nature of the observation and for dimension reduction technique. The combination of these two branches has led to a new topic in statistics which known as multilevel functional data analysis.

Let us describe a well-known dataset in multilevel functional data literature, namely, colon carcinogenesis dataset. The data were collected in an experiment conducted by nutrition research at Texas A&M university, in order to investigate the interplay between diet and colon cancer at a cellular level. To this end, several groups of rats were fed a particular diets of interests for specific period, exposed to a carcinogen (radiation) that induces colon cancer and subsequently sacrificed for sample collection. The colon was resected from the rats and examined for the interesting biomarkers such as the concentration of p27, a cell cycle inhibitor protein, and apoptosis index. These multiple biomarkers are measured for each cell in multiple colonic crypts(Sgambato *et al.* , 2000). By considering these measurements as a function of cell position along the crypt wall, the problem becomes a multilevel functional data where the functional measurements at the crypt level are nested within rats, who are nested within diet group (Grambsch *et al.* , 1995). Several publications have been motivated by application to these dataset from different aspects, see Morris *et al.* (2003); Morris & Carroll (2006); Baladandayuthapani *et al.* (2008), and Staicu *et al.* (2010), among others. A sample of 10 functional observations for the rats with the fish diet has been shown in top panel of Figure 1.1. Most research in multilevel functional data concern to achieve an answer for the key scientific questions which can be grouped as decomposition of variability, group comparisons, functional regression and clustering.

Formally consider the setting where $i = 1, 2, \cdots, n$ index the subjects (rat), j index the visits or units(crypt), then the observed data for the ith subject are

$\{Y_{ij}(t_{ijl}), X_{ij}(t_{ijl}), j = 1, 2, \ldots, m_i\}$ where $Y_{ij}(t)$ denotes the outcome or response, which can be discrete or continuous. One important aim of the experiment was to evaluate the relationship between the multiple biomarker measured together on the same cell and the coordinated response. By assuming that $X_{ij}(t)$ is a proxy measurement of the following underlying subject-specific and unit-within-subject-specific functional signal and has the following decomposition

$$X_{ij}(t) = \mu(t) + Z_i(t) + U_{ij}(t) + \epsilon_{ij}(t) \tag{3.1}$$

where $\mu(t)$ is the mean function, $Z_i(t)$ and $U_{ij}(t)$ are independent random components, $Z_i(t)$ represents the subject-specific deviation from the mean, $U_{ij}(t)$ is the unit-specific random effect from the subject mean, and $\epsilon_{ij}(t)$ displays the noise. In addition, it is assumed that $Z_i(\cdot)$ and $U_{ij}(\cdot)$ are square integrable random processes on the closed and bounded set T which for simplicity it is usually considered $[0, 1]$. Furthermore, for identifiability Z_i, U_{ij}, and ϵ_{ij} are uncorrelated random processes with mean zero and ϵ_{ij} has covariance function that $\text{cov}(\epsilon_{ij}(t), \epsilon_{ij}(t')) = \sigma^2_\epsilon$, for $t = t'$ and 0 otherwise. In order to make a relationship between the response and covariate in this multilevel functional setting we will introduce a generalized time-varying regression in **?**, when the considered model relates the current value of the response at time t of the jth unit for the ith subject to the current value of the jth predictor at the same time t within the subject ith by considering the decomposition of $X_{ij}(t)$ in (3.1).

3.1 Correlated multilevel functional data

So far, we have focused on multilevel functional data when the label 'correlated' refers to the dependency of the functional unit which in functional data the observation of each signal are inherently correlated. As in functional data, all observation for each unit is considered as the observational unit, the correlation within observation of a unit is preserved automatically. So, using label 'correlated' in multilevel functional data refers to the dependency between units of the same subject. There are many possible situation in multilevel functional data however there is a high dependency between units which it is expedient to take this dependency into account in order to achieve a more accurate model. Estimating the correlation between the multiple signal for the same subject is one of the primary interests in some publications in correlated multilevel functional data. For instance, Morris *et al.* (2001, 2002) and Dubin & Müller (2005) estimated the correlation between the functions of the same subject while Li *et al.* (2007); Baladandayuthapani *et al.* (2008); Staicu *et al.* (2010); Zhou *et al.* (2010) estimated and considered the correlation (better to say spatial correlation) between the functions of the same subject in the functional modeling when analyzing the colon carcinogenesis data.

In the colon carcinogenesis data, there is scientific belief of coordinated response (Morris *et al.* , 2002) at the crypt level such that the biological response in one crypt may affect response in neighboring crypts. In particular of how p27 affects apoptosis. This is the motivation of the unfinished joint work with Ana-Maria Staicu and Damla Şentürk to introduce the 'Generalized time-varying spatial regression of multilevel functional data'. The idea of this work is. how we can make an association between a generalized spatially correlated multilevel functional response and a spatially correlated multilevel functional covariate. Again let $i = 1, 2, \cdots, n$, $j = 1, 2, \cdots, m_i$ index the subjects and units respectively and s_{ij} represents the spatial location of the jth unit within ith subject. Also, the observed data for subject i are displayed by

$$[\{Y_{ij}(t_{ijl,sij}), X_{ij}(t_{ijl,sij})\}, t_{ijl} \in [0,1], s_{ij}, j = 1, 2, \cdots, m_i]$$

where $Y_{ij}(t_{ijl}) = Y_{ij}(t_{ijl,sij})$ is the discrete or continuous-valued trajectory and $X_{ij}(t_{ijl}) = X_{ij}(t_{ijl,sij})$ is the functional predictor corresponding to the subject i, unit j with spatial location s_{ij} and t_{ijl} is the time point(distance the cell i the crypt from the crypt bottom) of the observation which in this dataset represents the distance of the cell on the crypt wall from the bottom of the crypt. Assuming that $X_{ij}(t)$ is a proxy ,measurement of the following subject-specific, unit within subject-specific and subject-specific signal as follows (Staicu *et al.* , 2010)

$$X_{ij}(t) = \mu(t) + Z_i(t) + U_{ij}(t) + W_i(s_{ij}) + \epsilon_{ij}(t) \tag{3.2}$$

where $Z_i(t)$ and $U_{ij}(t)$ are as defined before with the same assumptions and $W_i(t)$ represents the subject-specific spatial process and is considered a second order stationary process in $D \subseteq \mathbb{R}$. In addition, it is assumed that $W_i(t)$ is uncorrelated with $Z_i(t)$, $U_{ij}(t)$ and $\epsilon_{ij}(t)$.

The above simple practice decomposition of $X_{ij}(t)$ enables us to separate the different types of effects of the spatially correlated multilevel predictor on the response. We suggest that the distribution of $Y_{ij}(t_{ijl})$ can be described as follows

$$E\{Y_{ij}(t)|Z_i(t), U_{ij}(t), W_i(s+s_{ij}), s \in [-\Delta, \Delta]\} =$$
$$g\left\{\beta_0(t) + \beta_1(t)Z_i(t) + \beta_2(t)U_{ij}(t) + \int_{-\Delta}^{\Delta} \gamma(s)W_i(s+s_{ij})\,ds\right\} \quad , \quad \Delta < s_{ij} < L - \Delta \tag{3.3}$$

for a known inverse link function $g(\cdot)$ and L in the length of D. In (3.3), $\beta_0(t)$ is the intercept, $\beta_1(t)$ is the time-varying effect of the subject-specific deviation Z_i, $\beta_2(t)$ is the effect of the unit-specific deviation $U_{ij}(t)$, and $\gamma(s)$ represents the spatial effect which quantify the effect of the neighborhood of the subject-specific spatial deviation W_i in a window of size Δ on the response at site s_{ij}. The function $\gamma(s)$ is a real-valued defined on $[-\Delta, \Delta]$ which for identifiability is assumed to be symmetry about zero with the property $\int_{-\Delta}^{\Delta} \gamma^2(s)\,ds = 1$.

The idea of estimation of β_0, β_1 and β_2 is mostly based on **?** which this approach is based on method-of-moment estimator of mean function and covariance functions. Simulation study shows that the estimator for these parameter functions work well. The attempt to find the estimator for $\gamma(s)$ is not promising and we are still working on the estimator.

4

Perspective

In this chapter we discuss some perspective suggested by the work done in Paper I, III and possible future work.

In Paper I, the analysis of horse lameness data through multinomial functional regression was conducted by using vertical acceleration signals. This is a general idea, and we must take full advantage of the existent data. So, as in the lameness of horse data, acceleration signals for other directions, namely, transverse and longitudinal, are collected, it would be expedient to consider the signals for all three direction when analyzing the data. This consideration could led to better result in classification.

In the lameness application the 85 signals (from 8 horses) were considered as independent signals. Ideally, we need to consider the effect of the horses in the analysis and this effect must be considered as random effect in the model.

In application of multinomial functional regression, we considered amplitude variation. Also in Chapter 2 we computed a W-score based on phase variation. But it could be a good idea to use a combination of phase variation and amplitude variation with the aim of distinguish.

Regarding the lameness of horse dataset, the data are collected for two degree of lameness in groups LH, RH, RF, RH. But in the analysis of the data, we did not consider the degree of lameness in the model. This is important as one of the aims considering the degree of lameness in the analysis.

In multinomial functional regression, we used the raw discrete wavelet transform as design matrix in the regression model. In functional data, we usually deal with data which are contaminated with noise. One potential way to remove the noise from the data could be using (soft or hard) threshold on the wavelet coefficients in order to identify those who are associated to the noise and then modify them. For very dense functional data and observation with noise, we could consider thresholding of the wavelet coefficient as design matrix in the model.

In order to maintain more information in wavelet transform, one could apply appropriate high- and low-pass filters to the data at each level to produce two sequences at each level. In this case no decimation occurs and so the two produced sequences

have the same length as the original sequence. Instead the filters are modified at each level by padding them to with zeros. This transform is well-known as th non-decimated wavelet transform (Nason & Silverman, 1995). This transform has been applied in statistical research in particular for non-parametric regression and time series analysis, but to the best of our knowledge has not applied in functional data analysis.

In some situations, i.e. colon carcinogenesis data, there is spatial dependency between units within subject. On the face of it, it is expedient to consider this dependency when decomposing the functional covariate as it has done in Staicu *et al.* (2010). So, the suggested concurrent model in Paper III can be extended for multilevel functional data with spatially dependency by considering the spatial effect in the model.

Bibliography

Baladandayuthapani, Veerabhadran, Mallick, Bani K, Young Hong, Mee, Lupton, Joanne R, Turner, Nancy D, & Carroll, Raymond J. 2008. Bayesian hierarchical spatially correlated functional data analysis with application to colon carcinogenesis. *Biometrics*, **64**(1), 64–73.

Chernick, Michael R. 2007. *Bootstrap methods: A guide for practitioners and researchers*. Vol. 619. Wiley-Interscience.

Crainiceanu, Ciprian M, Caffo, Brian S, Di, Chong-Zhi, & Punjabi, Naresh M. 2009. Nonparametric signal extraction and measurement error in the analysis of electroencephalographic activity during sleep. *Journal of the american statistical association*, **104**(486), 541–555.

Cuevas, Antonio, Febrero, Manuel, & Fraiman, Ricardo. 2006. On the use of the bootstrap for estimating functions with functional data. *Computational statistics & data analysis*, **51**(2), 1063–1074.

Dubin, Joel A, & Müller, Hans-Georg. 2005. Dynamical correlation for multivariate longitudinal data. *Journal of the american statistical association*, **100**(471), 872–881.

Efron, Bradley. 1979. Bootstrap methods: another look at the jackknife. *The annals of statistics*, 1–26.

Febrero-Bande, Manuel, & Oviedo de la Fuente, Manuel. 2012. Statistical computing in functional data analysis: the r package fda. usc. *Journal of statistical software*, **51**(4), 1–28.

Febrero-Bande, Manuel, Galeano, Pedro, & González-Manteiga, Wenceslao. 2010. Measures of influence for the functional linear model with scalar response. *Journal of multivariate analysis*, **101**(2), 327–339.

Grambsch, Patricia M, Randall, Bryan L, Bostick, Roberd M, Potter, John D, & Louis, Thomas A. 1995. Modeling the labeling index distribution:

an application of functional data analysis. *Journal of the american statistical association*, **90**(431), 813–821.

Halling Thomsen, Maj, Tolver Jensen, Anders, Sørensen, Helle, Lindegaard, Casper, & Haubro Andersen, Pia. 2010. Symmetry indices based on accelerometric data in trotting horses. *Journal of biomechanics*, **43**(13), 2608–2612.

Halvorsen, Kjetil. 2012. Elemstatlearn: data sets, functions and examples from the book:'the elements of statistical learning, data mining, inference, and prediction'(2012.04–0 edn). *See http://cran. r-project. org/web/packages/elemstatlearn.*

Li, Yehua, Wang, Naisyin, Hong, Meeyoung, Turner, Nancy D, Lupton, Joanne R, & Carroll, Raymond J. 2007. Nonparametric estimation of correlation functions in longitudinal and spatial data, with application to colon carcinogenesis experiments. *The annals of statistics*, 1608–1643.

Morris, Jeffrey S, & Carroll, Raymond J. 2006. Wavelet-based functional mixed models. *Journal of the royal statistical society: Series b (statistical methodology)*, **68**(2), 179–199.

Morris, Jeffrey S, Wang, Naisyin, Lupton, Joanne R, Chapkin, Robert S, Turner, Nancy D, Young Hong, Mee, & Carroll, Raymond J. 2001. Parametric and nonparametric methods for understanding the relationship between carcinogen-induced dna adduct levels in distal and proximal regions of the colon. *Journal of the american statistical association*, **96**(455), 816–826.

Morris, Jeffrey S, Wang, Naisyin, Lupton, Joanne R, Chapkin, Robert S, Turner, Nancy D, Hong, Meeyoung, & Carroll, Raymond J. 2002. A bayesian analysis of colonic crypt structure and coordinated response to carcinogen exposure incorporating missing crypts. *Biostatistics*, **3**(4), 529–546.

Morris, Jeffrey S, Vannucci, Marina, Brown, Philip J, & Carroll, Raymond J. 2003. Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis. *Journal of the american statistical association*, **98**(463), 573–583.

Nason, Guy P, & Silverman, Bernard W. 1995. The stationary wavelet transform and some statistical applications. *Lecture notes in statistics-new york-springer verlag-*, 281–281.

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

Ramsay, James O, & Dalzell, CJ. 1991. Some tools for functional data analysis. *Journal of the royal statistical society. series b (methodological)*, 539–572.

Ramsay, James O, & Silverman, Bernard W. 2002. *Applied functional data analysis: methods and case studies*. Vol. 77. Springer New York.

Schrack, Jennifer A, Zipunnikov, Vadim, Goldsmith, Jeff, Bai, Jiawei, Simonsick, Eleanor M, Crainiceanu, Ciprian, & Ferrucci, Luigi. 2013. Assessing the "physical cliff": detailed quantification of age-related differences in daily patterns of physical activity. *The journals of gerontology series a: Biological sciences and medical sciences*, glt199.

Sgambato, Alessandro, Cittadini, Achille, Faraglia, Beatrice, & Weinstein, I Bernard. 2000. Multiple functions of p27kip1 and its alterations in tumor cells: a review. *Journal of cellular physiology*, **183**(1), 18–27.

Sørensen, Helle, Tolver, Anders, Thomsen, Maj Halling, & Andersen, Pia Haubro. 2012. Quantification of symmetry for functional data with application to equine lameness classification. *Journal of applied statistics*, **39**(2), 337–360.

Staicu, Ana-Maria, Crainiceanu, Ciprian M, & Carroll, Raymond J. 2010. Fast methods for spatially correlated multilevel functional data. *Biostatistics*, **11**(2), 177–194.

Zhou, Lan, Huang, Jianhua Z, Martinez, Josue G, Maity, Arnab, Baladandayuthapani, Veerabhadran, & Carroll, Raymond J. 2010. Reduced rank mixed effects models for spatially correlated hierarchical functional data. *Journal of the american statistical association*, **105**(489), 390–400.

Papers

I

Multinomial Functional Regression with Wavelet and LASSO Penalization

Seyed Nourollah Mousavi
Department of Mathematical Sciences
University of Copenhagen

Helle Sørensen
Department of Mathematical Sciences
University of Copenhagen

Publication details

Revising to submit (2015).

Multinomial functional regression with wavelets and LASSO penalization

Seyed Nourollah Mousavi*
Department of Mathematical Sciences
University of Copenhagen, Denmark
nourollah@math.ku.dk

Helle Sørensen
Department of Mathematical Sciences
University of Copenhagen, Denmark
helle@math.ku.dk

Abstract

We study the situation with a categorical response variable (more than two classes) and a functional predictor and suggest to use a multinomial functional regression (MFR) model for the analysis. We combine the discrete wavelet transform and LASSO penalization for estimation, and the fitted model is used for classification of new curves with unknown class membership. We apply our MFR approach to two datasets, one regarding lameness detection for horses and another regarding speech recognition. In the applications, as well as in a simulation study, we compare the performance of our MFR approach to that of other methods for supervised classification of functional data.

Key words: Multinomial functional regression; Discrete wavelet transform; LASSO penalization; Supervised classification; Lameness data for horses; Phoneme data.

1 Introduction

Detection of lameness for horses and identification of the lame limb is a difficult task even for experienced veterinarians, and there is a need for objective methods that can be used as supplement to the usual clinical examination and visual inspection of the horse. In this paper we examine if acceleration signals collected during trot can be used for diagnostic purposes. Data are available from eight horses, each in nine conditions corresponding to no lameness and to low- and moderate-degree lameness on either of the four limbs. Another application comes from speech recognition where the aim is to predict which phoneme is spoken, based on a log-periodogram.

From a statistical point of view both problems can be framed as classification for functional data. The observation is a function (acceleration signal or log-periodogram), and we want to classify new observations into well-specified groups (lameness status or phoneme). We propose

*Corresponding author

a method based on multinomial functional regression (MFR) which, apart from the classification itself, also gives us information about which parts of the signals are used in the classification procedure. It combines wavelet expansions with LASSO regularization.

Methods for analysis of functional data have been developed since the 1960's and 1970's, and the development accelerated in the 1990's with the first edition of Ramsay and Silverman (2005) from 1997 as an important milestone. These days, functional data analysis (FDA) is a statistical discipline in itself which develops fast and vividly in many directions. This is illustrated by a large number of hits, more than 6000, for the phrase "functional data" in articles since 2014 on Google Scholar (`http://scholar.google.dk`). The development is driven by the technical development which has resulted in a vast amount of data of functional nature. Examples from the literature cover a broad range of scientific fields and include growth data, spectral data from food and plant science, medical data concerning CD4 counts for HIV patients, brain images and vascular geometry, weather, climate data and pollution data, and data on speech recognition.

There are several approaches in the literature to classification of functional data. Early work include Hall et al. (2001) and James and Hastie (2001) who used linear discriminant analysis (LDA) on scores from a principal component analysis (PCA) and on coefficients from spline expansions, respectively, and Ferraty and Vieu (2003) using a kernel approach. Later, PCA was combined with logistic regression for the case with two groups (Müller and Stadtmüller, 2005), the method of partial least squares (PLS) was accommodated to functional data (Preda et al., 2007), and methods based on functional depth were suggested (Cuevas et al., 2007; López-Pintado and Romo, 2006). Recently Tian and James (2013) suggested a dimension reduction approach that takes into account the association to the categorical variable, and Delaigle and Hall (2012) studied optimality properties of a nearest centroid classifier.

Another corner of FDA is devoted to regression problems with functional outcome and/or predictors. The situation with scalar response and functional covariates is of particular interest for this paper. In the simplest case we observe for each subject i a one-dimensional continuous response Y_i and a function $x_i : (0,1) \to \mathbb{R}$, and assume (among others) that the conditional expectation of Y_i given x_i is given by

$$\mathrm{E}[Y_i|x_i] = \alpha + \int_0^1 \beta(t)x_i(t)\,dt, \tag{1}$$

where α is an unknown intercept and $\beta : (0,1) \to \mathbb{R}$ is an unknown coefficient function.

Several estimation approaches have been suggested for this model. One method, often referred to as functional principal component regression (FPCR), consists of a functional principal component analysis of the x_i's followed by a regression on the first few, say K, scores (Cardot et al., 1999; Ramsay and Silverman, 2005). This yields β functions in the space spanned by the first K principal components (PCs), and it is thus implicitly assumed these PCs not only account for a large proportion of the variation between x_i's, but are also relevant for the association between Y and x. Lee and Park (2012) discussed a selection approach to choose the most informative PC basis using LASSO, i.e., imposing a L^1 penalty on β, and the effect of a quadratic penalty on β in FPCR was discussed by Randolph et al. (2012).

Another approach is to use a rich, flexible basis for β in combination with regularization methods. For example, Marx and Eilers (1999) and Cardot et al. (2003) used spline series expansions and added penalty terms to the log-likelihood function, and Goldsmith et al. (2011) and Wood (2011) used spline series expansions in a mixed-model set-up. Reiss and Ogden (2007) combined FPCR, functional partial least squares, and penalized splines. The paper by Zhao et al. (2012) is of particular importance for this paper and combined wavelet expansions with LASSO regression. This is an effective combination, since LASSO penalization by construction selects sparse models, and wavelets are known to offer sparse, yet precise, representations of many types of functions. The LASSO has also been used in combination with other basis systems in order to obtain sparse representations (James et al., 2009; Lee and Park, 2012).

Many of the above-mentioned methods also apply to exponential families, in particular to the case with binary response leading to functional logistic regression (Cardot and Sarda, 2005; Crainiceanu et al., 2009; Goldsmith et al., 2011; James, 2002; Müller and Stadtmüller, 2005). Most of the papers contain asymptotic results but there are only few examinations of finite-sample properties in non-Gaussian cases. An exception is the paper by Reiss et al. (2015) where Gaussian and logistic regression with image predictors are studied.

We will take the logistic regression set-up a step further and consider multinomial regression with functional covariates. Let x_i be as before, but consider categorical outcomes Y_i with M possible outcomes, $m \in \mathscr{M}$. Define $p_m(x)$ as the conditional probability of class m given the functional outcome,

$$p_m(x) = P(Y = m | X = x), \quad m \in \mathscr{M},$$

and assume that $p_m(x)$ is proportional to $\exp\left(\alpha_m + \int_0^1 \beta_m(t)x(t)\,dt\right)$ for class-specific intercepts α_m and class-specific coefficient functions β_m. Once the model has been fitted, it can be used for classification in the obvious way: Given a curve, we compute $\hat{p}_m(x)$ for all m and allocate the curve to the group with highest probability.

For estimation, we will follow the approach from Zhao et al. (2012) closely. More specifically, we select a family of wavelet bases and a resolution level, expand the covariate functions in the basis and use the wavelet coefficients as covariates in a multinomial regression with LASSO penalization. The LASSO tuning parameter and resolution level are selected by cross validation. The regression coefficients from the optimal multinomial regression are extracted and translated into estimated coefficient functions, $\hat{\beta}_m$.

In summary, the aim of the paper is to generalize the wavelet- and LASSO-based regression approach from Zhao et al. (2012) to the multinomial case and use it for classification. Our main application is about diagnosis of lameness among horses, but we also apply our method to a dataset on phonemes. This dataset has been widely used in speech recognition and was discussed by Hastie et al. (1995) and Ferraty and Vieu (2003), among others.

The rest of the paper is organized as follows. The motivating dataset on lameness is described in detail in Section 2. In Section 3 we go through the details about the functional multinomial regression, including formulation of the model and brief discussions about wavelets and the LASSO. Section 4 contains a thorough analysis of the lameness data, and Section 5 presents a simulation study. In Section 6 the phoneme data are classified using MFR. Finally, we discuss

the results and conclude in Section 7. Supplementary material including details on preprocessing of the lameness data and an introduction to wavelets is available in appendices A and B.

2 Detection of lameness: Motivation and data

Our main application is concerned with lameness detection for horses. It is well documented that there is large variation between different veterinarians' evaluation of the same horse and even between multiple evaluations undertaken by the same veterinarian (Keegan et al., 2010, 1998). Therefore, several research groups have worked with more objective/automated evaluations as supplement to the usual examination (Pfau et al., 2005; Weishaupt et al., 2004). A group from University of Copenhagen has worked with acceleration signals (Thomsen et al., 2010), and our interest in this paper is whether the acceleration signals can be used for diagnostic purposes.

The acceleration signals are collected while the horses are trotting. Trot is a two-beat gait where the diagonal pairs of legs (left-fore/right-hind and right-fore/left-hind) move forward at the same time with a moment of suspension between each beat. A complete gait cycle thus consists of two parts corresponding to stance on each of diagonal. A healthy horse is hypothesized to trot symmetrically such that the two parts are alike up to random variation. Lameness is known to disturb this symmetry as the horse tries to reduce the pressure onto the ground for the injured limb (Weishaupt et al., 2004). Pressure generates upwards acceleration, and the idea in the current study is that asymmetry, and thereby lameness, can be detected from acceleration signals.

2.1 Experimental design

The dataset consists of a total of 85 acceleration signals from eight horses, who went through a thorough clinical examination and showed no indication of lameness. The data was collected in two sub-experiments. In the first sub-experiment four horses were tested four times each in healthy condition (no lameness). In the second sub-experiment all eight horses were tested nine times, namely in healthy condition and after induction of two degrees of lameness on each of the four legs. The lameness was induced mechanically by equipping the horse with a modified horseshoe with a screw eliciting pressure on the sole of the hoof. The shoe makes stance on the limb painful, and the two degrees of lameness correspond to different levels of this pressure. Three signals are unavailable resulting in a total of 85 signals. The data have been analyzed in Sørensen et al. (2012) and Thomsen (2010).

We will not distinguish between the two degrees of lameness in this paper. This leaves us with five lameness groups, in the following referred to as NO, LF, LH, RF, RH corresponding to normal or healthy condition and lameness on left-fore, left-hind, right-fore and right-hind leg, respectively.

2.2 Data collection and preprocessing

We refer to Thomsen et al. (2010) for details on the technical description of the data collection process. In short, a three-axis accelerometer was placed at the back of the horse (close to body center of mass), and the horse was led by the hand in trot at approximately constant velocity. The horse was videotaped during measurement. The accelerometer measured the 3-dimensional accelerations of trunk movements at frequency 240 Hz. We will only use the acceleration in vertical direction in this paper.

Several preprocessing steps were carried out before the MFR analysis in order to reduce variation between gait cycles in each signal and variation between signals due to slightly different timing at the beginning of the signals. These steps are described in detail in appendix A, but in short they consist of the following steps: (a) eight gait-cycles were picked out, based on the video recording; (b) the signal was smoothed lightly to bring it on functional form; (c) seven gait cycles starting at a well-defined zero-crossing were identified and aligned, and thereafter averaged; (d) the average was shift aligned to a minus cosine curve. Time is measured in gait cycles, so clock time has been scaled separately for each signal. After preprocessing the data consist of 85 real-valued functions x_i defined on $(0,1)$. Each x_i is periodic, as it is expressed in a Fourier basis, and the first peak always corresponds to stance on the RF/LH diagonal.

The 85 preprocessed signals and the group-wise mean curves are shown in Figure 1. Notice how the signals in the NO group seem to be symmetric in the sense that the sub-signals on $(0,0.5)$ and $(0.5,1)$ corresponding to stance on the two diagonals are very similar, whereas this is not the case for the other groups. For example, in the LH group, the amplitude appears to be larger on $(0.5,1)$ compared to $(0,0.5)$. This is quite reasonable: When the horse has the special horse shoe attached to the left-hind limb, then one would expect it to hurt to stand on the right-fore/left-hind diagonal, and thus we would expect the horse to put less pressure on the ground with this diagonal compared to the other. This exactly corresponds to a smaller amplitude of the first compared to the second half of the signal.

3 Multinomial functional regression

In this section we describe the model and the estimation procedure in detail. We consider data (x_i, y_i), $i = 1,\dots,n$ from n individuals, and assume that they are outcomes from independent random variables (X_i, Y_i). The response variables Y_i are categorical with two or more possible outcomes. The state space is denoted $\mathscr{M}$, and we use m to denote outcomes. The explanatory variables X_i are functional and defined on a regular time interval, which for simplicity is taken as the unit interval, i.e., $(0,1)$. Hence, each observed variable is a function, $x_i : (0,1) \to \mathbb{R}$. In practice each x_i is observed at discrete sample points.

In our approach, dealing with the discrete wavelet transform, the number of observations must be a power of 2, but this is not really restrictive: In other dense cases, we could make linear interpolation between observation points, and thereby get a function that could be evaluated at a vector of time points with the desired length. This technique also makes it possible to deal

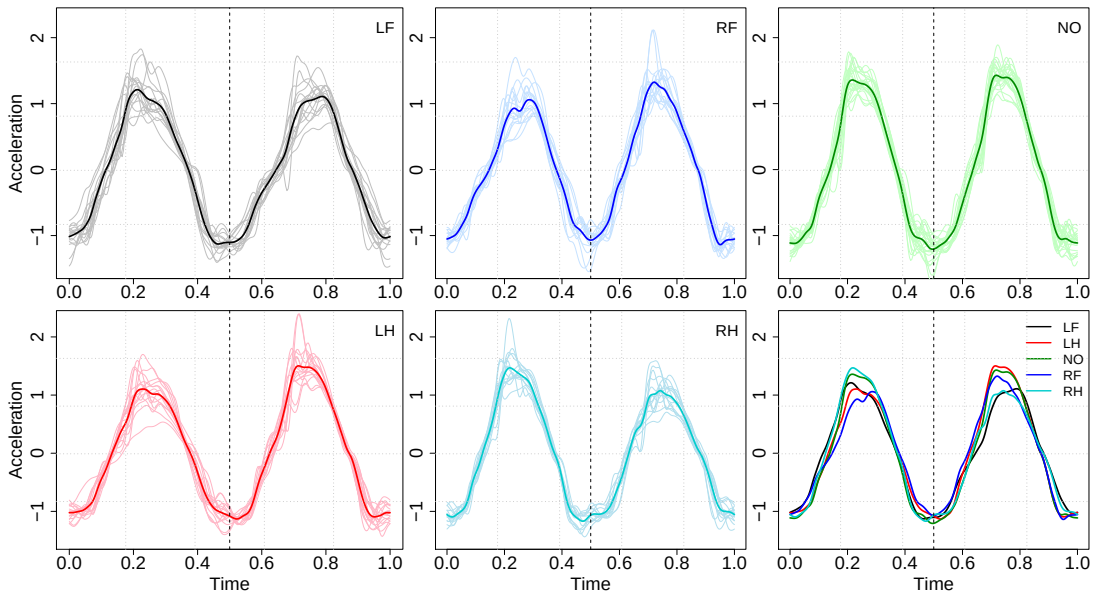

Figure 1: The 85 preprocessed acceleration signals (thin lines) organized after lameness status and mean curves (thick lines). The five mean curves are also shown in the bottom right panel.

with irregularly spaced observations and situations where the number of sample points differ between X_i's. For sparse data we could predict the trajectory, for example by the approach in Goldsmith et al. (2013), and then evaluate it at the desired time points. In the following we therefore assume that each X_i is sampled at N equally spaced time points ranging from 0 to 1.

3.1 Regression model and prediction

As in standard regression problems, we are interested in the association between X and Y. In many applications it would be most natural to think about this association and the data generating mechanism in terms of the conditional distribution of X given Y. For example, how does the distribution of acceleration signals differ between lameness groups? However, since the primary purpose of our work is classification, i.e., prediction of Y given X, we are rather interested in the conditional distribution of Y given X and therefore aim at estimation of a model for the conditional probabilities, i.e.,

$$p_m(x) = P(Y = m|X = x), \quad m \in \mathscr{M}.$$

More specifically, and as a natural extension of the conventional multinomial regression model, we consider unknown constants α_m and unknown coefficient functions $\beta_m : (0,1) \to \mathbb{R}$, and assume that

$$p_m(x) = \frac{e^{\alpha_m + \int \beta_m(t)x(t)dt}}{\sum_{l \in \mathscr{M}} e^{\alpha_l + \int \beta_l(t)x(t)dt}}. \tag{2}$$

Here, the integrals are from 0 to 1, and the quantities $\eta_m(x) = \alpha_m + \int \beta_m(t)x(t)dt$ will be referred to as linear predictors. Often, and this will also be the case in our lameness application, a reference group is selected, and the other groups are compared against the reference group. For two groups, m and l, the conditional log-odds for group m compared to group l given x is

$$\log \frac{p_m(x)}{p_l(x)} = \eta_m(x) - \eta_l(x) = (\alpha_m - \alpha_l) + \int_0^1 \big(\beta_m(t) - \beta_l(t)\big)x(t)\,dt. \tag{3}$$

The differences between coefficient functions are thus used to model how these odds are affected by x. The intercept term is more difficult to interpret, but notice that we can reparameterize by centering the x's:

$$\eta_m(x) = \kappa_m + \int \beta_m(t)\big(x(t) - \bar{x}(t)\big)dt \tag{4}$$

where $\bar{x}$ is the pointwise mean of $x_1, \ldots, x_n$, and $\kappa_m = \alpha_m + \int \beta_m(t)\bar{x}(t)dt$. Then

$$\log \frac{p_m(\bar{x})}{p_l(\bar{x})} = \kappa_m - \kappa_l,$$

and the κ's model the probabilities for the average signal.

Notice that model (2) is overparameterized since we may add the same constant c_0 to all α's and/or add the same constant c to all β's, and yet get the same probabilities. This ambiguity is dealt with in a natural way by the LASSO regularization, see Section 3.4.

Once we have fitted the model such that estimates $\hat{\alpha}_m$ and $\hat{\beta}_m$ are available for $m \in \mathscr{M}$ the fitted model can be used for classification. For a new function x, compute the estimated linear predictor

$$\hat{\eta}_m(x) = \hat{\alpha}_m + \int_0^1 \hat{\beta}_m(t)x(t)dt$$

and the corresponding probability $\hat{p}_m(x)$ for each group. As prediction of Y, choose the group with largest probability:

$$\hat{Y}(x) = \operatorname{argmax}_{m\in\mathscr{M}} \hat{p}_m(x).$$

Estimates of $\hat{\beta}_m(t)$ tell us how this prediction machine works: Which parts of the curves are used for prediction, and which are not?

3.2 Wavelets

The basic idea behind wavelets is to represent a complex function with simple functions at different scales and locations. The simple functions form a wavelet basis and are generated from a given wavelet by the operation of dilation and translation. The wavelet representation thus describes the features of the function at different locations and scales, and wavelets are particularly useful for describing non-stationarity and discontinuities. In this section we briefly recall the basics for wavelets and wavelet bases for $L^2(\mathbb{R})$. Appendix B gives more details,

and comprehensive accounts of wavelets can be found in Daubechies et al. (1992), Vidakovic (2009), and Nason (2010).

A wavelet transform has a mother wavelet, $\psi(t)$, and a father wavelet, $\phi(t)$, that are linked by the relationship, $\psi(t) = \sum_{k\in\mathbb{Z}} g_k \sqrt{2}\phi(2t-k)$. The set of the coefficients $\mathscr{G} = \{g_k\}_{k\in\mathbb{Z}}$ are the high-pass filter coefficients associated with the particular wavelet function.

For a given mother wavelet $\psi(t)$, the wavelet basis is given by $\{\psi_{j,k}(t)\}_{j,k\in\mathbb{Z}}$ where

$$\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k).$$

The indices j and k represent dilation and translation, respectively. The index k shows the coefficient's position and is known as the location parameter. The index j represents the detail level and is known as the scale parameter.

A discrete wavelet transform (DWT) is a linear transformation that operates on a discrete series $x_1, x_2, \cdots, x_N$ where N is a power of 2, transforming it into a numerically different series of the same length. Keep in mind that the observed series is generated from an underlying signal $f(t)$ such that $x_j = f(t_j)$ The idea is to filter the series, using the high- and low-pass filters associated with the wavelet, i.e., $\mathscr{G} = \{g_k\}_{k\in\mathbb{Z}}$ and $\mathscr{H} = \{h_k\}_{k\in\mathbb{Z}}$, to obtain the wavelet coefficients. Figure 2 shows the DWT computation as a cascade of filtering followed by a factor 2 subsampling; `H` and `L` represent high- and low-pass filters, respectively, and $\downarrow 2$ denotes subsampling.

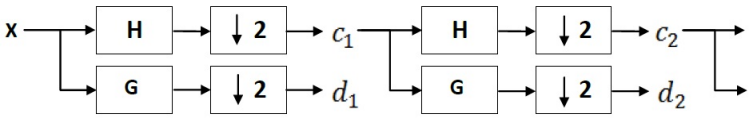

Figure 2: Graphical display of a multiscale transform.

For a sequence of length $N = 2^J$, there are J different detail levels, $j = 0, 1, \cdots, J-1$. At level j the wavelets cover one of 2^j sub-intervals, and $j = J-1$ is thus the finest detail level and $j = 0$ the coarsest detail level. For a fixed detail level j_0, a fine-scale representation of the function f at detail level $j = j_0$ is given by

$$f_{j_0}(t) = \sum_{k=0}^{2^{j_0}-1} c_{j_0,k}\phi_{j_0,k}(t) + \sum_{j=j_0}^{J-1}\sum_{k=0}^{2^j-1} d_{j,k}\psi_{j,k}(t). \tag{5}$$

Notice that coefficients associated with the father and mother wavelet are denoted by $c_{j,k}$ and $d_{j,k}$, respectively. The functions f_j and f_{j+1} belong to subspaces V_j and V_{j+1}, respectively, with $V_j \subset V_{j+1}$, and the difference between f_{j+1} and f_j thus consists of the details in $V_{j+1}\backslash V_j$. The detail level j_0 will used as a tuning parameter in the estimation procedure, see Section 3.5.

We have several reasons for choosing wavelets as our workhorse. First, wavelet analysis extracts information about local properties due to the compact support of the basis functions. This is in contrast to Fourier analysis which is better at describing global than local features. Second, precise approximations are often obtained with sparse wavelet representations, i.e.,

with representations that only contain relatively few non-negligible terms. Hence, important features can be highlighted by a few non-zero coefficients which makes wavelet analysis and LASSO penalization good companions.

We use We use the least asymmetric Daubechies wavelets for our analyses, see Figure 16 in Appendix B.

3.3 LASSO

The standard linear regression model for n scalar observations and p predictors can be written as $Y_i = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j + \varepsilon_i$ $(i = 1, 2, \ldots, n)$, where ε_i is a random error. Ordinary least squares (OLS), minimizing the residual sum of squares, is the simplest estimation method:

$$\hat{\beta}^{\text{OLS}} = \text{argmin}_\beta \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right\}. \tag{6}$$

If the remainder terms ε_i are iid. Gaussian, then OLS is equivalent to maximum likelihood estimation.

If the number of covariates is large in comparison to the sample size, OLS leads to over-fitting and models that are hard to interpret, and if $p > n$ the OLS solution is not unique. Therefore it is common to apply regularization techniques. The methods include dimension reduction methods like principal component analysis (PCA), partial least squares (PLS), sufficient dimension reduction (SDR), usage of information criteria like AIC and BIC, and shrinkage methods like ridge regression (RR), least absolute shrinkage and selection operator (LASSO), and elastic nets (Hastie et al., 2009).

We will use the LASSO introduced by Tibshirani (1996) for our estimation problem. In the standard regression problem this amounts to adding a term with the L^1 norm of the coefficient vector to the OLS criterion:

$$\hat{\beta}^{\text{LASSO}} = \text{argmin}_\beta \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{7}$$

Here, λ is a tuning parameter that controls the amount of shrinkage. The L^1 penalization has the effect that each β_j is shrinked toward the origin and some β_j are even driven to zero. Hence, LASSO results in a sparse estimate of β if λ is chosen large.

The response in our set-up is multinomial, and we will add the L^1 penalty term to the deviance (minus the log-likelihood) rather than to an OLS criterion, see the next section.

3.4 Penalized maximum likelihood

We are now ready to combine multinomial functional regression, wavelets, and LASSO and thus describe the estimation procedure in detail.

For a start, consider the tuning parameters, λ_0 and j_0, fixed. The observed covariate functions as well as the unknown coefficients functions are expressed in the wavelet basis at detail level j_0:

$$x_{i,j_0}(t) = \sum_{k=0}^{2^{j_0}-1} c_{i,j_0,k}\phi_{j_0,k}(t) + \sum_{j=j_0}^{J-1}\sum_{k=0}^{2^j-1} d_{i,j,k}\psi_{j,k}(t), \quad i = 1,\dots,n$$
$$\beta_{m,j_0}(t) = \sum_{k=0}^{2^{j_0}-1} c^*_{m,j_0,k}\phi_{j_0,k}(t) + \sum_{j=j_0}^{J-1}\sum_{k=0}^{2^j-1} d^*_{m,j,k}\psi_{j,k}(t), \quad m \in \mathscr{M} \tag{8}$$

where c, c^* and d, d^* are coefficients associated with the father and mother wavelet, respectively, and for the signals $x_i(t)$ and coefficient function $\beta(t)$, respectively. Recall that the wavelet basis is an orthonormal basis. Therefore the linear predictors can be expressed in terms of the wavelet coefficients by

$$\eta_m(x_i) = \alpha_m + \int \beta_m(t)x_i(t)dt = \alpha_m + \sum_{k=0}^{2^{j_0}-1} c_{i,j_0,k}c^*_{m,j_0,k} + \sum_{j=j_0}^{J-1}\sum_{k=0}^{2^j-1} d_{i,j,k}d^*_{m,j,k} = \alpha_m + Z_i\gamma_m \tag{9}$$

where Z_i is the $1\times N$ matrix (row vector) consisting of c and d coefficients for subject i, and γ_m is the $N\times 1$ matrix (column vector) consisting of c^* and d^* coefficients for group m in the same order. We recognize this structure for the linear predictors from the classical multinomial regression.

If we furthermore introduce the indicator variables $w_{im} = \mathbf{1}_{(y_i=m)}$, then the contribution from subject i to the likelihood can be written as

$$p_{y_i}(x_i) = \prod_{m\in\mathscr{M}} p_m(x_i)^{w_{im}} = \frac{\prod_{m\in\mathscr{M}}\left(e^{\alpha_m+Z_i\gamma_m}\right)^{w_{im}}}{\sum_{m'\in\mathscr{M}} e^{\alpha_{m'}+Z_i\gamma_{m'}}},$$

and the complete log-likelihood is

$$\log L\Big((\alpha_m,\gamma_m)_{m\in\mathscr{M}}\Big) = \log\prod_{i=1}^n p_{y_i}(x_i) = \sum_{i=1}^n\Big(\sum_{m\in\mathscr{M}} w_{im}(\alpha_m+Z_i\gamma_m) - \log\sum_{m\in\mathscr{M}} e^{\alpha_m+Z_i\gamma_m}\Big).$$

With L^1 penalty on the γ coefficients and fixed penalty parameter (λ_0) our penalized log-likelihood is thus

$$\begin{aligned} Q\Big((\alpha_m,\gamma_m)_{m\in\mathscr{M}}\Big) &= -\log L\Big((\alpha_m,\gamma_m)_{m\in\mathscr{M}}\Big) + \lambda_0\sum_{m\in\mathscr{M}}\sum_{r=1}^N|\gamma_{mr}| \\ &= -\sum_{i=1}^n\Big(\sum_{m\in\mathscr{M}} w_{im}(\alpha_m+Z_i\gamma_m) - \log\sum_{m\in\mathscr{M}} e^{\alpha_m+Z_i\gamma_m}\Big) + \lambda_0\sum_{m\in\mathscr{M}}\sum_{r=1}^N|\gamma_{mr}| \end{aligned}$$

which is minimized w.r.t. (α_m,γ_m), $m\in\mathscr{M}$.

As already mentioned, the model is overparameterized. From equation (9) we notice that parameter sets $(\alpha_m - c_0, \gamma_m - c)_{m\in\mathscr{M}}$ and $(\alpha_m,\gamma_m)_{m\in\mathscr{M}}$ give rise to the same linear predictors and

thus to the same likelihood for all $c_0 \in \mathbb{R}$ and all $c \in \mathbb{R}^N$ (not depending on m). The penalty term is not the same, however, and regularization therefore takes care of the ambiguity and makes it possible to obtain uniquely determined parameter estimates. For fixed $(\alpha_m, \gamma_m)_{m\in\mathscr{M}}$, the smallest LASSO penalty term among parameter sets $(\alpha_m, \gamma_m - c)_{m\in\mathscr{M}}$ is obtained for c being the vector of pointwise medians (Friedman et al., 2010, Theorem 1):

$$c_r = \text{median}\big\{(\gamma_{m,r})_{m\in\mathscr{M}}\big\}, \quad r = 1,\ldots,N.$$

We therefore center the γ's such that the median is zero for all coordinates. The intercepts, which are not penalized, are centered such that they have mean zero. This is the unique solution to the penalized optimization problem.

Recall that γ_m consists of the c^* and d^* coefficients for β_m in the wavelet expansion, see (8). Hence, the estimated vector $\hat{\gamma}_m$ gives us an estimated coefficient function $\hat{\beta}_m$ through performing the inverse discrete wavelet transform.

In practice, we use the implementation of penalized multinomial regression from the R package `glmnet` (Friedman et al., 2010). The algorithm is based on cyclical coordinate descent methods and has LASSO penalization as a special case of elastic net penalization. The implementation in `glmnet` is very efficient.

3.5 Selection of tuning parameters

The computations above were for fixed tuning parameters, j_0 and λ, but they must be selected as part of the analysis. We use K-fold cross validation for that purpose. More precisely, we divide the index set $I = \{1,\ldots,n\}$ into K non-overlapping parts, $I = \cup_{k=1}^K I_k$, and consider the corresponding sub-datasets, $D_k = \{(x_i, y_i)\}_{i\in I_k}$. For fixed values of j_0 and λ and for sub-dataset k, we proceed as follows: First we use the data *not* included in D_k, i.e., $\{(x_i, y_i)\}_{i\in I\setminus I_k}$ to fit the regression model with tuning parameters j_0 and λ. This yields estimates $\hat{\alpha}_m^{(-k)}$ and $\hat{\beta}_m^{(-k)}$, $m \in \mathscr{M}$, and therefore also an estimated relation between a signal x and fitted probabilities, $\hat{p}_m^{(-k)}(x)$. Second, the deviance for the data in D_k, using the fitted regression without D_k, is computed as

$$\text{DEV}_k(j_0, \lambda) = -2\sum_{i\in I_k} \log p_{y_i}^{(-k)}(x_i),$$

where we have emphasized the dependence of the tuning parameters on the left hand side. This is repeated for all K sub-dataset, and the "mean cross-validated deviance" (MCVD) is computed as the average deviance per observation:

$$\text{MCVD}(j_0, \lambda) = \frac{1}{K}\sum_{k=1}^K \text{DEV}_k(j_0, \lambda). \tag{10}$$

The pair (j_0, λ) that makes MCVD the smallest is selected for further analysis.

In practice we profile over j_0: For fixed j_0 we minimize over λ; this is implemented directly in the `glmnet` package. We repeat this for the J possible values of j_0 and thereby find the optimal pair of tuning parameters.

3.6 Summary of estimation routine

In summary, estimation is carried out as follows. For each possible j_0:

1. Covariate functions x_i are expanded in the wavelet basis at detail level j_0.
2. The wavelet coefficients are used in as covariates in a multinomial regression with LASSO penalization. Cross validation yields an optimal value of λ and the corresponding cross validation deviance.

This is repeated for the possible values of j_0 as to minimize the cross validation deviance. The regression coefficients, $\hat{\alpha}_m$ and $\hat{\gamma}_m$, from the optimal multinomial regression are extracted and finally $\hat{\gamma}_m$ are translated into estimated coefficient functions, $\hat{\beta}_m$, using inverse DWT.

4 Detection of lameness: Multinomial functional regression

We are now ready to apply MFR to our data concerning horse lameness. Recall the data from Figure 1 consisting of 85 acceleration signals from five groups corresponding to lameness on the left-fore (LF), left-hind (LH), right-fore (RF) or right-hind (RH) limb, or healthy/normal (NO). The signal and group response correspond to x_i and y_i, respectively, so the set of possible values of y_i is $\mathscr{M} = \{\text{LF}, \text{LH}, \text{NO}, \text{RF}, \text{RH}\}$. Each signal is sampled at $N = 256$ equidistant time-points in $(0,1)$.

We are particularly interested in the ability of MFR to make predictions of Y for a new signal x, but we are also interested in the coefficient functions in order to understand the association between acceleration and lameness. Notice that the independence assumption (among all observations) is not quite reasonable in this application since the 85 signals come from only 8 horses, and observations from the same horse are likely to be dependent. We comment on this in Section 7.

Recall that each signal starts with stance on the RF/LH diagonal, so stance on the injured limb happens on the interval $(0, 0.5)$ for RF and LH signals and on the interval $(0.5, 1)$ for LF and RH signals. Due to symmetry of trot we would expect LF and RF signals to be similar except for an interchange of the two halves, and similarly for LH and RH signals. For NO signals we would expect the two halves to be similar. It is natural to require the regression model to resemble these symmetry properties.

In order to make this more precise, we introduce for any function $f : (0,1) \to \mathbb{R}$ its "twin function", that changes the order of f's restriction to $(0, 0.5)$ and f's restriction to $(0.5, 1)$:

$$\tilde{f}(t) = \begin{cases} f(t+0.5), & t \in (0, 0.5) \\ f(t-0.5), & t \in (0.5, 1) \end{cases}$$

Notice that $\tilde{f}$ is left undefined at 0.5.

We furthermore introduce the "symmetry group" $\tilde{m}$ for any $m \in \mathscr{M}$ as follows:

$$\widetilde{\mathrm{LF}} = \mathrm{RF}, \quad \widetilde{\mathrm{LH}} = \mathrm{RH}, \quad \widetilde{\mathrm{NO}} = \mathrm{NO}, \quad \widetilde{\mathrm{RF}} = \mathrm{LF}, \quad \widetilde{\mathrm{RH}} = \mathrm{LH}.$$

With this notation, the natural symmetry restrictions can be expressed as $p_m(x) = p_{\tilde{m}}(\tilde{x})$ for any signal x and any group $m \in \mathscr{M}$, and it is thus natural to assume

$$\alpha_{\mathrm{LF}} = \alpha_{\mathrm{RF}}, \quad \alpha_{\mathrm{LH}} = \alpha_{\mathrm{RH}} \tag{11}$$

$$\beta_{\mathrm{LF}} = \tilde{\beta}_{\mathrm{RF}}, \quad \beta_{\mathrm{LH}} = \tilde{\beta}_{\mathrm{RH}}, \quad \beta_{\mathrm{NO}} = \tilde{\beta}_{\mathrm{NO}}, \tag{12}$$

where, strictly speaking, the symmetry restrictions on the β's are on the set $(0, 0.5) \cup (0.5, 1)$.

We first run an analysis without these symmetry restrictions and examine to which extent the model by itself detects the symmetry and thus produces estimates that are in accordance with (11) and (12). In Section 4.5 we impose the restrictions in the estimation procedure.

4.1 Wavelet expansion of acceleration signals

The first step in the analysis is to compute the wavelet coefficients for each signal by DWT. The coefficients are stored in an $n \times N$ matrix which is subsequently used as a design matrix in the regression. To keep track of the order, coefficients are ordered with the $c_{j_0,k}$ coefficients (corresponding to the father wavelet) first and the $d_{j,k}$ coefficient (corresponding to the mother wavelet) last.

Figure 3 illustrates the wavelet decomposition at detail level $j_0 = 2$ for five signals from the same horse, one signal from each lameness group. The signals are shown in the lower right panel, the upper left panel shows the father wavelet coefficients $\{c_{2,k}\}_{k=0}^{3}$, and the remaining panels show the mother wavelet coefficients $\{d_{j,k}\}_{k=0}^{k=2^j-1}$ for $j = 2, 3, 4, 5$. The figure gives an interesting overview of the design matrix: The coefficients $c_{2,k}$ and $d_{j,k}$ for small j have larger amount on the half part with smaller amplitude compared to the part with larger amplitude, whereas coefficients $d_{j,k}$ for j large are associated with high frequency patterns of the signal.

4.2 Selection of tuning parameters (j_0 and λ)

We now address the selection of tuning parameters, in particular the choice of j_0. The left part of Figure 4 shows the MCVD criterion (10), i.e., the mean cross-validated deviance, as a function of λ for the eight possible values of j_0. We see that the minimum deviance (over λ) does not differ much for j_0 equal to 0, 1 and 2, but is larger for $j_0 \geq 3$. The right panel of the figure shows the estimated difference in coefficient function between the LF group and the normal group, i.e., $\hat{\beta}_{\mathrm{LF}} - \hat{\beta}_{\mathrm{NO}}$ for $j_0 = 0, \ldots, 4$. For each j_0 the optimal λ for that j_0 was used. We see that the estimated functions for the smaller values of j_0 do not differ much in smoothness, but the estimate corresponding to $j_0 = 4$ includes many more local features, and for larger values j_0 the estimated functions are even wilder (not shown). The pictures for the LH, RF and RH groups exhibit the same characteristics (not shown). When signals are classified into groups

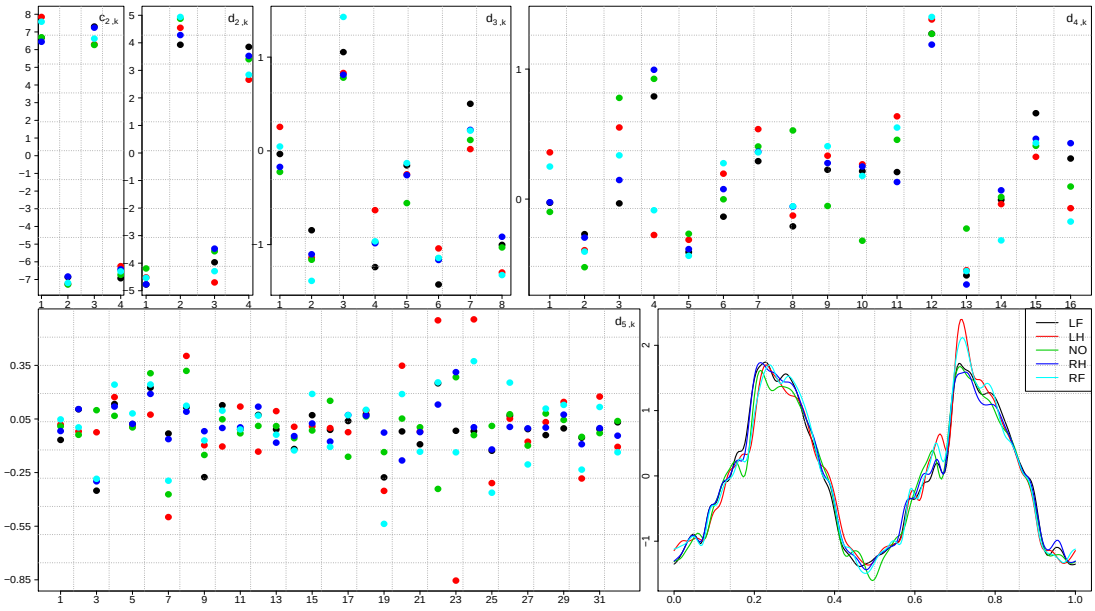

Figure 3: Coefficients in the wavelet decomposition at detail level $j_0 = 2$ of a signal from each group from the same horse (upper panels and lower left panel). The bottom right panel shows the signals.

according to the fitted MFR models for $j_0 = 0,\ldots,4$, we get 3, 3, 7, 8, and 10 misclassified signals, respectively, clearly indicating that a too large detail level leads to overfitting of the data.

Based on Figure 4, we use $j_0 = 0$ and the corresponding optimal LASSO tuning parameter, $\lambda = 10^{-5.27}$, for the analysis.

4.3 Estimates in the fitted model

The results below come from the MFR with $j_0 = 0$ and $\lambda = 10^{-5.27}$. There are 4–5 non-zero c^* and d^* coefficients per group in the fitted model. The left panel in Figure 5 shows the five estimated coefficient functions, $\hat{\beta}_m$. As in standard regression settings, the β's determine the effect of x_i as predictor on the distribution of the response Y_i, but in our case over intervals. In intervals where $\beta_m(t) \approx 0$, changes in $x_i(t)$ have no or little effect on the probability $P(Y_i = m)$, whereas x_i has a larger effect on the probability in intervals with $|\beta_m(t)|$ large. We notice that the coefficient function for the NO group has less fluctuations compared to the other groups.

In the right panel in Figure 5 we take NO as a reference group and consider for each $m \in \{\mathrm{LF},\mathrm{LH},\mathrm{RF},\mathrm{RH}\}$ the deviation $\delta_m = \beta_m - \beta_{\mathrm{NO}}$, that describes the effect of a signal x on the log-odds for group m compared to NO, cf. (3). We see that the largest deviation from zero occurs in the interval with stance on an injured limb. It makes good sense that this part of the

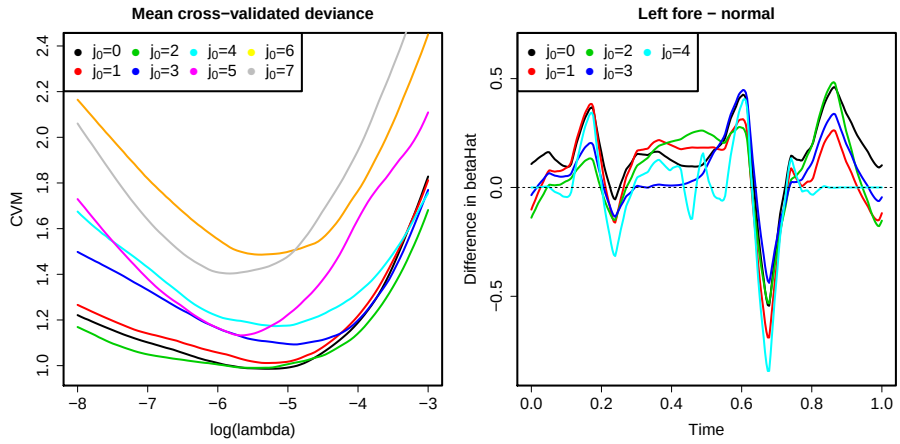

Figure 4: Sensitivity of wavelet tuning parameter, j_0. The left plot shows the MCVD criterion as a function of the LASSO tuning parameter λ for the possible values of j_0. The right part shows estimates of $\beta_{\text{LF}} - \beta_{\text{NO}}$ for $j_0 \leq 4$ and the corresponding optimal values of λ.

movement is used the most to distinguish horses with lameness from healthy horses. The sign of $\hat{\delta}_{\text{LF}}$ and $\hat{\delta}_{\text{RH}}$ are the same on most of $(0,1)$, which is natural as LF and RH come from the same diagonal. The same is true for $\hat{\delta}_{\text{RF}}$ and $\hat{\delta}_{\text{LH}}$. We also notice that deviations are numerically larger for LF and RF compared to LH and RH. This is not surprising for veterinarian experts since fore limb lameness is well-known to disturb the gait pattern more than hind-limb lameness. The intercept estimates, in the α parameterization as well as in the κ parameterization, see (4), are given in the following table:

	Group, m				
	LF	LH	NO	RF	RH
$\hat{\alpha}_m$	2.06	−1.05	−5.64	8.89	−4.27
$\hat{\kappa}_m$	−0.31	0.18	1.00	−0.45	−0.42

The estimates $\hat{\alpha}_m$ and $\hat{\beta}_m(t)$ are not automatically supplemented by standard errors or confidence intervals. They could be produced by resampling methods, though, and in Section 5 we conduct a simulation study where we, among others, point to the sampling variation in the estimates (see Figure 11).

Now, remember the natural symmetry restrictions (11) and (12), and consider first the estimated coefficient functions. The shape of $\hat{\beta}_{\text{NO}}$ over $(0, 0.5)$ and $(0.5, 1)$ is almost the same, and the estimates $\hat{\beta}_{\text{LH}}$ and $\hat{\beta}_{\text{RH}}$ are close to symmetric. The estimates $\hat{\beta}_{\text{LF}}$ and $\hat{\beta}_{\text{RF}}$ corresponding to fore limbs lameness have deep valleys around 0.7 and 0.2, respectively, very well in line with symmetry, however the valley is followed by a sharper peak for RF compared to LF. Altogether, the estimated coefficient functions obey the symmetry properties in (12) fairly well.

The estimates of α do not seem to fulfil restrictions (11) very well. In order to evaluate the total

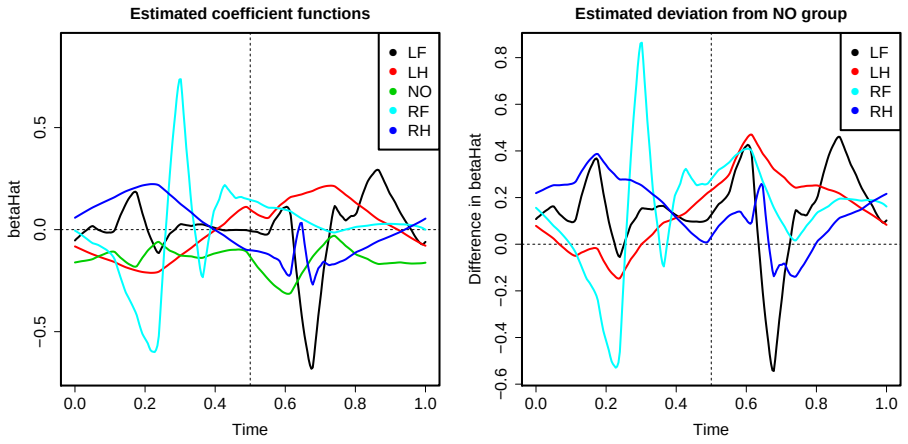

Figure 5: Results from the MFR with no symmetry restrictions on coefficient functions, and with $j_0 = 0$. Left: Estimates $\hat{\beta}_m$ for $m \in \mathscr{M}$. Right: Estimated deviations $\hat{\delta}_m = \hat{\beta}_m - \hat{\beta}_{\text{NO}}$ for $m \in \{\text{LF}, \text{LH}, \text{RF}, \text{RH}\}$.

symmetry features of the fitted model, we did the following: First we computed the predicted probability $\hat{p}_{y_i}(x_i)$ for the true (observed) group. Then we constructed the twin signal $\tilde{x}_i$ and computed the predicted probability $\hat{p}_{\tilde{y}_i}(\tilde{x}_i)$, i.e., the predicted probability that the twin signal comes from the symmetry group to the observed group. If the fitted model was completely symmetric in the sense of (11) and (12), then these two probabilities would coincide for all signals. We have therefore plotted the predicted probabilities against each other in Figure 6. The plot shows that there is a high degree of, albeit not complete, symmetry, as the points scatter around the line with intercept zero and slope one. Notice that the predicted probability is larger for the original data than for the twin data for the majority of the signals. This is hardly surprising as the model, except for penalization, was fitted as to maximize the predicted probabilities for the original data.

4.4 Leave-one-curve-out classification

In Section 4.2 we reported that only 3 of the 85 signals were misclassified when we used $j_0 = 0$ and the corresponding optimal λ. However, this success rate is overly optimistic as all curves were included as training as well as test data. In order to get a more realistic validation we now carry out a leave-one-curve-out procedure. This corresponds to K-fold cross validation with $K = n = 85$, see Section 3.5.

The results are reported in Table 1. We see that 68 signals are classified in the correct group, corresponding to a misclassification rate (MCR) of 20%. Seventeen signals are misclassified. Eleven of these are either NO signals that are classified in one of the lameness groups or signals that are wrongly classified in the NO group. Five signals are classified at the correct diagonal (but wrong limb); for example two RF signals are classified as LH signals. Only one signal is

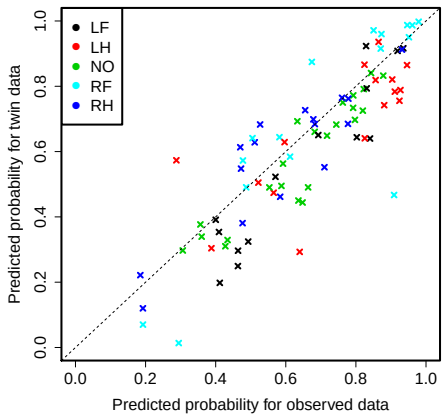

Figure 6: Predicted probabilities from the model fitted without symmetry restrictions. The probabilities $p_{y_i}(x_i)$ were computed on observed signals x_i and the true group, and $p_{\tilde{y}_i}(\tilde{x}_i)$ were computed on twin signals and symmetry groups.

Table 1: Results from the leave-one-out classification (no symmetry restrictions imposed).

	Predicted group				
True group	LF	LH	NO	RF	RH
LF	13	0	2	0	1
LH	1	12	2	1	0
NO	1	1	19	1	1
RF	0	2	1	13	0
RH	1	0	2	0	11

classified to a wrong diagonal.

Figure 7 shows the maximum predictive probability for each signal, i.e., the largest value among $\hat{p}_m(x_i)$, $m \in \mathscr{M}$. The points are organized after the true status and coloured according to the correctness of the classification. Green points correspond to signals for which the predicted group is the correct one, blue points correspond to signals for which the predicted group is wrong but the predicted diagonal is correct, and red points correspond to the remaining signals (mainly curves for which either the true or predicted status, but not both, is NO). The figure shows that misclassification mainly occurs when the maximum probability is relatively small, e.g. below 50%. In these cases, at least one other probability is not negligible, and it turns out that the correct group has the second largest probability for 11 of the 17 misclassified signals.

Recall that the 85 observations come from only eight different horses. It seems likely that the MCR from the leave-one-out analysis above underestimates the actual MCR as the training

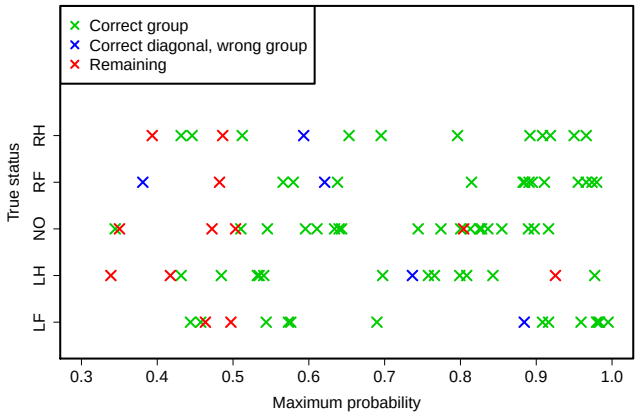

Figure 7: The largest probability among $\hat{p}_m(x_i)$ for $m \in \mathscr{M}$ in the leave-one-out study without symmetry restrictions for the 85 acceleration signals, coloured after the correctness of the prediction. The true status is shown on the y axis.

data for each test signal contains signals from the horse that generated the test signal. As supplement, we therefore ran a "leave-one-horse-out" analysis where, for each horse, we fitted the MFR model based on signals from the other horses, and used that model for classification. Then 63 of the 85 signals were classified correctly, corresponding to a MCR of 26%.

4.5 Incorporation of symmetry restrictions

In the above analysis the five coefficient functions and intercepts were allowed to vary freely with no constraints across the groups. We now incorporate the symmetry restrictions (11) and (12). The restrictions correspond to certain restrictions on the c^* and d^* coefficients in the wavelet expansions, but instead of implementing those restrictions directly, we consider an extended dataset, where each observation (x_i, y_i) is supplemented by an imaginary "twin observation" $(\tilde{x}_i, \tilde{y}_i)$ consisting of the twin signal of x_i and the symmetry group of y_i. The estimates from the MFR on the augmented data consisting of all (x_i, y_i) and $(\tilde{x}_i, \tilde{y}_i)$ will automatically satisfy the symmetry restrictions. Notice that, obviously, an observation and its twin are not independent.

The MFR on the augmented data has 3–6 non-zero c^* and d^* coefficients per group in the fitted model. Figure 8 shows the estimated coefficient functions (left) and the deviations from $\hat{\beta}_{\mathrm{NO}}$ (right). By construction, the behavior of $\hat{\beta}_{\mathrm{RF}}$ on $(0, 0.5)$ is exactly same as the behavior of $\hat{\beta}_{\mathrm{LF}}$ on $(0.5, 1)$, and vise versa. Similarly for $\hat{\beta}_{\mathrm{RH}}$ and $\hat{\beta}_{\mathrm{LH}}$, whereas $\hat{\beta}_{\mathrm{NO}}$ is identical on $(0, 0.5)$ and on $(0.5, 1)$. The intercept estimates are listed in the following table, in the α parameterization and in the κ parameterization:

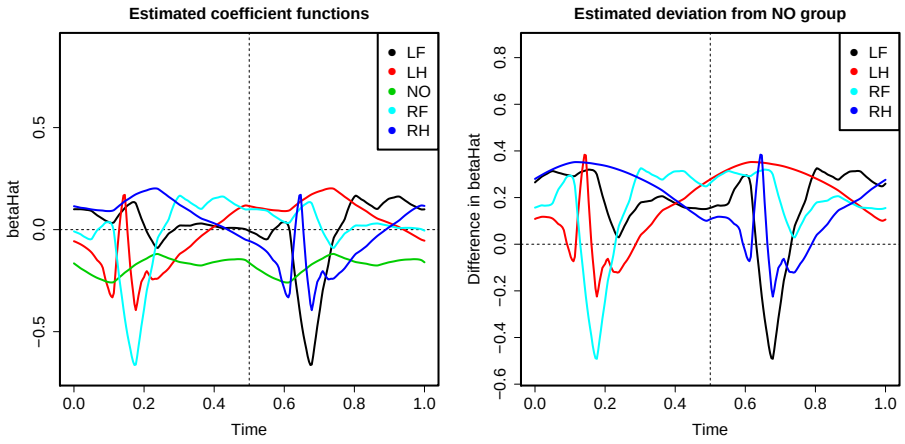

Figure 8: Results from the MFR on the augmented data (with symmetry restrictions imposed), and with $j_0 = 0$. Left: Estimates $\hat{\beta}_m$ for $m \in \mathscr{M}$. Right: Estimated deviations $\hat{\delta}_m = \hat{\beta}_m - \hat{\beta}_{\text{NO}}$ for $m \in \{\text{LF}, \text{LH}, \text{RF}, \text{RH}\}$.

	Group		
	LF, RF	LH, RH	NO
$\hat{\alpha}_m$	5.40	−2.03	−6.75
$\hat{\kappa}_m$	−0.36	−0.13	1.00

Table 2 show the results from a leave-one-out classification carried out on the augmented data. For each observed signal x_i ($i = 1, \ldots, 85$), we fitted the multinomial functional regression

Table 2: Results from the leave-one-out classification on the augmented data.

	Predicted group				
True group	LF	LH	RF	RH	NO
LF	13	0	2	0	1
LH	1	13	1	1	0
NO	1	1	18	2	1
RF	0	2	1	13	0
RH	1	0	2	0	11

model to the augmented data consisting of all data except (x_i, y_i) and $(\tilde{x}_i, \tilde{y}_i)$, and used the model fit to predict the outcome of y_i. The results are summarized in Table 2. Sixty-eight signals are classified correctly, corresponding to a MCR of 20%. The numbers in the table are close to those in Table 1 from the analysis on the original (not augmented data). A closer study reveals that the predicted group changes for six curves, all either curves from the NO group or curves that were predicted as NO curves in one of the analyses. A "leave-one-horse-out" gave a MCR of 32%, somewhat larger than the 26% from the similar analysis on the non-augmented data,

see Section 4.4. This is perhaps slightly surprising as we would have expected that asymmetry between the group estimates was due to random variation.

4.6 Comparison to other classification methods

This paper is, to the best our knowledge, the first study on MFR, so it is not possible to compare the estimates of the coefficients functions to estimates from other MFR approaches. There are, however, several classification methods for functional data in the literature, and we applied two of them to the lameness data: Linear discriminant analysis on principal component scores (PC-LDA) and curve discrimination (CurDis), cf. Ferraty and Vieu (2003).

The PC-LDA approach consists of a functional principal component analysis, see Section 5.1, followed by a standard LDA on the first few principal scores (PC scores), both carried out on the training data. Then the PC scores are computed for test signals and used as input to the LDA. We selected the number of principle components as to explain at least 95% of variation. This criterion gave us 13 principal components.

CurDis is a non-parametric method, where the posterior probability for a test curve x is computed as

$$\hat{p}_m(x) = \frac{\sum_{i=1}^n \mathbf{1}_{\{Y_i=m\}} K(h^{-1}d(X_i,x))}{\sum_{i=1}^n K(h^{-1}d(X_i,x))} \tag{13}$$

where K is a kernel function, h is the bandwidth, d is a distance measure between two functions, and we sum over the training sample. We used the kernel function $K(u) = (1-u^2)1_{[0,1]}(u)$ and the L^∞ distance in this study. The bandwidth was chosen as $h = 0.57$; then the denominator in (13) was non-zero for all signals.

We carried out leave-one-out validation for MFR (as already reported), PC-LDA, and CurDis. For each method we computed the overall misclassification rate (MCR) as well as the sensitivity (Sens) and specificity (Spec) for each group defined as follows:

$$\text{Sens(m)} = \text{Prob}\{\hat{Y} = m | Y = m\}, \quad \text{Spec(m)} = \text{Prob}\{\hat{Y} \neq m | Y \neq m\}.$$

The results are listed in Table 3. It shows that sensitivity and specificity are high for the MFR approach in all groups, whereas the other methods have relatively low values in several groups. Also, the rate of correctly classified signals ($1 - \text{MCR}$) is largest for MFR.

5 Simulation Study

In this section we conduct a simulation study to evaluate the variability of the estimated coefficient functions and the prediction results. We take the lameness data as starting point and generate datasets by means of eigenfunctions and eigenvalues estimated from these data. This simulation approach is similar to that of Swihart et al. (2014).

Table 3: Results from the leave-one-out analysis performed with three different methods. The MFR was fitted without symmetry restrictions.

	LF		LH		NO		RF		RH		
	Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec	$1-\text{MCR}$
PC-LDA	0.50	0.94	0.75	0.93	0.74	0.79	0.56	0.94	0.57	0.93	0.63
CurDis	0.56	0.97	0.19	0.94	1.00	0.76	0.75	0.91	0.64	0.96	0.66
MFR	0.81	0.96	0.75	0.96	0.83	0.89	0.81	0.97	0.77	0.97	0.80

5.1 Eigen-decomposition of lameness data

Suppose $X(t)$, $t \in [0,1]$ is a square-integrable random function with mean function $\mu(t)$ and covariance function $K(s,t)$, that is, $\mu(t) = E(X(t))$ and $K(s,t) = Cov(X(s),X(t))$. Mercer's theorem (Indritz, 1963) gives the spectral decomposition of the covariance function:

$$K(s,t) = \sum_{l=1}^{\infty} \lambda_l \nu_l(s) \nu_l(t) \tag{14}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots$ are the ordered non-negative eigenvalues, and the ν_l's are the corresponding orthonormal eigenfunctions (with respect to the L^2 norm). The ν_l's form a basis for the functional space $L^2(0,1)$, so the Karhunen-Loève (KL) expansion of the random function $X(t)$ is

$$X(t) = \mu(t) + \sum_{l=1}^{\infty} \xi_l \nu_l(t) \tag{15}$$

where the ξ_l's are uncorrelated random variables, and ξ_l has mean zero and variance λ_l. These random variables are called principle component scores (PC scores). For our application we need estimates of such an eigen-decomposition.

There are several approaches in the FDA literature to estimate the functional principle components (FPCs) from functional data observed with noise: 1) Smoothing of the raw signals followed by estimation of the FPCs; 2) Bi-variate smoothing of the raw covariance function followed by estimation of the FPCs; 3) Extraction of the FPCs from the raw covariance function followed by smoothing of the eigenfunctions. In what follows, we use the second approach to estimate the FPCs.

The first step is to compute smooth estimates of μ and K. A penalized cubic spline smoother is applied to the raw data mean to obtain a smooth $\hat{\mu}(t)$, and a fast bivariate penalized spline smoother (sandwich smoother) is applied to the raw covariance matrix (Xiao et al., 2013). In both cases, the smoothing parameter is selected via generalized cross validation (GCV). The curves are observed with noise, and the measurement error variance can be estimated as the average difference between the middle 60%, say, of diagonal elements of the raw and smoothed covariance (Goldsmith et al., 2013).

In the second step, the eigenfunction and eigenvalues are estimated from the smoothed covariance function. In practice the number of principle components must be chosen via some criterion. It is common to use the cumulative percent variance (CPV) criterion, that is, choose L components where $L = \min\{l \geq 1 : \sum_{j=1}^{l} \lambda_j / \sum_{j=1}^{N} \lambda_j \geq P\}$ for some predefined $P \in (0,1)$ representing the percentage of explained variance. Other possibilities include Akaike's information criterion (Yao et al., 2005) or cross validation (Rice and Silverman, 1991).

We applied the above procedure to the lameness data, separately for each lameness group. We used $L = 9$ as this was the smallest number such that the CPV criterion was at least 0.95 in all groups. Thereby we obtained five sets of eigenvalues and eigenfunctions, and five measurement error standard deviations:

$$\left\{\hat{\lambda}_l^{(m)}, \hat{v}_l^{(m)}\right\}_{l=1}^{L} \text{ and } \hat{\sigma}^{(m)}, \quad m \in \mathscr{M}.$$

The eigen-decompositions will form the basis for the simulations, as explained below.

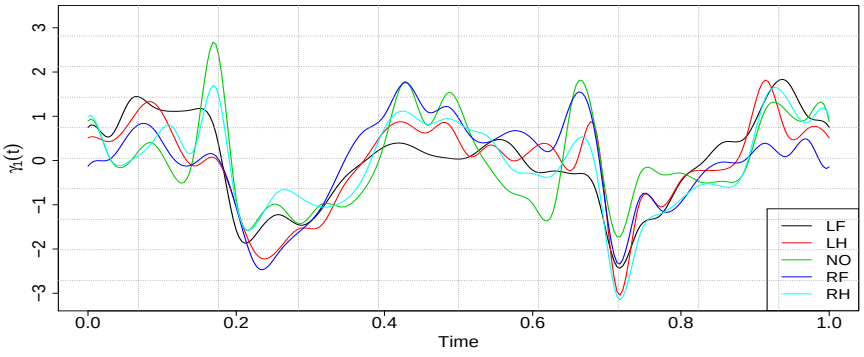

Figure 9: The first eigenfunctions estimated for each group separately.

For illustration, Figure 9 shows the first eigenfunctions extracted from the data. Generally, the first principle component reflects general variation in the amplitude of vertical acceleration and has the highest peaks and deepest valleys around 0.2 and 0.7. The percent of variation explained by the first eigenfunction varied between 40% (NO) and 48% (LF). The estimated standard deviations varied between 0.017 and 0.020 and were thus similar for the five groups.

5.2 Data generation and simulation scenarios

Our simulation study involves five groups corresponding to the lameness groups. The predictor curves were constructed using the following model:

$$X_i^{(m)}(t) = \hat{\mu}^{(m)}(t) + \sum_{l=1}^{L} \xi_{i,l}^{(m)} \hat{v}_l^{(m)}(t) + \varepsilon_i(t), \quad m \in \mathscr{M},\ i = 1,\ldots,n_m,\ t \in (0,1). \tag{16}$$

Here m represents the group, n_m is the number of simulated curves in group m, L is the truncation lag, $\hat{\mu}^{(m)}(t)$ is the smoothed group mean, and $\hat{v}_l^{(m)}$ is the estimated lth principal component

for group m, computed as described in Section 5.1. The PC scores $\xi_{i,l}^{(m)}$ were generated independently with $\xi_{i,l}^{(m)} \sim N(0, \lambda_l^{(m)})$. The measurement noise process was generated on an equally spaced grid of length N on $(0,1)$ with terms being independent and drawn from $N(0,\sigma^2)$ for various values of σ^2 (see below). We chose N to a power of 2, $N = 2^J$, since we need that for the estimation routine. Notice that the two first terms in (16), i.e. the mean and PC terms, are group-specific, whereas the distribution of the measurement noise is the same for all groups.

In the simulation study we varied the sample sizes (n_m), the number of observed points per curve (N), and the measurement error standard deviation (σ) as follows:

- **Sample sizes**: (a) small sample size, $n_m = 20$ and (b) large sample size, $n_m = 40$. Two datasets of this size are generated for each simulation; one is used as training data, the other as test data.
- **Number of observation per curve**: (a) sparse observations, $N = 2^5 = 32$ and (b) dense observations, $N = 2^8 = 256$.
- **Measurement error**: (a) without noise, $\sigma = 0$, (b) small noise, $\sigma = .05$, and (c) large noise, $\sigma = .25$.

This gives a total of 12 possible designs. Moreover, in order to compare directly with the analysis in Section 4, an additional scenario (scenario 1) was considered with sample sizes and sampling density as in the actual dataset, i.e., $n = (16, 16, 23, 16, 14)$ and $N = 256$, and no measurement error. An overview of the simulation scenarios is given in Table 4.

Table 4: Overview of the 13 simulation scenarios.

Scenario	1	2	3	4	5	6	7	8	9	10	11	12	13
n_m	As data	20	20	20	20	20	20	40	40	40	40	40	40
N	256	256	256	256	32	32	32	256	256	256	32	32	32
σ	0	0	.05	.25	0	.05	.25	0	.05	.25	0	.05	.25

A simulated dataset consists of a training dataset with n_m curves in group m, and a test dataset of the same size. The training and test data are independent and were drawn in exactly the same way. A simulated dataset from scenario 1 is shown to the right in Figure 10, with the original data to the left for comparison. The simulated data indeed seem to catch the same features and to exhibit the same degree of variability between curves as the observed data.

For each scenario we simulated 100 datasets. For each dataset we fitted the MFR model to the training data and used the fitted model to predict the groups for the training data as well for the test data. We ran all analyses with detail levels from 0 to 4, and with and without imposing the symmetry constraints as explained in Section 4.5. In any case, we evaluated the prediction ability by the misclassification rate (MCR). For scenario 1 we furthermore computed sensitivity and specificity for each group.

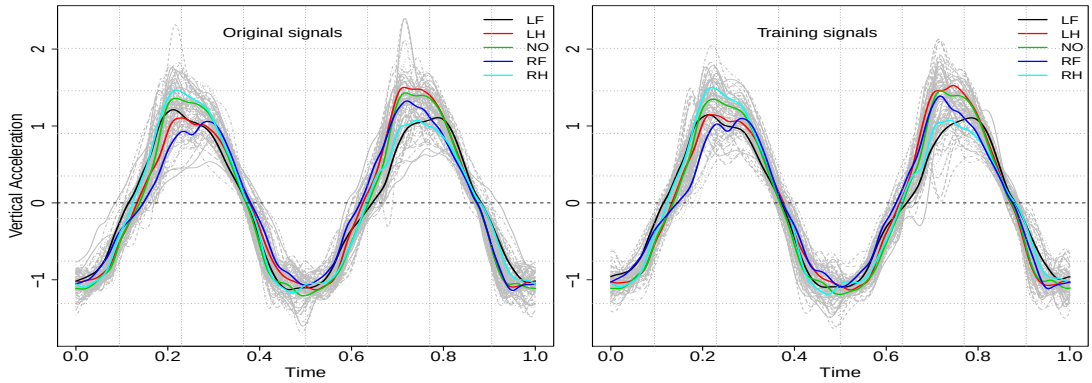

Figure 10: Left: The original curves along with group mean functions. Right: Training data from a simulated dataset in scenario 1.

5.3 Results

First, let us get a feeling for the variability in estimates of the coefficient functions. Figure 11 shows $\hat{\beta}_m$ from the 100 fitted MFR's from scenario 8 ($n_m = 40$, $N = 256$, $\sigma = 0$). The estimates are based on the training data and without symmetry restrictions on the coefficient functions. The realizations show the same patterns with peaks and valleys roughly at the same locations, and with $\hat{\beta}_{\text{NO}}$ being more flat than the coefficient functions for the other groups. The variability of $\hat{\beta}_m(t)$ is larger in intervals with larger numerical values of $\hat{\beta}_m(t)$. The peaks and valleys are generally larger on the part of the signal corresponding to stance on the injured limb, i.e., the part where the data signal has the smaller amplitude. For example, $\hat{\beta}_{\text{LF}}$ is numerically larger on (0.5,1) compared to (0,0.5). Although not imposed in the estimation process, the estimates for the hind limbs (LH and RH) are close to symmetric in the sense of (12). The symmetry is not as distinguished for the fore limbs (LF and RF).

Then, let us turn to the classification results. Average misclassification rates (MCR's) in percent are reported in Table 5 for training data and in Table 6 for test data. The classification results are better for the training data compared to the test data, just as we would expect. Perhaps more surprisingly, the classification results are quite robust to the choice of detail level.

The comparison of the different scenarios is better carried out by inspecting Figure 12 which shows boxplot of the MCR for test data in scenarios 2–13. The results are for detail level $j_0 = 0$. Hence, the graphs correspond to the first and fifth columns with results in Table 6. Except for the combination of sparse data and large measurement noise (scenarios 7 and 13), all average misclassification rates are below 20%. For sparse data and large measurement noise classification results are bad with 32–62% incorrectly classified curves.

We get the expected results when we compare the scenarios: (a) A larger training sample improves classification; (b) smaller measurement errors, or no measurement noise at all, improves classification; (c) dense observation of the functional predictors improves classification, at least in the presence of measurement noise. Except for the scenarios with sparse data and large mea-

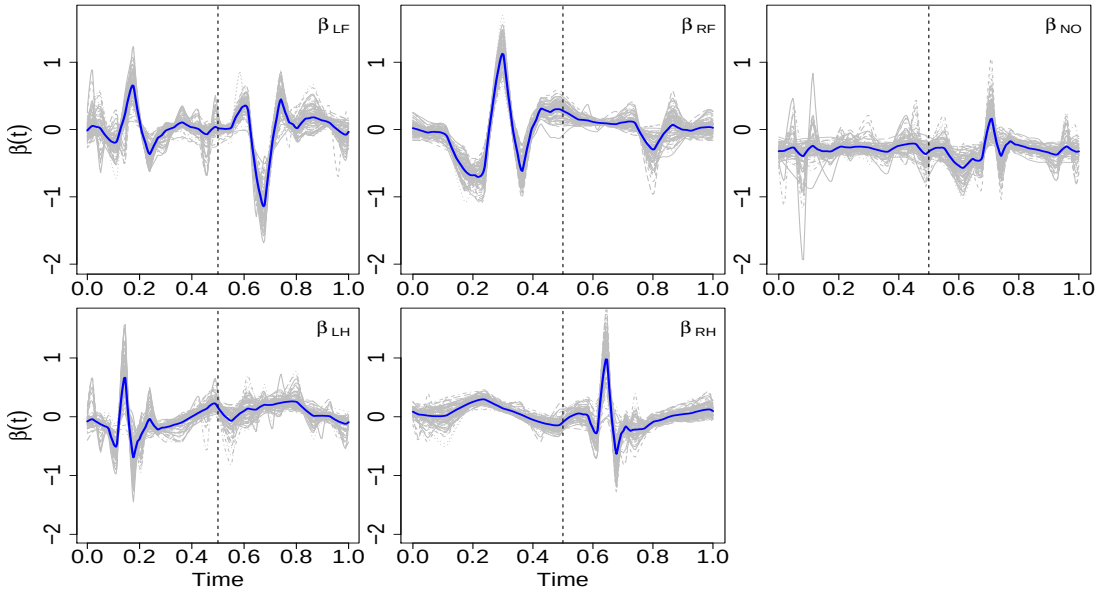

Figure 11: Estimated coefficient functions, $\hat{\beta}(t)$, for the five groups for 100 simulations. In all plots, the mean of the estimated functions is highlighted in blue.

Table 5: Average misclassification rate (over 100 simulated datasets) in percent for training data.

	Simulated Data				**Twins of Simulated Data**			
	Detail Level (j_0)				**Detail Level (j_0)**			
Scenario	**0**	**1**	**2**	**3**	**0**	**1**	**2**	**3**
1	2.95	3.11	4.16	4.74	3.66	3.75	4.25	3.98
2	2.53	2.51	3.26	3.26	3.32	3.27	3.66	3.77
3	2.51	2.21	2.79	3.30	3.35	3.52	3.90	3.88
4	7.22	7.27	7.80	7.98	8.01	8.62	9.00	9.33
5	2.24	2.35	2.69	3.26	2.13	2.25	2.21	2.80
6	4.00	3.68	4.26	4.84	4.04	3.67	3.92	4.58
7	22.39	21.87	22.09	21.16	50.93	50.52	51.03	49.94
8	2.01	1.98	2.17	2.51	1.44	1.23	1.22	1.00
9	2.21	2.29	2.29	2.75	1.84	1.87	1.71	1.62
11	0.94	0.64	0.56	0.92	1.27	1.16	1.09	1.26
10	6.89	6.83	7.55	7.34	8.60	8.88	9.35	9.11
12	2.48	2.54	2.35	3.04	4.04	3.85	3.83	4.24
13	23.79	23.70	23.79	23.88	50.22	50.30	50.37	50.40

surement error, introduction of twin data improves classification slightly.

For comparison we also applied the PC-LDA and CurDis methods (see Section 4.6) to sim-

Table 6: Average misclassification rate (over 100 simulated datasets) in percent for test data.

	Simulated Data				**Twins of Simulated Data**			
	Detail Level (j_0)				**Detail Level (j_0)**			
Scenario	**0**	**1**	**2**	**3**	**0**	**1**	**2**	**3**
1	11.48	12.21	12.20	13.82	9.75	10.00	10.54	10.75
2	9.94	10.55	10.21	11.59	9.17	9.35	9.94	10.06
3	10.73	11.10	11.00	11.84	9.44	9.60	10.32	10.37
4	18.55	18.75	18.35	19.32	16.50	16.97	17.51	17.89
5	9.84	10.30	10.76	10.54	8.38	8.38	8.52	9.12
6	12.71	13.25	13.51	13.06	11.73	11.85	11.89	12.39
7	37.45	37.91	37.83	36.00	61.19	61.24	60.84	61.07
8	6.24	6.43	6.58	7.21	5.66	5.50	5.51	5.27
9	6.62	6.74	6.83	7.32	6.20	5.95	5.87	5.72
10	14.06	14.04	14.16	14.53	13.60	13.73	14.03	14.04
11	5.40	5.37	5.03	5.99	5.19	5.07	5.16	5.49
12	8.44	8.61	8.56	8.98	8.62	8.53	8.55	8.90
13	32.76	32.73	32.62	32.20	58.34	58.49	58.62	58.58

ulated data. Table 7 shows the average sensitivity and specificity in each group and the rate

Table 7: Average sensitivity and specificity and rate of correctly classified curves (all in percent) for three classification methods. The numbers are based on 100 simulated datasets from scenario 1.

	Group										
	LF		LH		NO		RF		RH		
	Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec	1−MCR
PC-LDA	83	99	78	96	91	89	89	97	86	99	86
CurDis	67	99	53	98	98	81	81	96	78	97	77
MFR	84	98	83	97	93	94	90	98	91	99	89

of correctly classified curves ($1-\text{MCR}$) over 100 simulated datsets from scenario 1 (n_m as in lameness data, $N = 256$, $\sigma = 0$). MFR has the largest rate of correctly classified curves, and overall also the best performance with respect to sensitivity and specificity. We should mention here that the denominator of (13) was zero for some signals, and that these signals were discarded in the CurDis analysis. For scenarios with sparse data this occured to many/most data, regardless of the bandwith (results not shown).

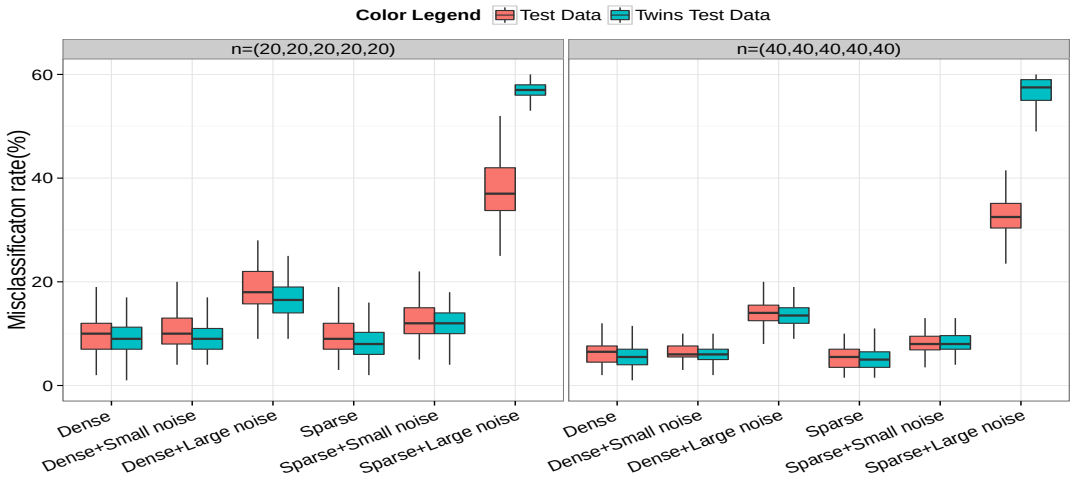

Figure 12: Boxplots of misclassification rates for test data in scenarios 2–13 with $j_0 = 0$ (with and without twin data).

6 Application to phoneme data

In this section we apply the MFR approach to a dataset which has been widely used in the signal classification literature as well as for research in speech recognition. The data originally come from the TIMIT database described in Hastie et al. (1995) and are available in the `ElemStatLearn` package in R (Halvorsen, 2012).

The data are log-periodograms corresponding to recording continuous speech of 50 male speakers. There are a total of 4509 speech frames from five different phonemes transcribed as follows: "sh" as in *she*, "dcl" as in *dark*, "iy" as the vowel in *she*, "aa" as the vowel in *dark*, and "ao" as the first vowel in *water*. The distribution of the speech frames on phonemes are as follows:

sh: 872, iy: 1163, dcl: 757, ao: 1022, aa: 695.

Each speech frame has a duration of 32 milliseconds and is sampled at 16 kHz, and thus originally consists of 512 observations. We use the data from Halvorsen (2012) with 4509 log-periodogram of length 256. Figure 13 displays a random sample of 5 log-periodograms for phoneme.

The interesting question is whether it is possible to predict the phoneme from a given log-periodogram. Several classification approaches have been discussed (Ferraty and Vieu, 2003; Hastie et al., 1995). Hastie et al. (2009, Ch. 4) discussed logistic regression for two classes, namely, "aa" and "ao". We will apply MFR on data with all five phonemes, and investigate the classification properties as well as the behaviour of the coefficients functions. Results from Ferraty and Vieu (2003) will be used for comparison.

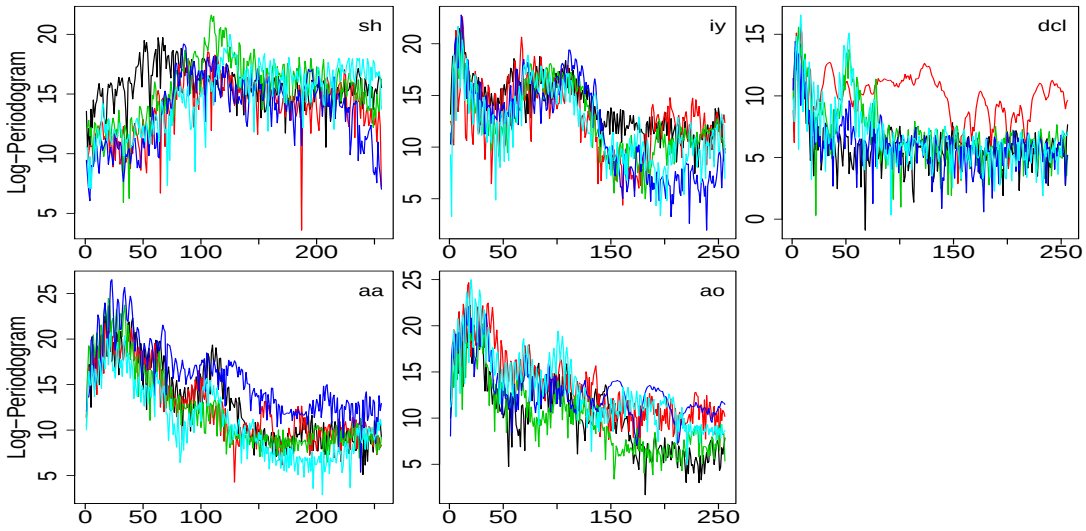

Figure 13: A sample of 5 log-periodograms per category.

We divided the data randomly into training data (40%) and test data (60%), fitted the MFR model with $j_0 = 0$ (the best detail level) to the training data, and used the model to predict the phoneme for the test data. This was repeated 50 times. The mean and standard deviation for the MCR over the 50 samples were 0.072 and 0.0034, respectively.

The estimated coefficients functions are shown in Figure 14 along with the mean functions highlighted in blue. The plot reveals that most information to predict the phoneme is in the first half of the signals. We therefore repeated the analysis using only the first 128 observations from each signal. This gave a mean and standard deviation of MCR of 0.073 and 0.0031, respectively, over 50 samples. Hence, the prediction is just as good when we use the fist half of the frame as when we use the complete frame.

Ferraty and Vieu (2003) used a subset of the phoneme data, available from the webpage of Ferraty and Vieu (2006) and consisting of 2000 log-periodograms of length 150 (still with with known phoneme class). They reported classification results from the following methods:

- PDA/Ridge : Penalized discriminant analysis (Hastie et al., 1995)
- MPLSR : Multivariate partial least-square regression (Martens, 1989)
- NPCD/PCA: Non-parametric curve discriminant based on FPCA (Hall et al., 2001)
- NPCD/MPLSR : NPCD including the MPLSR method in its semi-metric (Ferraty and Vieu, 2003).

In order to compare the MFR approach directly to those methods we ran the MFR classification on the smaller dataset. Keep in mind that the length of the data for each signal in this case is 150.

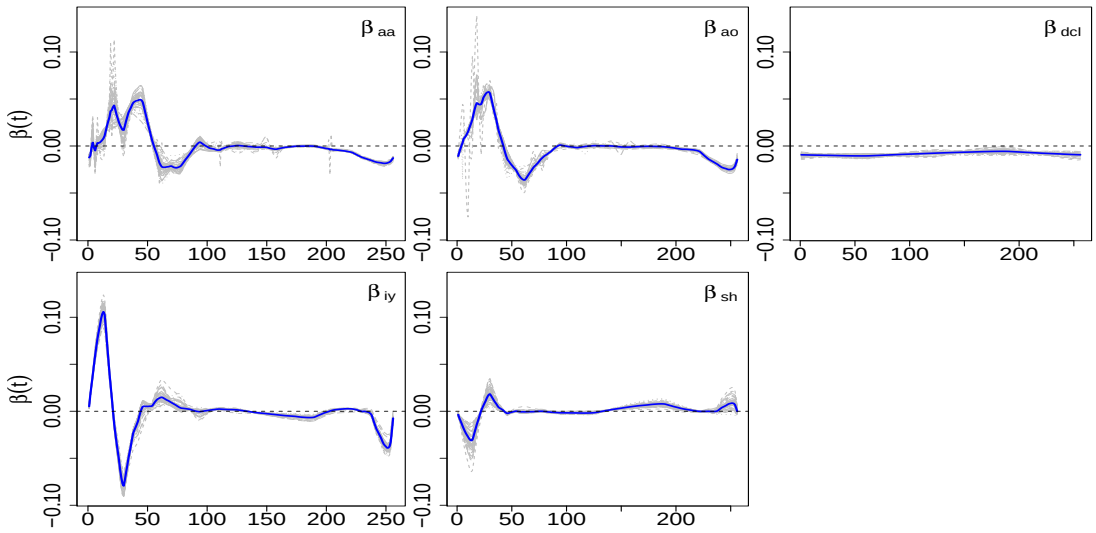

Figure 14: Estimates of coefficients functions and the corresponding mean functions highlighted in blue. Results come from 50 samples of test data.

In order to obtain signals of length equal to a power of 2, we used a spline basis of piecewise linear function with a knot at each interior time point of the data and evaluated the spline at a vector with length $2^7 = 128$.

The median of MCR, again over 50 test samples, are shown in Table 8 for all methods. The values from the above-mentioned approaches are taken directly from Figure 4 in Ferraty and Vieu (2003). The results show that for this example, the MFR approach performs the best and improves the test error rate to 7.3%.

Table 8: Median of misclassification rate over 50 sample test data.

	MPLSR	NPCD/MPLSR	NPCD/PCA	PDA/Ridge	MFR
MCR	0.095	0.084	0.104	0.082	0.073

7 Conclusion and discussion

In this paper, we combined the discrete wavelet decomposition and LASSO penalization to fit a regression model with functional predictor and multinomial response. More specifically, the wavelet decomposition was computed for each signal at an appropriate resolution level, and

the wavelet coefficients were used as the predictors in a multinomial regression with LASSO penalization on the regression coefficients to deal with the curse of dimensionality. The same approach has been used for regression with scalar response and functional predictors (Zhao et al., 2012) and regression with scalar or binary outcome and image predictors (Reiss et al., 2015), but to the best of our knowledge, this is the first work about multinomial functional regression (MFR). We applied the MFR approach to a famous dataset from speech recognition and to a dataset regarding detection of lameness among horses. In both cases, classification based on the MFR method gave good results compared to other classification methods for functional data.

We chose the combination of wavelets and LASSO because the LASSO forces most coefficients to zero while wavelets are known to offer precise approximations with sparse representations. However, other basis systems such as B-splines, ramp functions or harmonic functions could be used in combination with an appropriate penalty function.

In our application we used the raw wavelet coefficients from the DWT. However, functional data are often observed with noise, and one could remove noise by attempting to identify the wavelet coefficients associated to the noise and then modify these coefficients by hard or soft thresholding (Donoho, 1995).

If one would like to retain more information and not miss potentially interesting differences, then one could apply appropriate high- and low-pass filters to the data at each level to produce two sequences at the next level. In this transform, no decimation occurs and so the two sequences have the same length as the original sequence. This transform is known as non-decimated wavelet transform.

In the lameness application the 85 signals were treated as independent although they came from only eight different horses. This is not quite appropriate and we would like to include a horse effect in the model. It would be easy to include horse as a fixed effect in the multinomial regression, but it would be more appropriate to include it as random. A similar model would be appropriate for the phoneme data as the 4509 data signals come from only 50 different speakers. Most likely, inclusion of random effects would not change the prediction notably, but it could make a change for the variability of estimated coefficicients and thereby for inference about the coefficient functions.

In summary, our MFR approach turned out to give good classification results in the two applications, but there are still modifications that could be interesting to pursue.

Acknowledgements

We thank Maj Halling Thomsen (Department of Large Animal Sciences, University of Copenhagen) for letting us use her data on lameness.

References

Cardot, H., Ferraty, F., and Sarda, P. (1999). Functional linear model. *Statistics & Probability Letters*, 45(1):11–22.

Cardot, H., Ferraty, F., and Sarda, P. (2003). Spline estimators for the functional linear model. *Statistica Sinica*, 13:571–591.

Cardot, H. and Sarda, P. (2005). Estimation in generalized linear model for functional data via penalized likelihood. *Journal of Multivariate Analysis*, 92:24–41.

Crainiceanu, C. M., Staicu, A.-M., and Di, C.-Z. (2009). Generalized multilevel functional regression. *Journal of the American Statistical Association*, 104:1550–1561.

Cuevas, A., Febrero, M., and Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, 22:481–496.

Daubechies, I. et al. (1992). *Ten Lectures on Wavelets*, volume 61. SIAM.

Delaigle, A. and Hall, P. (2012). Achieving near perfect classification for functional data. *Journal of the Royal Statistical Society (Series B)*, 74:267–286.

Donoho, D. L. (1995). De-noising by soft-thresholding. *Information Theory, IEEE Transactions on*, 41(3):613–627.

Ferraty, F. and Vieu, P. (2003). Curves discrimination: a nonparametric functional approach. *Computational Statistics & Data Analysis*, 44:161–173.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Goldsmith, J., Bobb, J., Crainiceanu, C. M., Caffo, B., and Reich, D. (2011). Penalized functional regression. *Journal of Computational and Graphical Statistics*, 20:830–851.

Goldsmith, J., Greven, S., and Crainiceanu, C. (2013). Corrected confidence bands for functional data using principal components. *Biometrics*, 69(1):41–51.

Hall, P., Poskitt, D., and Presnell, B. (2001). A functional data-analytic approach to signal discrimination. *Technometrics*, 43:1–9.

Halvorsen, K. (2012). *ElemStatLearn: Data sets, functions and examples from the book: "The Elements of Statistical Learning, Data Mining, Inference, and Prediction" by Trevor Hastie, Robert Tibshirani and Jerome Friedman.* R package.

Hastie, T., Buja, A., and Tibshirani, R. (1995). Penalized discriminant analysis. *The Annals of Statistics*, pages 73–102.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., and Tibshirani, R. (2009). *The Elements of Statistical Learning*. Springer.

Indritz, J. (1963). *Methods in Analysis*. Macmillan New York:.

James, G. and Hastie, T. (2001). Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society (Series B)*, 63:533–550.

James, G. M. (2002). Generalized linear models with functional predictors. *Journal Of The Royal Statistical Society (Series B)*, 64:411–432.

James, G. M., Wang, J., and Zhu, J. (2009). Functional linear regression that's interpretable. *Annals of Statistics*, 37:2083–2108.

Keegan, K., Dent, E., Wilson, D., Janicek, J., Kramer, J., Lacarrubba, A., Walsh, D., Cassells, M., Esther, T., Schiltz, P., Frees, K., Wilhite, C., Clark, J., Pollit, C., Shaw, R., and Norris, T. (2010). Repeatability of subjective evaluation of lameness in horses. *Equine Veterinary Journal*, 42:92–97.

Keegan, K., Wilson, D., Wilson, D., Smith, B., Gaughan, E., Pleasant, R., Lillich, J., Kramer, J., Howard, R., Bacon-Miller, C., Davis, E., May, K., Cheramie, H., Valentino, W., and van Harreveld, P. (1998). Evaluation of mild lameness in horses trotting on a treadmill by clinicians and interns or residents and correlation of their assessments with kinematic gait analysis. *American Journal of Veterinary Research*, 59:1370–1377.

Lee, E. R. and Park, B. U. (2012). Sparse estimation in functional linear regression. *Journal of Multivariate Analysis*, 105(1):1–17.

López-Pintado, S. and Romo, J. (2006). Depth-based classification for functional data. In Liu, R. Y., Serfling, R., and Souvaine, D. L., editors, *Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications*, volume 72 of *DIMACS: Series in Discrete Mathematics and Theoretical Computer Science*, pages 103–120. AMS.

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 11(7):674–693.

Martens, H. (1989). *Multivariate Calibration*. John Wiley & Sons.

Marx, B. D. and Eilers, P. H. C. (1999). Generalized linear regression on sampled signals and curves: a P-spline approach. *Technometrics*, 41:1–13.

Meyer, Y. (1995). *Wavelets and Operators: Volume 1*. Cambridge University Press.

Müller, H.-G. and Stadtmüller, U. (2005). Generalized functional linear models. *The Annals of Statistics*, 33:774–805.

Nason, G. (2010). *Wavelet Methods in Statistics with R*. Springer.

Pfau, T., Witte, T., and Wilson, A. (2005). A method for deriving displacement data during cyclical movement using an inertial sensor. *The Journal of Experimental Biology*, 208:2503–2514.

Preda, C., Saporta, G., and Lévéder, C. (2007). PLS classification of functional data. *Computational Statistics*, 22:223–235.

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

Randolph, T. W., Harezlak, J., and Feng, Z. (2012). Structured penalties for functional linear models-partially empirical eigenvectors for regression. *Electronic journal of statistics*, 6:323.

Reiss, P. and Ogden, R. (2007). Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association*, 102:984–996.

Reiss, P. T., Huo, L., Zhao, Y., Kelly, C., and Ogden, R. T. (2015). Wavelet-domain regression and predictive inference in psychiatric neuroimaging. *Annals of Statistics*, page to appear.

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society (Series B)*, pages 233–243.

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

Swihart, B. J., Goldsmith, J., and Crainiceanu, C. M. (2014). Restricted likelihood ratio tests for functional effects in the functional linear model. *Technometrics*, 56(4):483–493.

Thomsen, M., Jensen, A., Sørensen, H., Lindegaard, C., and Andersen, P. (2010). Symmetry indices based on accelerometric data in trotting horses. *Journal of Biomechanics*, 43:2608–2612.

Thomsen, M. H. (2010). *Objective assessment of lameness in horses: Development and evaluation of symmetry indices based on a single tri-axial accelerometer*. PhD thesis, Univsersity of Copenhagen.

Tian, T. S. and James, G. M. (2013). Interpretable dimension reduction for classifying functional data. *Computational Statistics and Data Analysis*, 57:282–296.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, pages 267–288.

Vidakovic, B. (2009). *Statistical Modeling by Wavelets*. John Wiley & Sons.

Weishaupt, M., Wiestner, T., Hogg, H., Jordan, P., and Auer, J. (2004). Compensatory load redistribution of horses with induced weightbearing hindlimb lameness trotting on a treadmill. *Equine Veterinary Journal*, 36:727–733.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (series B)*, 73:3–36.

Xiao, L., Li, Y., and Ruppert, D. (2013). Fast bivariate p-splines: the sandwich smoother. *Journal of the Royal Statistical Society (Series B)*, 75(3):577–599.

Yao, F., Müller, H.-G., Wang, J.-L., et al. (2005). Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, 33(6):2873–2903.

Zhao, Y., Ogden, R. T., and Reiss, P. T. (2012). Wavelet-based LASSO in functional linear regression. *Journal of Computational and Graphical Statistics*, 21:600–617.

A Preprocessing

We now describe the proprocessing steps transforming the raw data (not shown) to the signals shown in Figure 1. Each acceleration signal was collected over a distance of at least 24 metres, and the number of gait cycles as well as the length of the discrete data vary between horses in the raw data. The purpose of the preprocessing is to make data signals comparable and remove noise. The preprocessing takes a raw data signal on discrete form and produces a function on the unit interval, which is interpreted as the average acceleration over one gait cycle and which has time zero in the suspension phase right before stance at the RF/LH signal. In the process we remove variation between gait cycles due to slightly varying velocity and random pertubations of the movement pattern as well as variation due to uncertain selection of start and end point of the part used for analysis. Notice that preprocessing was carried out separately for each acceleration signal without using any information at all from the other signals.

The details are described in detail below and illustrated in Figure 15 for a signal from a horse with lameness induced on the right-fore leg. The process consists of the following steps:

1. The raw data signal was filtered to remove micro structure noise using a Butterworth filter with cut-off frequency 50 Hz.

2. Based on video recordings, data from eight gait cycles was selected such that the first top corresponds to stance phase on the RF/LH diagonal. The number of observations varies across signals from 1120 to 1440 with an average of 1269. An example signal is shown in the top of Figure 15. Notice that we use gait cycles as time unit, such that the time domain is $(0, 8)$.

3. The signal was smoothed with a very large b-spline basis (500 basis elements). The data was thus converted to functional form, but almost not smoothed, and the functional version is indistinguishable from the discrete data in Figure 15. A non-periodic basis was chosen since the start and end of the eight gait signal were chosen with uncertainty and therefore not corresponds to exactly the same place in the movement.

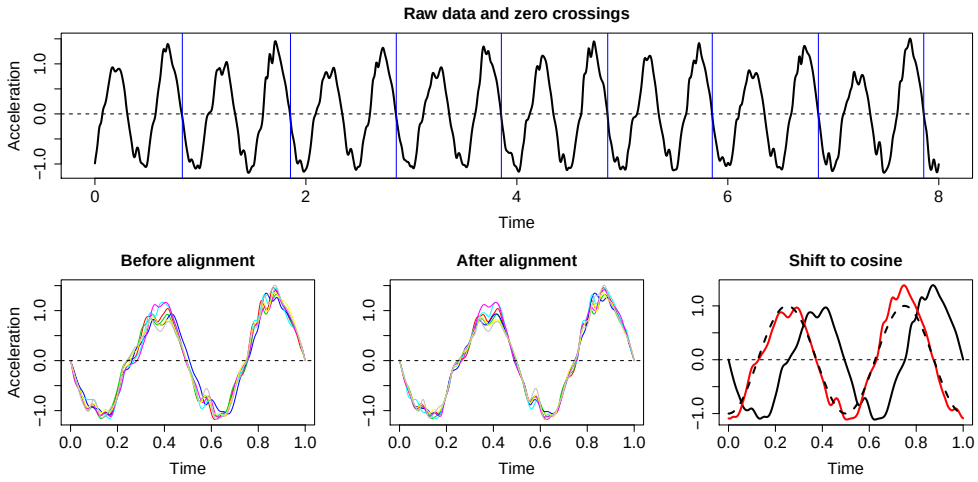

Figure 15: Preprocessing for a signal from a horse with lameness on the right-fore limb. Top panel: Data from eight gait cycles before and after smoothing to a b-spline basis (indistinguishable, black curve) and the time points for zero crossing used to partition the signals into gait cycles (blue lines). Bottom, left and middle: Data from seven gait cycles before and after alignment. Bottom right: Average of aligned sub-signals (black, solid), the function $g(t) = -\sin(2\pi t)$ (black, dashed), and the average shift aligned to g (red).

4. The functional version of the signal was cut into nine pieces by identifying the zero crossing after stance on the LF/RH diagonal. Zero crossing was chosen as feature since it was always well-identified (unique), as opposed to minimum or maximum, say. The zero crossings are illustrated by the vertical blue lines in Figure 15.

5. The first and last piece do not consist of complete cycles and were discarded. The remaining seven pieces were considered as replications and for technical reasons expressed in a large Fourier basis (101 basis elements for each part). Time was rescaled to $(0,1)$ for each of the sub-signals. The seven replications are shown on top of each other in the lower left part of Figure 15.

6. As it often happens with functional data, the seven gait cycles are slightly misaligned: Sub-signals have the same features but they occur at slightly different time points. Such phase variation may blur analysis of the amplitudes, and we therefore used continuous registration (Ramsay and Silverman, 2005, Chapter 7) to align the sub-signals. The aligned sub-signals are shown in the lower, middle panel in Figure 15.

7. The pointwise average over the seven alignes sub-signals was computed. It is shown as the solid, black curve in the lower right panel of Figure 15.

8. By construction, time zero for the average signal corresponds to zero-crossing after stance on the LF/RH diagonal. This does not correspond to a well-defined physiological feature of the gait pattern and is therefore not comparable from signal to signal.

Minimum acceleration, on the other hand, occurs during the suspension phase (if any). We therefore shift aligned the average signal to a shifted sine curve, $g(t) = -\sin(2\pi t)$. This curve is shown as dashed black in the lower right panel Figure 15, and the shift aligned average as the red curve. The red curve is the end product of our preprocessing steps.

Altogether the preprocessing steps transforms the original signal to real-valued functions x_i defined on $(0,1)$. The 85 functions were shown in Figure 1. In practice, for the multinomial regression, we evaluated the function in $N = 256$ time points from 0 to 1. Importantly, the first top always corresponds to stance on the RF/LH diagonal.

B Introduction to wavelets

Wavelet theory is about representing arbitrary functions in terms of simpler, fixed building blocks at different scales and positions. This has been found to be useful in many areas. In the wavelet literature, there are two common ways to introduce wavelets: one is through the continuous wavelet transform, the other is through multiresolution analysis. Of course, there are connections between the two.

Multiresolution analysis (MRA) provides a natural framework for the understanding of wavelet bases and was initiated by Mallat (1989) and Meyer (1995). MRA also provides a framework for examining functions at different scales. A multiresolution analysis of $L^2(\mathbb{R})$ is defined by a sequence of closed subspaces V_j of $L^2(\mathbb{R})$, $j \in \mathbb{Z}$, with the following properties:

1. $V_j \subset V_{j+1}$
2. $v(t) \in V_j \Leftrightarrow v(2t) \in V_{j+1}$ and $v(t) \in V_0 \Leftrightarrow v(t-k) \in V_0$, $k \in \mathbb{Z}$
3. $\bigcup_{j=-\infty}^{+\infty} V_j$ is dense in $L^2(\mathbb{R})$ and $\bigcap_{j=-\infty}^{+\infty} V_j = \{0\}$
4. A *scaling function*, $\phi \in V_0$ with a non-vanishing integral exists such that the sequence $\{\phi(t-k),\ k \in \mathbb{Z}\}$ is an orthonormal basis of V_0

The MRA definition implies that $\{\phi_{j,k} = 2^{j/2}\phi(2^j t - k),\ k \in \mathbb{Z}\}$ is an orthonormal basis of V_j. Hence, the function $\phi(t) \in V_0 \subset V_1$ can be represented as a linear combination of functions of the $\phi_{1,k}$ functions, i.e., $\phi(t) = \sum_{k\in\mathbb{Z}} h_k \sqrt{2}\phi(2t-k)$ for some coefficients h_k, $k \in \mathbb{Z}$, which are often referred to as the low-pass filter. The scaling function ϕ is also called the *father wavelet*.

For each MRA, a *mother wavelet*, $\psi(t)$, can be defined to explain the detail at each level, i.e., the set of $L^2(\mathbb{R})$ functions that are elements of V_{j+1} but not V_j. Consider the detail space W_j to be the orthogonal complement of the space V_j in V_{j+1}, and denote it $V_j^{\perp}$. Then $V_{j+1} = V_j \oplus W_j$. In this situation, $\{\psi_{j,k} = 2^{j/2}\psi(2^j t - k),\ k \in \mathbb{Z}\}$ forms an orthonormal basis for W_j. Also, as $\psi(t) \in V_1$, it can be represented as a linear combination of the functions from V_1, i.e., $\psi(t) = \sum_{k\in\mathbb{Z}} g_k \sqrt{2}\phi(2t-k)$, where the set of the coefficients $\mathscr{G} = \{g_k\}_{k\in\mathbb{Z}}$ are the high-pass filter coefficients associated with the particular wavelet function being used.

In summary, a wavelet transform has a mother wavelet, $\psi(t)$, and a father wavelet, $\phi(t)$, that are linked by the relationship, $\psi(t) = \sum_{k\in\mathbb{Z}} g_k\sqrt{2}\phi(2t-k)$. Consider for instance the Haar wavelet functions,

$$\phi(t) = \begin{cases} 1 & x \in [0,1] \\ 0 & \text{otherwise} \end{cases}, \quad g_k = \{g_0 = 1/\sqrt{2}\ ,\ g_1 = -1/\sqrt{2}\ ,\ g_k = 0\ ,\ k \geq 2\}$$

$$\psi(t) = \sum_{k\in\mathbb{Z}} g_k\sqrt{2}\phi(2t-k) = \phi(2t) - \phi(2t-1) = \begin{cases} 1 & x \in [0,1/2) \\ -1 & x \in [1/2,1) \\ 0 & \text{otherwise} \end{cases}$$

The number of detail levels is determined by the number of entries in the series. The length of the series for a discrete wavelet transform (DWT) must be a power of 2. The DWT proposed by Mallat (1989) is an efficient algorithm to calculate the wavelet coefficients of a discrete series. The idea of DWT is to filter the series using the high- and low-pass filter associated with the wavelet.

Daubechies et al. (1992) introduced two families of compactly supported wavelets with different degree of smoothness: the *extremal phase* wavelets and the *least asymmetric* wavelets. As these wavelets have compact support, the associated high- and low-pass filters $\mathscr{G}$ and $\mathscr{H}$ have a finite number of coefficients. For a pre-set detail level, j_0, a fine-scale representation of a function at detail level $j = j_0$ is given by

$$f_{j_0}(t) = \sum_{k=0}^{2^{j_0}-1} c_{j_0,k}\phi_{j_0,k}(t) + \sum_{j=j_0}^{J-1}\sum_{k=0}^{2^j-1} d_{j,k}\psi_{j,k}(t).$$

The coefficients in the fine-scale representation are computed using the pyramid algorithm (Mallat, 1989) which computes the transform in $O(N)$ calculations. Understanding the concept and idea behind the computation of the wavelet helps us to understand the operation of the wavelet transform in depth. The coefficients $\{c_{j,k}\}$ are known as the smooth coefficients, and they are used to represent global features of the series or function $f(t)$.

The formula to generate the coefficients of the father wavelet is $c_{j-1,k} = \sum_l h_{l-2k}c_{j,l}$, where the low-pass coefficients $\mathscr{H} = \{h_k\}_{k\in\mathbb{Z}}$ satisfy the condition $\sum_k h_k^2 = 1$. Conversely, the mother wavelet coefficients, $\{d_{j,k}\}$, are known as the detail coefficients and the local features in the sequence or the function. These coefficients represent the difference between the global representation and the true function or sequence. The coefficients $\{d_{j,k}\}$ are computed by $d_{j-1,k} = \sum_l g_{l-2k}c_{j,l}$.

The formulas for computing the coefficients $\{c_{j,k}\}$ and $\{d_{j,k}\}$ are quite similar but involve the low-pass and high-pass coefficients, respectively. The low-pass and high-pass wavelet coefficients are linked by the relationship, $g_{L-1-k} = (-1)^k h_k$, where L is the length of the filter. We note that by introducing the negative signs into the filter, the detail coefficients will be returned. Also, the coefficients $\{c_{j,k}\}$ are used again to compute the father and mother wavelet coefficients for the next decomposition level.

Apparently, the correlation between the detail coefficients is low for the Haar wavelet, and this may results in a poor performance in discriminant analysis. Therefore a longer filter where

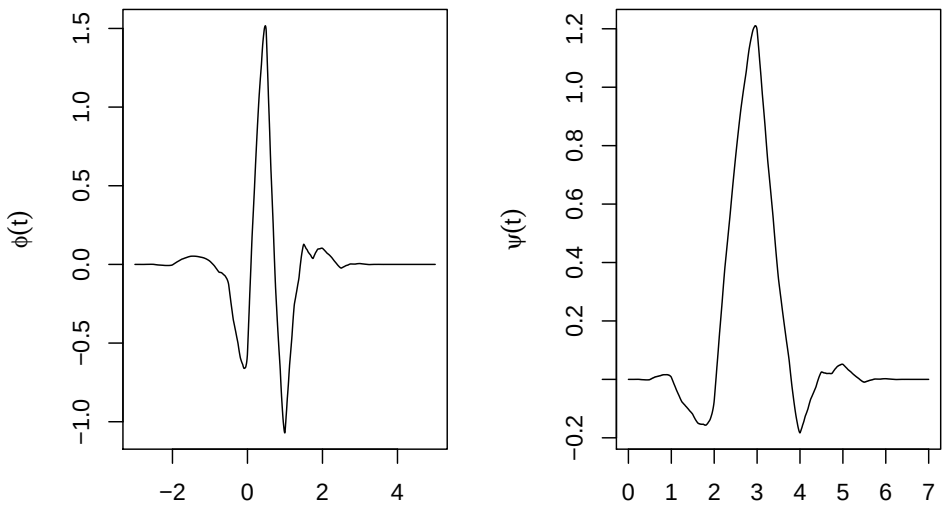

Figure 16: The Daubechies least asymmetric wavelet with filter number 4. Mother wavelet is shown to the left and father wavelet to the right.

the correlations are more stable, may improve the performance. For this reason, we used the family of Daubechies' least asymmetric wavelets in this paper, i.e., the *Daubechies-n* wavelets, where the so-called filter number n signifies the number of remarkable non-zero coefficients h_k. More specifically we used the Daubechies-4 least asymmetric wavelets. Figure 16 shows the Daubechies-4 least asymmetric scaling function or father wavelet $\phi(t)$ to the left and the mother wavelet $\psi(t)$ to the right.

There are at least three packages in R for wavelet analysis: `wavelets`, `wavethresh`, and `waveslim`. The second one is a comprehensive package that performs 1D, 2D and 3D real and complex-valued wavelet transforms, non-decimated transforms, wavelet packet transforms, non-decimated wavelet packet transforms, multiple wavelet transforms, wavelet shrinkage for various kinds of data, locally stationary wavelet time series, non-stationary multiscale transfer function modeling, and density estimation. We used this package for our work.

II

Functional logistic regression: A comparison of three methods

Seyed Nourollah Mousavi
Department of Mathematical Sciences
University of Copenhagen

Helle Sørensen
Department of Mathematical Sciences
University of Copenhagen

Publication details

Ready to submit (2015).

Functional logistic regression: A comparison of three methods

Seyed Nourollah Mousavi
Department of Mathematical Sciences
University of Copenhagen, Denmark
nourollah@math.ku.dk

Helle Sørensen
Department of Mathematical Sciences
University of Copenhagen, Denmark
helle@math.ku.dk

Abstract

Functional logistic regression is becoming more popular as there are many situations where we are interested in the relation between functional covariates (as input) and a binary response (as output). Several approaches have been advocated, and this paper goes into detail about three of them: dimension reduction via functional principle component analysis, penalized functional regression, and wavelet expansions in combination with LASSO penalization. We discuss the performance of the three methods on simulated data and also apply the methods to data regarding lameness detection for horses. Emphasis is on classification performance, but we also discuss estimation of the unknown parameter function.

Key words: Functional logistic regression; Discrete wavelet transform; LASSO penalization; Functional principle component analysis; Supervised classification; Penalized functional regression; Lameness data for horses.

1 Introduction

Functional data consist of curves or images, and occur more and more often and in many different scientific fields, e.g. medicine, economics, biology, chemistry. Functional data are observed discretely but are generated from underlying random functions. Hence, the data points per curve (or image) are highly dependent, and it is expedient to use approaches that take this dependency into account when analyzing the data. This has led to the sub-field of statistics called functional data analysis (FDA). Many existing strategies in FDA handle the high-dimensionality with basis expansions, either using fixed or data-driven bases. Textbooks on FDA include Ramsay and Silverman (2005), Ferraty and Vieu (2006), and Horváth and Kokoszka (2012).

If there are only few observations per curve, then the data are naturally considered and analyzed as longitudinal data (Diggle et al., 2002), either with a parametric model for the development over time or with time as a categorical variable (if the observation times are identical across subjects), and random subject effects. With more sampling points per curve, functional approaches become more in compliance with the data structure. There are both similarities and

differences between FDA and longitudinal data analysis; see for example Rice (2004) for a comparison of FDA and longitudinal data analysis from a smoothing perspective.

An increasing number of studies in FDA is concerned with the relationship between one or more functional covariates and an outcome variable which can be scalar, binary, or categorical. For scalar outcome there is a plethora of literature, see Ramsay and Dalzell (1991), Cardot et al. (1999), Yao et al. (2005), James et al. (2009), Goldsmith et al. (2011b), and Zhao et al. (2012), among others. The primary aims of these studies were to explain the variation in the data and predict future values of the outcome using the information from the functional covariates.

There are fewer studies related to binary outcomes, and the methods have not been compared thoroughly in the literature. Ratcliffe et al. (2002) suggested a logistic regression analysis for foetal heart rate data. They used Fourier expansions for the functional covariates and the parameter function and a modified fisher-scoring algorithm for computation of the maximum likelihood estimate. Several papers used functional principal component analysis (PCA) followed by standard logistic regression with the principal component scores as entries in the design matrix (Aguilera et al., 2006, 2008; Wei et al., 2014). The high-dimensional multicollinearity in the functional covariates is thus overcome by the dimension reduction obtained by PCA. The main difference between the papers lies in the approach for the PCA. In Section 3.4, we will pay special attention to logistic regression in combination with the so-called PACE technique (Yao et al., 2005) where estimation of the scores is based on conditional expectations. This method was also used by Wei et al. (2014) to gene detection in a case-control study of pancreatic cancer. Goldsmith et al. (2011b) used penalized functional regression for analysis of white-matter tract profiles in multiple sclerosis (the division of participants into patients and controls constitutes the binary response). They combined a functional PCA for the functional covariate and a B-spline expansion for the parameter function, including a difference penalty for regularization. We will go into details about this method in Section 3.5. Aguilera et al. (2011) introduced a penalized spline approach to functional principal logistic regression in order to obtain a smooth and reasonable interpretable estimate of the parameter function. Mousavi and Sørensen (2015) proposed to use a combination of wavelets and LASSO penalization for multinomial functional setting with application on detecting of horse lameness. In Section 3.3 we use a binomial version of this method.

In this paper we describe, compare and discuss three of the above-mentioned methods, namely, functional logistic regression using functional principle component, penalized functional regression, and functional logistic regression based on wavelets and LASSO penalization. The competing methods are used to analyze a dataset regarding detection of lameness of horse, and they are also tested in a simulation study. Our aim is to compare the three approaches with primary focus on the methods' ability to correctly classify new observations. In addition, we will discuss the quality of the estimates of the parameter function.

The remainder of the paper is organized as follows. In Section 2, we review the standard logistic regression model. Functional logistic regression and the methods of interest are explained in Section 3 with focus on estimation of the parameter function. Bootstrap for functional data is discussed in Section 4. Simulation studies are discussed in Section 5, and Section 6 contains an analysis of data on lameness of horses. Final remarks are provided in Section 7.

2 Logistic regression

There are many situations where we are interested in the relation between input variables and a binary output. For classification, for example, we need a rule that takes the input variables and delivers a guess of the output. Random variation implies that there is no "perfect rule", and it is thus expedient to seek a stochastic model which is able to account for noise. Indeed, we are interested in finding the conditional distribution of the binary response given the input variables. In other words, if Y is the outcome and X is the collection of covariates, we wish to model the conditional probability $P(Y=1|X)$ as a function of X, describing the effect of X on the outcome distribution. If we use the maximum likelihood method for estimation, then we have the usual asymptotic results available for statistical inference.

To be specific, let Y_i and $X_i = (X_{i1}, X_{i2}, \cdots, X_{ip}) \in \mathbb{R}^p$ be the random Bernoulli variable and the vector of p explanatory variables for the ith individual, respectively, and assume that Y_i is one with probability $\pi_i = \pi(X_i)$ and zero with probability $1-\pi_i = 1-\pi(X_i)$. The logistic regression model is

$$\pi_i = P(Y_i = 1|X_i) = \frac{\exp\{\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}\}}{1+\exp\{\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}\}}, \quad i = 1,2,\ldots,n, \tag{1}$$

where $\beta_0, \beta_1, \cdots, \beta_p$ are the parameters to be estimated. Equivalently, the use of the logit transformation enables us to write the model in terms of a linear relation as follows:

$$\texttt{logit}(\pi_i) = \log\Big(\frac{\pi_i}{1-\pi_i}\Big) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}, \quad i = 1,2,\ldots,n. \tag{2}$$

Let $l_i = \texttt{logit}(\pi_i)$ be the linear predictor for individual i such that $l_i = \log(\frac{\pi_i}{1-\pi_i}) = \alpha + \sum_{j=1}^{p} X_{ij}\beta_j$ for $i = 1,2,\ldots,n$. The parameter β_j is associated to the jth explanatory and is interpreted as the additive change in the linear predictor when the jth explanatory variable increases one unit and the rest remain constant. The matrix form of Eq. (2) can be written as $L = \alpha\mathbf{1} + \boldsymbol{X}\boldsymbol{\beta}$ where $L = (l_1, l_2, \ldots, l_n)^T$ is the vector of linear predictors, $\mathbf{1} = (1,1,\ldots,1)^T$ is the n-vector of ones, $\boldsymbol{X}$ is the $n \times p$ matrix where each row represents the covariates for one individual, and $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is the vector of parameters.

To find the maximum likelihood estimates of the parameters, we would differentiate the log likelihood with respect to the parameters, set the partial derivatives to zero, and solve the equations. The achieved equations are transcendental equations, and there is no closed form solution, so it is necessary to use a numerical method to obtain an estimate. There are several methods for numerical optimization such as Newton-Raphson and Fisher scoring, among others.

3 Functional logistic regression

To better understand the setting for functional logistic regression, we first describe functional linear regression and the common approaches that are used to estimate the parameters.

3.1 Functional linear models

Functional regression attempts to model and estimate an association between X and Y when one or both variables have functional form, and is therefore a natural extension of classical regression. Suppose that the predictor X is a square-integrable random function on a compact interval $T \subset \mathbb{R}$, Z is a vector of scalar covariates, and $Y \in \mathbb{R}$ is a scalar response. Then the most common functional linear regression model is

$$Y = \alpha_0 + Z\alpha + \int_T X(t)\beta(t)\,dt + \varepsilon. \tag{3}$$

Here α_0 is a scalar intercept parameter, $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_q)^T$ is the vector of coefficients for the scalar covariates, the coefficient function β is a square-integrable function on T, and the error term ε is $N(0, \sigma^2)$ random variable. Often, for simplicity, T will be scaled to interval $[0, 1]$. As in standard linear regression, β describes the relationship between X and Y, now on the the compact interval, T. The region where $|\beta(t)|$ is larger indicates that changes in $X(t)$ on this interval has greater predictive power on Y.

One central goal in functional linear regression is to find an appropriate estimate of the regression coefficient function $\beta(t)$ and make inference for $\beta(t)$. For example, we are interested in confidence regions for $\beta(t)$ and in testing whether the functional predictor $X(t)$ has influence on the response. For a start, it is natural to look for estimates of $\beta(t)$, α_0, and α that minimize the sum of squared residuals, i.e. minimize

$$\sum_{i=1}^{n} \left(Y_i - \alpha_0 - Z_i^T \alpha - \int_T \beta(t) X_i(t)\,dt \right)^2$$

but without restrictions on $\beta(t)$ there may be infinitely many solutions.

In practice, the functional predictors X are observed at N discrete points, and quite often N is much larger than n, the number of curves. The main issue is that the function $\beta(t)$ is an infinite-dimensional object which must be estimated from an finite sample, and the problem thus calls for regularization or dimension reduction. We notice that the classical multivariate regression approaches do not necessarily produce a meaningful estimator of the coefficient function in functional case: If we apply dimension reduction methods from multivariate data analysis such as Principal Component Regression (PCR), Partial Least Square (PLS), or Sufficient Dimension Reduction (SDR) directly to the discrete data, the functional nature of the data will be ignored.

A standard approach, that maintains the functional nature of $\beta(t)$, is to represent $\beta(t)$ with basis functions that (hopefully) can approximate $\beta(t)$ well. The basis can either be a fixed basis (not constructed from the data), for example a Fourier basis, a B-spline basis, a wavelet basis, or a ramp basis, or it can be a basis constructed by the eigenfunctions of the covariance operator. The eigenfunctions can be estimated from the data as described in Rice and Silverman (1991), Capra and Müller (1997), and Goldsmith et al. (2013). Functional regression that use eigenfunctions to expand the coefficient function $\beta(t)$ is known as functional principle component regression (FPCR) and has been described in Cardot et al. (1999); Müller and Yao (2008) and Delaigle

et al. (2009). In these papers, the basis functions used to expand the functional covariates $X_i(t)$ and the coefficient function $\beta(t)$ are the same, and it is implicitly assumed that $\beta(t)$ and X_i have similar smoothness properties. A slightly different approach consists of using fixed basis functions to expand $\beta(t)$ and $X_i(t)$, where the type and/or the number of basis functions can be different for $X_i(t)$ and $\beta(t)$. For instance, Marx and Eilers (1999) and Cardot et al. (2003) used spline bases for expansion and added a penalty term to the log-likelihood function, Goldsmith et al. (2011a) and Wood (2011) proposed a spline-series expansion in a mixed-model setting, and Zhao et al. (2012) suggested to use wavelet bases for the functional covariates $X_i(t)$ as well as the parameter function $\beta(t)$ along with LASSO penalization.

In general, the basis functions should be chosen to reflect the characteristics of the signals, for example, the Fourier basis is appropriate to model periodic functions. In other situations B-spline and Wavelet basis are more appropriate, and also have the advantage that the basis functions have finite support. It is often desirable that the expansion is "economical", interpreted as sparsity of the coefficients. This means that just a few non-zero coefficient give a good approximation of the function. For example, wavelet bases are known to be able to represent functions, even functions with discontinuities, accurately and parsimoniously with few terms.

Now, assume that we have selected $\psi(t)$ and $\theta(t)$ as bases for expansion of the coefficient function and the sampling trajectory, respectively. We write

$$\psi(t) = (\psi_1(t), \cdots, \psi_{K_x}(t))^T \quad \text{and} \quad \theta(t) = (\theta_1(t), \cdots, \theta_{K_\beta}(t))^T$$

for the bases. The trajectories and the parameter function are thus be expressed as

$$X_i(t) = \sum_{k=1}^{K_x} c_{ik}\psi_k(t) = \mathbf{c}_i^T \psi(t) \tag{4}$$

and

$$\beta(t) = \sum_{\ell=1}^{K_\beta} b_\ell \phi_\ell(t) = \phi^T(t)\mathbf{b}. \tag{5}$$

Then model (3) for ith curve can be expressed as

$$\begin{aligned} Y_i &= \alpha_0 + Z_i^T\alpha + \int_T X_i(t)\beta(t)\,dt + \varepsilon_i = \alpha_0 + Z_i^T\alpha + \sum_{k=1}^{K_x}\sum_{\ell=1}^{K_\beta} c_{ik}\{\int_T \phi_\ell(t)\psi_k(t)\,dt\}b_\ell + \varepsilon_i \\ &= \alpha_0 + Z_i^T\alpha + \mathbf{c}_i^T \int_T \psi(t)\theta^T(t)\,dt\ \mathbf{b} + \varepsilon_i = \alpha_0 + Z_i^T\alpha + \mathbf{c}_i^T\mathbf{J}_{\psi\phi}\mathbf{b} + \varepsilon_i, \end{aligned} \tag{6}$$

where

$$\mathbf{J}_{\psi\phi} = \int_T \psi(t)\phi^T(t)\,dt. \tag{7}$$

By defining $\zeta = (\alpha_0, \alpha_1, \cdots \alpha_q, b_1, b_2, \cdots, b_{K_\beta})^T$ and $\mathbf{W} = [\mathbf{1}\ \mathbf{Z}\ \mathbf{C}\mathbf{J}_{\psi\phi}]$, model (6) simply becomes

$$\mathbf{Y} = \mathbf{W}\zeta + \boldsymbol{\epsilon} \tag{8}$$

and so a least squares estimate of ζ solves the equation

$$\mathbf{W}^T\mathbf{W}\hat{\zeta} = \mathbf{W}^T\mathbf{Y}. \tag{9}$$

If $K_\beta + q + 1 > K_x$, the least squares solution "overfit" and therefore, regularization is in general needed. A convenient method of regularization is to truncate the basis by choosing a value K_β such that then fit ζ by least square, but the problem may still suffer from multi-collinearity. For a given K_x, increasing K_β increases roughness of the estimate of $\beta(t)$ quite dramatically. We often want to control this roughness, suggesting a small number of basis functions for $\beta(t)$. On the other hand, we also want to capture the complexity of $\beta(t)$, suggesting a large K_β. The standard solution to this problem is to adopt a regularization approach with a penalty on roughness of coefficients functions in the model fitting process, thus imposing smoothness of the estimate.

More specifically, we define a penalized residual sum of squares

$$\mathbf{PENSSE}_\lambda(\alpha_0, \alpha, \beta) = \sum_{i=1}^{n} \left(Y_i - \alpha_0 - Z_i^T\alpha - \int_T X_i(t)\beta(t)\,dt \right)^2 + \lambda \int_T (L\beta(t))^2\,dt \tag{10}$$

where L is a linear differential operator. The specific choice of L depends on the nature of the data. Two common operators that have been used in functional data literature are the following: (1) Curvature or departure from linearity, $L(\beta)(t) = D^2\beta(t)$, and (2) the harmonic acceleration operator, $L(\beta)(t) = (2\pi/a)^2 D\beta(t) + D^3\beta(t)$ which is used for periodic functions with period a. The smoothing parameter λ can be chosen either subjectively or by an automatic method such as cross-validation. A small value of penalty parameter leads to a rough estimate of $\beta(t)$, and a large value of λ results in a smooth (ultimately, flat) estimate of $\beta(t)$.

Basis expansions and penalization go well together: Suppose that the covariate functions $X_i(t)$ are expanded by K_x terms relative to basis functions ψ_k and that the regression function $\beta(t)$ is expanded by K_β terms relative to basis functions θ_ℓ as in (4) and (5), respectively. For the penalty term, we observe that $L\beta(t) = \sum_{\ell=1}^{K_\beta} b_\ell \{L\theta_\ell(t)\}$ and hence we can write

$$\int_T \{L\beta(t)\}^2\,dt = \sum_{\ell=1}^{K_\beta}\sum_{\ell'=1}^{K_\beta} b_\ell \left(\int_T \{L\theta_\ell(t)\}\{L\theta_{\ell'}(t)\}\,dt \right) b_{\ell'} = \mathbf{b}^T\mathbf{R}\mathbf{b}$$

where $\mathbf{R}$ is an $K_\beta \times K_\beta$ matrix with its (ℓ, ℓ')-th element equal to the parenthesis in the expression just above.

Therefore the penalized residual sum of squares can be written

$$\mathbf{PENSSE}_\lambda(\alpha_0, \alpha, \beta) = ||\mathbf{Y} - \alpha_0 - \mathbf{Z}^T\alpha - \mathbf{C}\mathbf{J}_{\psi\phi}\mathbf{b}||^2 + \lambda\mathbf{b}^T\mathbf{R}\mathbf{b} \tag{11}$$

where $\mathbf{J}_{\psi\theta}$ was defined in (7). Given a fixed value of λ, there exists a closed form solution for the parameters α_0, α and $\mathbf{b}$. Define $\zeta = (\alpha_0, \alpha, \mathbf{b}^T)^T$ and $\mathbf{W}$ as the $n \times (1 + q + K_b)$ coefficient matrix $[\mathbf{1}\ \ \mathbf{Z}\ \ \mathbf{C}\mathbf{J}_{\psi\theta}]$. Also define $\mathbf{R}_0$ from $\mathbf{R}$ as

$$\mathbf{R}_0 = \begin{pmatrix} \mathbf{0}_{(q+1)\times(q+1)} & \mathbf{0}_{(q+1)\times K_\beta} \\ \mathbf{0}_{K_\beta\times(q+1)} & \mathbf{R}_{K_\beta\times K_\beta} \end{pmatrix} \tag{12}$$

Then the vector $\hat{\zeta}$ that minimizes $\mathbf{PENSSE}_\lambda$, satisfies

$$(\mathbf{W}^T\mathbf{W}+\lambda\mathbf{R}_0)\hat{\zeta}=\mathbf{W}^T\mathbf{Y}.$$

In summary, the major point debated here is the need for regularization for estimation of the coefficient function $\beta(\cdot)$ to ensure existence and control smoothness. This can be done using either restricted basis function or roughness penalty.

3.2 Functional logit regression

We now modify the functional linear regression to the case of a binary response variable. For each of n individuals, the observed response Y_i is assumed to come from a Bernoulli distribution with success probability π_i, i.e., $\pi_i=\pi(X_i(t))=Pr\{Y_i=1|X_i(t)\}$. If we use the logit link function, then the natural extension of model (3) is

$$\text{logit}(\pi_i)=\log\left(\frac{\pi_i}{1-\pi_i}\right)=\alpha_0+\mathbf{Z_i}^T\alpha+\int X_i(t)\beta(t)\,dt. \tag{13}$$

Just as before, α_0 is the intercept parameter, $\alpha=[\alpha_1,\alpha_2,\cdots,\alpha_q]^T$ is the vector of coefficients for scalar covariates $\mathbf{Z_i}=[Z_{i1}\ Z_{i2}\ \cdots Z_{iq}]^T$ and $\beta(t)$ is the coefficient function for the functional covariate $X_i(t)$. As in the functional linear model case, it is impossible to obtain the estimate of parameter function $\beta(t)$ without further assumptions or restrictions on $X_i(t)$ and $\beta(t)$ (Ramsay and Silverman, 2005).

Therefore, we consider expansions of $X_i(t)$ and $\beta(t)$:

$$\beta(t)=\phi^T(t)\mathbf{b}\ \ ,\ \ \mathbf{X}(t)=\mathbf{C}\psi(t).$$

Substituting for $\beta(t)$ and $\mathbf{X}(t)$, the regression model (13) becomes

$$\log\left(\frac{\pi}{1-\pi}\right)=\alpha_0\mathbf{1}+\mathbf{Z}\alpha+\int\mathbf{C}\psi(t)\phi^T(t)\mathbf{b}\,dt=\alpha_0\mathbf{1}+\mathbf{Z}\alpha+\mathbf{C}\mathbf{J}_{\psi\phi}\mathbf{b}=[\mathbf{1}\ \mathbf{Z}\ \mathbf{C}\mathbf{J}_{\psi\phi}]\zeta \tag{14}$$

where π is the vector of success probabilities.

This model is now similar to a standard logistic regression model, and the maximum likelihood estimator of the parameter can be found by using numerical optimization methods. For instance, Ratcliffe et al. (2002) applied the Fisher scoring algorithm.

It is well-known that estimates of the parameters in multiple logistic regression models are not reliable when there is a high degree of dependency between the covariate variables (multicollinearity), i.e. when the columns of the design matrix are highy correlated (Aguilera et al., 2008; Ryan, 2008). In functional logistic regression, because of the nature of the functional data, there will often be a high correlation between the columns of the matrix $\mathbf{C}\mathbf{J}_{\psi\phi}$ in Eq. (14). The rest of this section will discuss different ways to overcome this problem.

In the following we will, for simplicity, consider a set-up with no scalar covariates. We also introduce the notation l_i for the linear predictor for observation i. Hence, the model under consideration is

$$l_i = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \int X_i(t)\beta(t)\,dt. \tag{15}$$

Notice that we from now on use the notation α for the intercept parameter. Assuming independence of the outcomes, the likelihood for the model is

$$L(\alpha,\beta(t)) = \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i} = \prod_{i=1}^{n} \frac{e^{Y_i(\alpha+\int X_i(t)\beta(t)\,dt)}}{1+e^{\alpha+\int X_i(t)\beta(t)\,dt}}.$$

3.3 Functional logistic regression using wavelet basis and LASSO penalization

Zhao et al. (2012) proposed to combine wavelet bases with LASSO penalization for a functional regression with continuous response, and Mousavi and Sørensen (2015) modified the approach to deal with classification of functional data. They used a multinomial functional regression model and converted the infinite-dimensional problem to a finite-dimensional problem with a sparse matrix of wavelet coefficients. The method was applied to the lameness data (see Section 6) and phoneme data. In this paper we will use the method for binary responses, and accordingly refer to the method as functional logistic regression with wavelets and LASSO (FLRWLASSO).

Let $\phi(t)$ and $\psi(t)$ be a father and mother wavelet, respectively, that are linked by the relationship, $\psi(t) = \sum_{k\in\mathbb{Z}} g_k\sqrt{2}\phi(2t-k)$. The set of the coefficients $\mathscr{G} = \{g_k\}_{k\in\mathbb{Z}}$ are high-pass filter coefficients. For a given mother wavelet, the wavelet bases $\{\psi_{j,k}(t)\}_{j,k\in\mathbb{Z}}$ can be constructed by dilation and translation as follows:

$$\psi_{j,k}(t) = 2^{j/2}\psi(2^j - k).$$

The indices j and k represent dilation and translation, respectively. For a given function $f(t)$ and a fixed detail level j_0, the basis expansion of f in terms of the wavelet bases is given by

$$f_{j_0}(t) = \sum_{k=0}^{2^{j_0}-1} c_{j_0,k}\phi_{j_0,k}(t) + \sum_{j=j_0}^{J-1}\sum_{k=0}^{2^j-1} d_{j,k}\psi_{j,k}(t) \tag{16}$$

where $N = 2^J$ and N represents the number of observations of the function f. Notice that N must be a power of 2.

If we assume that the curves $X_i(t)$ and $\beta(t)$ belong to the finite dimensional space generating by the wavelet basis, then $X_i(t)$ and $\beta(t)$ can be expressed in terms of the mother and father wavelets at detail level j_0 as follows:

$$\begin{aligned} X_{i,j_0}(t) &= \sum_{k=0}^{2^{j_0}-1} c_{i,j_0,k}\phi_{j_0,k}(t) + \sum_{j=j_0}^{J-1}\sum_{k=0}^{2^j-1} d_{i,j,k}\psi_{j,k}(t), \quad i = 1,\ldots,n \\ \beta_{j_0}(t) &= \sum_{k=0}^{2^{j_0}-1} c^*_{j_0,k}\phi_{j_0,k}(t) + \sum_{j=j_0}^{J-1}\sum_{k=0}^{2^j-1} d^*_{j,k}\psi_{j,k}(t). \end{aligned} \tag{17}$$

With these basis expansions for $X_i(t)$ and $\beta(t)$, and due to orthogonality of the wavelet basis functions, the linear predictor from equation (15) becomes

$$l_i = \alpha + \int X_i(t)\beta(t)\,dt = \alpha + \sum_{k=0}^{2^{j_0}-1} c_{i,j_0,k}c^*_{j_0,k} + \sum_{j=j_0}^{J-1}\sum_{k=0}^{2^j-1} d_{i,j,k}d^*_{j,k} = \alpha + B_i\gamma$$

where B_i is a row vector of length N involving of father and mother wavelet coefficients of signal $X_i(t)$, and γ is the vector of c^* and d^* coefficients for the parameter function $\beta(t)$. In matrix form we write $\mathbf{L} = \alpha\mathbf{1} + \mathbf{B}\gamma$, and the likelihood function becomes

$$L(\alpha,\gamma) = \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i} = \prod_{i=1}^{n} \frac{e^{Y_i(\alpha+\int X_i(t)\beta(t)\,dt)}}{1+e^{\alpha+\int X_i(t)\beta(t)\,dt}} = \prod_{i=1}^{n} \frac{e^{Y_i(\alpha+B_i\gamma)}}{1+e^{\alpha+B_i\gamma}}. \tag{18}$$

Notice that no regularization or smoothing has been imposed so far, and that the number of unknown parameters, including the intercept, is $N+1$. With functional data the number of observation for each subject, N, is often much larger than the number of subjects, n ($N \gg n$). In these situations unpenalized ML estimation is not possible as the likelihood equation have several solutions. Moverover, overfitting makes the interpretation of estimates difficult. Therefore we need to apply regularization. It is common to use either a ridge penalty or a LASSO penalty, adding an L^2 or L^1 penalty term on the coefficients onto the log-likelihood.

It is well-known that ridge regression reduces the variability and improves the accuracy of the estimates as the coefficients are shrunk towards zero. This is of particular value in presence of multicollinearity. Howvever, ridge regression is not concerned with variable selection and does not provide a parsimonious model with few parameters. On the other hand, the Least Absolute Shrinking and selection Operator (LASSO) coined by Tibshirani (1996) shrinks some of the coefficients all the way to zero, thereby delivering a sparse solution with just a few non-zero coefficients. In other words, a variable selection is effectively performed.

We use the LASSO, adding a L^1 penalty on the γ coefficients onto the minus log-likelihood, so the objective function is

$$Q(\alpha,\gamma) = -\log L(\alpha,\gamma) + \lambda\sum_{r=1}^{N}|\gamma_r| = -\sum_{i=1}^{n}\Big(Y_i(\alpha+B_i\gamma) - \log(1+e^{\alpha+B_i\gamma})\Big) + \lambda\sum_{r=1}^{N}|\gamma_r|$$

which should be minimized with respect to α and γ. The parameter λ in the objective function Q is a tuning parameter that controls the amount of the shrinkage and should be selected through cross validation. Keep in mind that the objective function Q also depends on the detail level parameter, j_0, which should be chosen via cross-validation as well.

There is no a closed from solution to the minimization problem, so we need to employ an optimization algorithm. Various optimization algorithms have been suggested, such as quadratic programming (Tibshirani, 1996), the LARS algorithm (Efron et al., 2004), and coordinate descent algorithm (Wu and Lange, 2008). The coordinate descent algorithm has advantaged since it can be performed in $\mathcal{O}(nN)$ calculation and with a practical limit of of variables as large as 1000000. In practice we have used the `cv.glmnet` function in the R-package `glmnet` (Friedman et al., 2010) for our numerical studies. It includes cross validation for selection of the tuning parameter λ.

3.4 Functional principle component approach

Let us have a closer look at the principal component approach where the basis functions are eigenfunctions estimated from the data. At first we consider the n individual trajectories as independent realization from a random process $\{X(t), t \in T\}$ with mean function $\mu(t)$ and covariance function $\Sigma_X(s,t) = \text{Cov}(X(s), X(t))$. Mercer's theorem gives us the eigen-decomposition of the covariance function $K(s,t) = \sum_{k=1}^{\infty} \lambda_k \phi_k(s)\phi_k(t)$, where ϕ_k and $\lambda_1 \geq \lambda_2 \geq \ldots$ are the orthogonal eigenfunctions and ordered eigenvalues, respectively. The Karhunen-Loève expansion of the random function $X_i(t)$ can be written as

$$X_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(t) \tag{19}$$

where ξ_{ik} is the kth FPC score for the ith individual and given by $\xi_{ik} = \int (X_i(t) - \mu(t))\phi_k(t)\,dt$. Notice that the FPC scores are uncorrelated random variables with mean 0 and $\text{var}(\xi_{ik}) = \lambda_k$. From the definition of ξ_{ik} we can say that it shows the similarity between the deviation of random function $X_i(t)$ from the mean function and the kth eigenfunction, $\phi_k(t)$.

We now express the time-varying coefficient $\beta(t)$ and the covariate function $X_i(t)$ as truncated expansions in the Karhunen-Loève basis:

$$\beta(t) = \sum_{k=1}^{K} b_k\phi_k(t), \quad X_i(t) = \mu(t) + \sum_{k=1}^{K} \xi_{ik}\phi_k(t).$$

The number of eigenfunctions, K, can be chosen either using the cumulative percent variance method or cross validation. Thanks to the orthogonality of eigenfunctions $\phi_k(t)$, model (15) now becomes

$$l_i = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \tilde{\alpha} + \sum_{k=1}^{K} b_k\xi_{ik}. \tag{20}$$

Here $\tilde{\alpha} = \alpha + \int \beta(t)\bar{X}(t)\,dt$, i.e, the mean function $\bar{X}(t) = \frac{1}{n}\sum_{i=1}^{n} X_i(t)$ is absorbed into the intercept. This model can be written in matrix form as $L = \tilde{\alpha}\mathbf{1} + \xi\mathbf{b}$ where $\mathbf{L} = (l_1, l_2, \ldots, l_n)^T$, $\mathbf{1} = (1,1,\ldots,1)^T$, $\mathbf{b} = (b_1, b_2, \ldots, b_K)^T$, and $\xi = (\xi_{ik})_{n\times K}$ with entries ξ_{ik}, i.e., the kth FPC score for the ith signal.

Once the FPC scores have been estimated, the unknown coefficients b_k can be estimated with a multiple logistic regression, and the estimated coefficient function $\beta(t)$ is then constructed by the basis expansion, $\hat{\beta}(t) = \sum_{k=1}^{K} \hat{b}_k\hat{\phi}_k(t)$. From now on we refer to this approach as functional logistic regression based on functional principal component analysis (FLRFPCA).

It remains to be explained how the FPC scores are computed. In principle, with estimates of the eigenfunctions $\{\hat{\phi}_k(t)\}_{k=1}^{K}$ and the mean function $\bar{x}(t)$, the FPC scores can be obtained by numerical integration as follows:

$$\hat{\xi}_{ik} = \int (X_i(t) - \bar{x}(t))\hat{\phi}_k(t)\,dt \approx \frac{1}{N}\sum_{j=1}^{N} (x_i(t_j) - \bar{x}(t_j))\,\hat{\phi}_k(t_j). \tag{21}$$

These estimates are precise when the observations are dense, but for the sparse data numerical integration are no longer appropriate. As an alternative, Yao et al. (2005) suggested principal analysis via conditional expectation (PACE) for longitudinal data. PACE gives better results for sparse and/or irregular functional data, and we will implement this approach in our simulations and application in Sections 5 and 6. The steps of PACE algorithm are described in detail in appendix A.

3.5 Penalized Functional Regression

Penalized Functional Regression (PFR) is discussed by Goldsmith et al. (2011a) for generalized functional linear models. PFR consists of three steps. First, the random functions X_i are approximated by the finite series expansion $X_i(t) = \sum_{k=1}^{K_x} c_{ik}\psi_k(t)$, where $\psi(t) = \{\psi_1(t), \ldots, \psi_{K_x}(t)\}$ is the set of the first K_x eigenfunctions of the smoothed covariance matrix $\Sigma_X(s,t) = \text{Cov}[X_i(s), X_i(t)]$. Second, a truncated power series basis or a B-spline basis is used to represent the coefficient function, hence $\beta(t) = \sum_{k=1}^{K_\beta} b_k\phi_k(t) = \phi^T(t)\mathbf{b}$ for the selected basis $\phi(t) = \{\phi_1(t), \phi_2(t), \cdots, \phi_{K_\beta}(t)\}$. Third, a penalized log-likelihood is minimized. Notice that the first step in PFR is identical to the first step in FPCR, so the difference lies in the exansion of $\beta(t)$ and penalization.

When $X_i(t)$ and $\beta(t)$ are expansions of the first few eigenfunctions $\{\psi_k\}_{k=1}^{K_x}$ and $\{\phi_k(t)\}_{k=1}^{K_\beta}$ respectively, we can rewrite the model as

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \int X_i(t)\beta(t)\,dt = \alpha + \mathbf{c}_i^T\mathbf{J}_{\psi\phi}\mathbf{b} \tag{22}$$

where $\mathbf{c}_i = (c_{i1}, \cdots, c_{iK_x})^T$, $\mathbf{b} = (b_1, \cdots, b_{K_b})^T$, and W is an $K_x \times K_b$ matrix with the (k,ℓ)-th element $\mathbf{J}_{\psi\phi\,k\ell} = \int \psi_k(t)\phi_\ell(t)\,dt$. The original PFR paper (Goldsmith et al., 2011a) used a truncated power series basis, whereas PFR with B-splines and difference penalties was implemented in a modified version of the R `refund` package (Crainiceanu et al., 2014).

Notice that two different approaches have been discussed in the literature on penalized splines: (1) a B-spline basis with equally spaced knots and difference penalties (Eilers and Marx, 1996) and (2) a truncated power series basis with unequally spaced knots usually based on the quintile of the time observation and a ridge penalty (Ruppert et al., 2003). Eilers and Marx (2010) showed that B-splines and difference penalties are easily adopted to smoothing of periodic data. This can be done by wrapping around the basis functions at the 'end' to the 'begining' and also changing the difference penalty in a similar way. They also mention that there is no evidence of any advantage of penalized truncated power series functions over the penalized B-splines. We will use B-splines with difference penalty as implemented in the function `pfr` in the R `refund` package (Crainiceanu et al., 2014) for our numerical studies.

4 Bootstrap for functional logistic regression

The simulation study in Section 5 shows that the estimator of β is subject to great uncertainty and that it is not very reliable as estimator of the true data generating mechanism. Nevertheless,

for a given dataset, we are indeed interested in describing the uncertainty of the estimator. It is not possible to make explicit inference for $\beta(t)$ so we will rely on bootstrap methods. Febrero-Bande et al. (2010) used the smoothed bootstrap approach for functional linear models with scalar response. We suggest to use the procedure for functional logistic regression as well, and we will apply to the lameness data in Section 6.

As in the previous sections we consider n observations, i.e. n binary observations collected in the vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ and n covariate functions collected in $\boldsymbol{X}(t) = (X_1(t), \ldots, X_n(t))$. The bootstrap procedure consists of the following steps:

1. Fit the functional logistic regression from equation (15) using one of the methods from Section 3. Let $\hat{\alpha}$ and $\hat{\beta}(t)$ denote the estimates of the intercept α and the coefficient function $\beta(t)$, respectively.

2. Make K_B standard bootstrap samples of size n from the original covariate curves, and denote the new samples $\boldsymbol{X}^b(t) = \{X_1^b(t), X_2^b(t), \ldots, X_n^b(t)\}, b = 1, 2, \ldots, K_B$.

3. Perform smoothed bootstrap by adding a multivariate Gaussian process to the bootstrap samples. More specifically, draw Z^b from a multivariate Gaussian process with mean zero and covariance matrix $s\Sigma_X$, where s is a smoothing parameter, and Σ_X is an estimate of the covariance matrix for the functional covariate, and let $\tilde{\boldsymbol{X}}^b(t) = \boldsymbol{X}^b(t) + Z^b$, for $b = 1, 2, \ldots, K_B$.In order to choose the smoothing parameter Febrero-Bande et al. (2010) showed through a simulation study that an appropriate choice could be $s \in (0.15, 0.25)$ for functional datasets with sample size $n \leq 100$, while $s \in (0.1, 0.2)$ is appropriate for sample sizes $100 < n < 200$, and for larger sample sizes unsmoothed data may be used.

4. Compute $\boldsymbol{\pi}^b = (\pi_1^b, \pi_2^b, \ldots, \pi_n^b)$ where π_i^b is the probability that the binary response Y_i variable takes value one given functional observation $x_i^b(t)$,

$$\pi_i^b = P\{Y_i = 1 | X(t) = X_i^b(t)\} = \frac{\exp\{\hat{\alpha} + \int X_i^b(t)\hat{\beta}(t)\,dt\}}{1 + \exp\{\hat{\alpha} + \int X_i^b(t)\hat{\beta}(t)\,dt\}}.$$

   Notice that we use the non-smoothed bootstrap data for this computation.

5. Use π_i^b, $i = 1, 2, \ldots, n$ to generate random binary data Y_i^b with success probability π_i^b, and denote the response vector $\boldsymbol{Y}^b = (Y_1^b, Y_2^b, \ldots, Y_n^b)$.

6. Fit the functional logistic model to the bootstrap data with $\boldsymbol{Y}^b$ as response and $\tilde{\boldsymbol{X}}^b(t)$ as covariate functions. Denote the estimate of the the coefficient function $\hat{\beta}^b(t)$for $b = 1, 2, \ldots, K_B$.

7. Compute $d_b = d(\hat{\beta}(t), \hat{\beta}^b(t))$, $b = 1, 2, \ldots, K_B$ where $d(\cdot, \cdot)$ is a metric associated with a norm. Define d_α as the $(1-\alpha)$ quantile in the empirical distribution of $d_1, \ldots, d_{K_B}$.

8. Define and plot the bootstrap confidence "ball" of level $(1-\alpha)$ as those bootstrap estimates whose distance to $\hat{\beta}(t)$ is smaller than d_α, i.e., $CB_{(1-\alpha)} = \{\hat{\beta}^b(t) \,\big|\, d_b \leq d_\alpha\}$.

9. Plot the estimated parameter functions in $CB_{1-\alpha}$ together with the estimated coefficient function from original data $\hat{\beta}(t)$, but in different colors.

5 Simulation study

In this section, a simulation study is performed in order to further evaluate and compare the aforementioned approaches. We try two different approaches for simulation; both inspired from the literature on functional regression. Another option would be to resemble an existing dataset by borrowing parameter values estimated from the data as in Mousavi and Sørensen (2015).

5.1 Simulation from a functional logistic regression model

Our first simulation approach consists of two steps: First simulate functional covariates X_i (all from the same distribution); then simulate the response Y_i from a logistic regression model with a fixed $\beta(t)$. In other words, the simulation model is in accordance with the model used for classification. More specifically, for the first step, we consider equally-spaced time points $\{t_j \in [0,10], j = 1,2,\cdots,256\}$ with length $2^8 = 256$, and generate 150 functional predictors using basis expansions on the form

$$X_i(t_j) = \sum_{k=1}^{13} c_{ik}\phi_k(t_j) \quad , \quad i = 1,2,\ldots,150 \quad , \quad j = 1,2,\ldots,256 \quad , \quad t_{ij} \in [0,10] \tag{23}$$

Here the basis functions $\{\phi_k(t)\}_{k=1}^{13}$ are cubic B-splines corresponding to nine equally space interior knots over the interval $[0,10]$, and c_{ik} are random basis coefficients generated as follows: The 150×13 matrix C is a product ZU where Z is a 150×13 matrix of iid. standard normal variables, and U is a 13×13 matrix of iid. random values with uniform distribution on $[0,1]$. This method of generating the functional covariates is adapted from the work of Escabias et al. (2004). Left panel of Figure 1 shows a sample of 10 random functional covariates $X(t)$. On top of this, we allow for measurement errors on the functional data and thus consider the curves contaminated with noise, i.e. $W_i(t_j) = X_i(t_j) + \delta_i(t_j)$ where $\delta_i(t_j) \sim n(0,\sigma_X^2)$. We use $\sigma_X = 0$ (no noise) and $\sigma_X^2 = 0.5$ as standard deviation in our study.

In the second step the binary response Y_i is generated by the following model:

$$\text{logit } Pr\{Y_i = 1|X_i(t)\} \;=\; \int_0^{10} X_i(t)\beta(t)\,dt \quad , \quad i = 1,2,\ldots,150 \tag{24}$$

where $\beta(t)$ is the parameter function. Notice that the intercept α is zero. We consider three different choices of true parameter functions, namely, $\beta_1(t) = \sin(t\pi/3)$, $\beta_2(t) = (t/2.5)^2/5$, and $\beta_3(t) = -p(t|2,0.3) + 3p(t|5,0.4) + p(t|7.5,0.5)$, where $p(\cdot|\mu,\sigma)$ represents the normal density with mean μ and standard devision σ. These functions are adapted from the work of Goldsmith et al. (2011b) with minor changes in order to generate datasets compatible with the binary model. The true parameter functions $\beta_j(t)$ are displayed in the right panel of Figure 1. The curves in the left part of the figure are coloured according to the value of Y_i as obtained from (24) with $\beta_1(t)$. For two curves the W process is also shown (dashed curves).

We simulated 100 datasets $\{Y_i, X_i(t_j), W_i(t_j),\ j = 1,2,\ldots,256,\ i = 1,2,\ldots,150\}$ of the above type for each $\beta_j(t)$, $j = 1,2,3$. For each dataset, 100 observations were used as training data

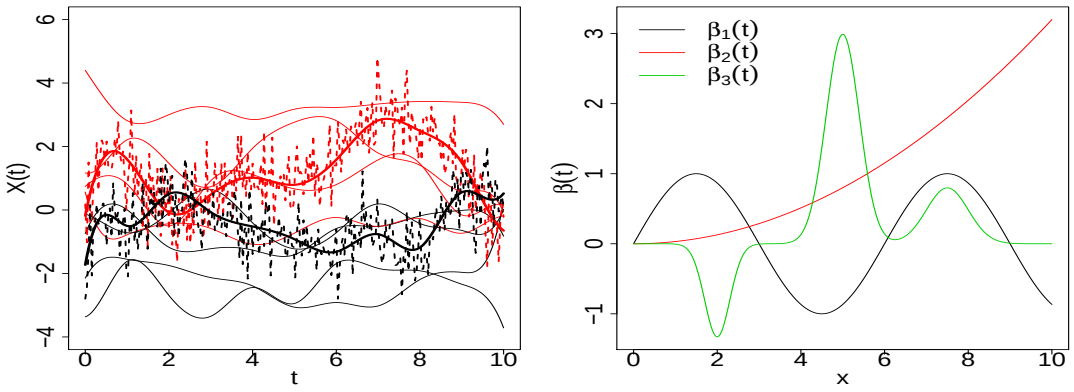

Figure 1: Left: Sample of 10 random functional covariates X_i generated from (23). Curves are coloured according to the outcome of Y (black and red for 0 and 1, respectively), generated from (24) with parameter function β_1. For the two highlighted curves the W process is also shown (dashed curves). Right: Parameter functions used for simulation of the response.

in order to fit the logistic model, and the fitted parameter function as well as the corresponding predicted responses were extracted. The remaining 50 observations were used as test data, i.e. the response was considered unknown and the fitted model was used to predict it. This set-up is classical in the evaluation of procedures for supervised classification.

All three approaches from Section 3 (FLRFPCA, PFR, FLRWLASSO) were tested with all three β_js and both with and without measurement noise. In the FLRPCA approach the number of eigenfunction was selected to explain 98% of the variation, which in practice gave around 5 eigenfunctions. In the FLRWLASSO approach the detail level j_0 was selected by cross validation, and varied between the different scenarios (different true β_js, with/without noise). In general, data with more oscillation requires a larger value of j_0. For PFR, we used $K_\beta = 30$, i.e. 30 basis functions to represent the parameter function.

The misclassification rates are listed in Table 1, and also illustrated in Figures 2 and 3 for test and training data, respectively. As expected misclassification rates are larger for test compared to training data. All approaches are successful in handling datasets with noise as the misclassification rate does not increase much when functions are contaminated with noise. The misclassification rates for the datasets generated by $\beta_2(t)$ are the lowest, however, the rates for data generated by $\beta_1(t)$ are rather large compared to $\beta_2(t)$ and $\beta_3(t)$. The misclassification rates for test data are similar for the three methods with PFR performing slightly better than FLRFPCA and FLRWLASSO across the six scenarios.

Notice that for the simulation set-up just described the simulation model and the regression model are of the same type (the regression model is true), and another aspect is therefore the ability of the different methods to reproduce the parameter functions $\beta(t)$. Unfortunately, it turns out that neither of the methods does a good job in that respect. FLRFPCA gives reasonable estimates for $\beta_1(t)$, but has large scale problems for $\beta_2(t)$ and is not able to reproduce the

shape for $\beta_3(t)$. PFR generally has severe problems with the scale of parameter functions, but is to some extent able to reproduce the shape of $\beta(t)$. FLRWLASSO has problems with both shape and scale. We conclude that the estimates of β are not reliable as estimates of the true data generating mechanism. However, the estimates can still be useful in applications as they describe the corresponding "prediction machine", and thereby includes information about which parts of the functional data that hold information associated to the response.

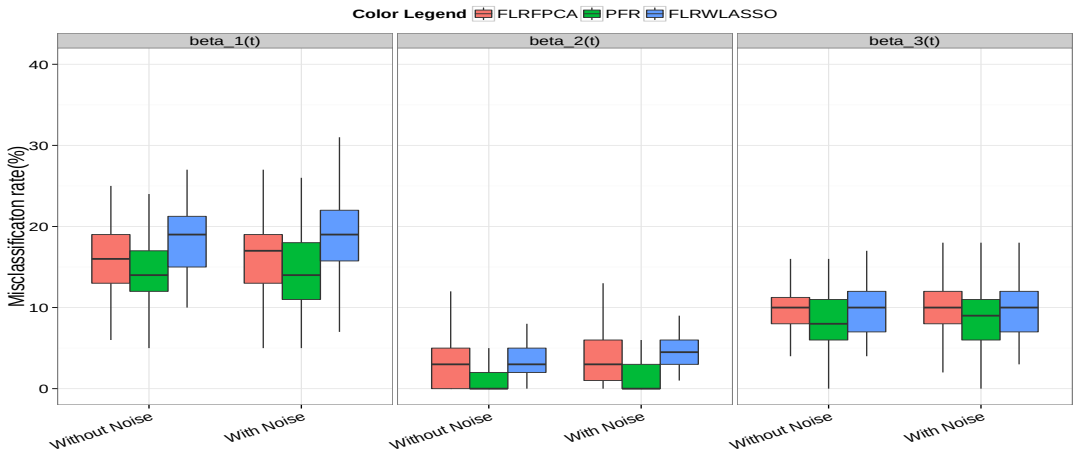

Figure 2: Boxplots of the misclassification rates for training data in the different scenarios and for each of the three methods. Each boxplot is based on 100 simulated datasets, each containing 50 test curves.

Table 1: Average misclassification rate (over 100 simulated datasets) in percent for each method and each simulation scenario. Data are simulated as described in Section 5.1.

		FLRFPCA		FPR		FLRWLASSO	
		$\sigma_X^2 = 0$	$\sigma_X^2 = 0.5$	$\sigma_X^2 = 0$	$\sigma_X^2 = 0.5$	$\sigma_X^2 = 0$	$\sigma_X^2 = 0.5$
$\beta_1(t)$	Test	18.78	19.22	18.46	18.70	20.56	20.96
	Train	15.75	16.24	14.08	14.21	18.78	18.41
$\beta_2(t)$	Test	5.92	5.76	4.18	4.26	4.80	5.36
	Train	3.31	3.72	1.33	1.80	3.67	4.73
$\beta_3(t)$	Test	11.90	11.72	11.80	11.64	11.60	11.98
	Train	9.91	10.07	8.57	8.96	9.98	9.74

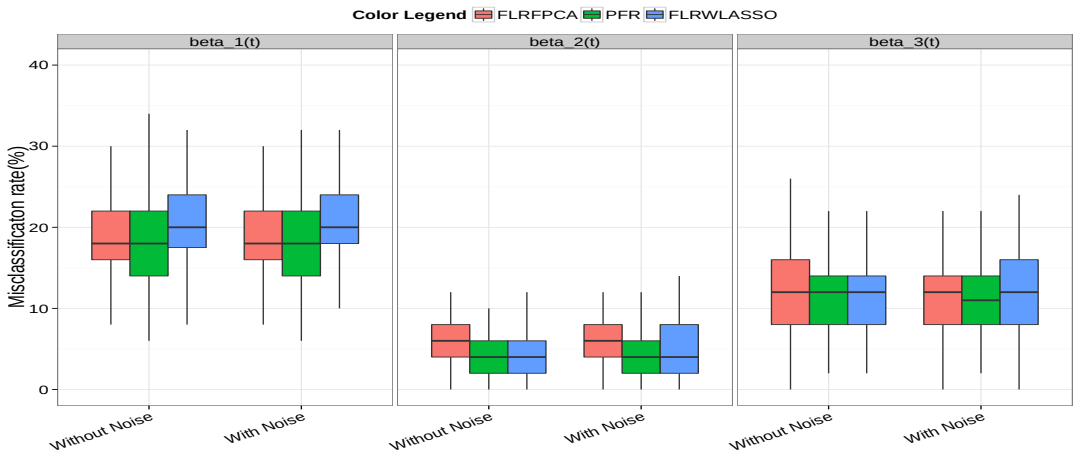

Figure 3: Boxplots of the misclassification rates for test data in the different scenarios and for each of the three methods. Each boxplot is based on 100 simulated datasets, each containing 50 test curves.

5.2 Stratified simulation from two groups

Our second simulation approach simulates curves from two different distributions corresponding to $Y = 0$ and $Y = 1$, respectively. This simulation set-up is adopted from the work of Aguilera et al. (2011).

Each simulated dataset contains a total of 250 curves of two different classes of sample curves. in the first class 125 random curves are generated as $x(t) = uh_1(t) + (1-u)h_2(t) + \varepsilon(t)$ while in the second class 125 random curves are generated $x(t) = uh_1(t) + (1-u)h_3(t) + \varepsilon(t)$. Here u and $\varepsilon(t)$ are iid. uniform and standard normal random variables, respectively, and $h_1(t) = \max\{6 - |t-11|, 0\}$, $h_2(t) = h_1(t-3)$, and $h_3(t) = h_1(t+3)$. The sample curves are generated at 101 equally spaced timepoints on the interval $[1, 21]$, and the binary response Y is considered as 0 for curves belonging to the first class and 1 for the curves from the second class. Notice that there is no true parameter function in this set-up as the curves are simulated conditionally on the response (not the opposite). Figure 4 shows a sample of 20 random functional covariates $X_i(t)$.

We divided each dataset into training data (150 curves) and test data (100 curves), and proceeded as in Section 5.1. A total of 200 such datasets were simulated, and the average misclassification rates are shown in Table 2. From a classification point of view there is not much difference between the different approaches. Recall that there is no true $\beta(t)$ for these simulation. Nevertheless, each procedure delivers an estimate of $\beta(t)$ that is used for classification, and it turns out that FLRWLASSO leads to a stable estimate whereas the estimates from FLRFPCA and PFR are extremely variable.

Our conclusions based on our simulation studies (Sections 5.1 and 5.2) are the following: (1)

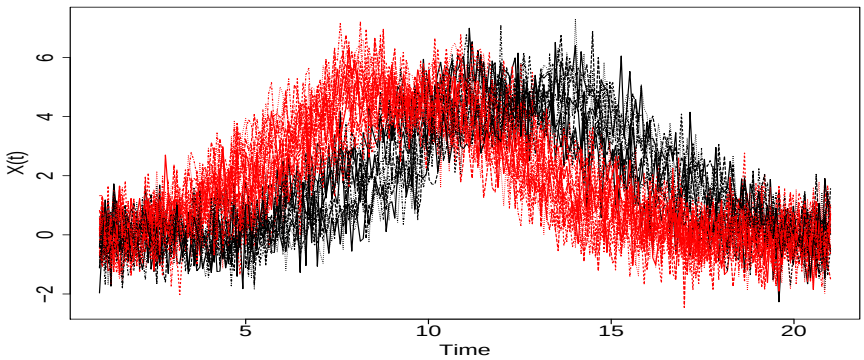

Figure 4: Sample of 30 random signal generated from two different classes. Curves are coloured according to the classes.

Functional logistic regression can be useful for classification — even if the true data generating model is not a functional logistic regression model; (2) the classification results do not differ much between the three considered approaches; (3) the estimated regression model is not necessarily reliable as an estimate of the true conditional distribution of the response given the curve, but can be useful for identification of intervals with strong association between curve and response. Finally, computation time is an important aspect, in particular if one wants to use bootstrap or other resampling methods for statistical inference. As will be illustrated in Section 6, FLRWLASSO is roughly a factor 25 faster than the other methods which gives that method a large advantage over the others.

Table 2: Average misclassification rate (over 200 simulated datasets) in percent for each method. Data are simulated as described in Section 5.2.

	FLRFPCA	FPR	FLRWLASSO
Test data	2.07	2.29	2.50
Train data	1.03	0.83	1.12

6 Lameness detection for horses

Lameness is a common problem for sports horses. Detection of lameness at an early stage could prevent chronic lameness (Stashak, 2002; Thomsen et al., 2010), but low-degree lameness is difficult to detect with clinical inspection, so supplementary methods for lameness detection and identification of the lame limb would be welcome.

Walk, trot and canter are the most common gaits, and the first two are symmetric. It is well known that lameness disturbs the symmetry, so continuous monitoring of activities from these gaits would be expected to be informative about the lameness status of the horse. Horses can be monitored with accelerometers which measure the activity through electrical signals that can be converted to proxy measurements for acceleration. Thomsen et al. (2010) recorded acceleration

data from trotting horse with this technology; with and without stimulation of lameness. In the following we will give a brief description of these data, and then apply FLRFPCA, PFR, and FLRWLASSO for supervised classification in three different scenarios.

6.1 Data collection and lameness groups

A 10G, three-axis accelerometer was used to record the signal of acceleration in three directions (vertical, transversal, longitodinal). The accelerometer was put on the lowest point of the back of horse which is the closest surface location to the body center of mass. More details on the data collection process can be found in Thomsen et al. (2010) and Sørensen et al. (2012).

Eight horses with no indication of lameness were used in two sub-experiments to generate a total of 85 acceleration signals in five lameness groups. Lameness was induced mechanically by equipping the horse with a modified horseshoe with a screw eliciting pressure on the sole of the hoof. The shoe was attached to one of the four hoofs and horses were also tested without the shoe; amounting to five groups.

Experience is that acceleration is similar for lameness on limbs from the same diagnal pair of limbs. Therefore, and in order to have larger groups, we only consider three groups in the following: Normal (NO) consisting of 23 signals from horses with no shoe attached; left diagonal (LD) consisting of 30 signals from horses with the shoe attached to left fore or right hind limb; and right diagonal (RD) consisting of 32 signals from horses with the shoe attached to right fore or left hind limb. We will study three scenarios:

1. Left diagonal vs. right diagonal (LD/RD)
2. Normal vs. left diagonal (NO/LD)
3. Normal vs. right diagonal (NO/RD).

We will only use the acceleration in the vertical direction in the current study. The raw acceleration signals consist of data from eight complete gait cycles (between 1121 and 1440 observations). Before the analysis we carried out several pre-processing steps in order to reduce variation between gate cycles in each signal and variation between signals due to different timing at the beginning. These steps of preprocessing have been explained in Appendix A in Mousavi and Sørensen (2015).

After pre-processing the data consists of 85 signals on $(0, 1)$, and each signal represents one gait cycle. Importantly, and for all signals, the first half correponds to stance on the right diagonal whereas the second half corresponds to stance on the left diagonal. The signals are shown in Figure 5, divided into groups. A close look reveals that for the healthy horses (NO), the first and second halfes are similar, whereas this symmetry is disturbed in groups LD and RD. More specifically, signals from the RD group generally have smaller amplitude on $(0, 0.5)$ compared to $(0.5, 1)$, and vice versa for the LD group. This is because horses tend to put less pressure, and thus generate less upward acceleration, when they stand on the lame compared to the healthy diagonal.

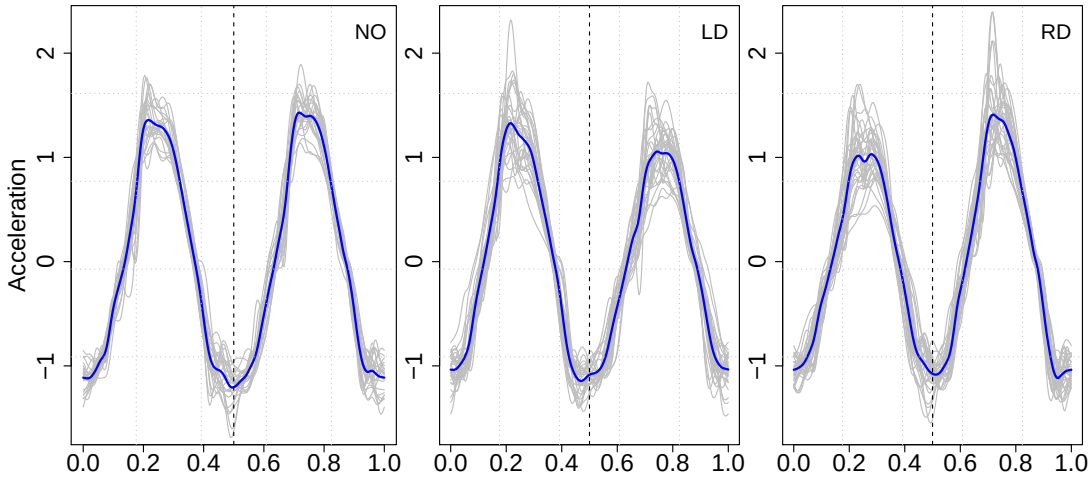

Figure 5: Vertical acceleration signals (grey lines) from the normal (NO), left diagonal (LD) and right diagonal (RD) groups. The blue lines are the average curves.

6.2 Analysis

We now apply FLRFPCA, PFR and FLRWLASSO in the three scenarios. The tuning parameters λ and j_0 for FLRWLASSO are selected by cross validation. This gives $j_0 = 0$ in scenario LD/RD, $j_0 = 1$ in other two scenarios, and values 0.0067, 0.0059 and 0.0114 for λ in scenario 1, 2 and 3, respectively. The number of eigenfunctions for FLRFPCA was selected such that 99% of the variation was explained; this gave 7 eigenfunctions in all three scenarios. For the PFR approach, 30 B-spline basis was used to expand parameter function $\beta(t)$. The primary aim of the analysis is classification, but we will also study estimates of the parameter function $\beta(t)$, and comment on execution times.

First, as is common for evaluation in supervised classification, we use the leave-one-curve-out approach. That is, all data (signals and groups) except for the ith observation are used as training data to fit the model, then signal i is used as test data, and the prediction is compared to the true group. This is repeated for all signals, i.e. $i = 1, 2, \ldots, n$, where n is the number of curves in the scenario under consideration. The results are displayed in Table 3 and Table 4 where the first shows the true and predicted groups for each scenario and each approach, and the latter shows the misclassification rates in percent. The tables show that from a classification point of view, the approaches FPR and FLRWLASSO are very similar, whereas FLRPCA is less good.

Second, let us examine the estimated parameter function $\hat{\beta}(t)$ for each scenario and each approach, where in each case all available observations have been used for estimation. The estimates are shown in Figure 6. The shape of $\hat{\beta}(t)$ is roughly the same for FPR and FLRWLASSO, except perhaps for some large jumps in the NO vs. LD scenario fitted with FLRWLASSO. The scales for $\hat{\beta}$ vary a lot between estimation methods and scenarios, with FLRWLASSO being the most stable across scenarios. The majority of the estimates are positive on roughly one

Table 3: Results from leave-one-curve-out classification.

	FLRFPCA			**FPR**			**FLRWLASSO**		
Scenario		Predicted			Predicted			Predicted	
LD	True Group	LD	RD	True Group	LD	RD	True Group	LD	RD
vs	LD	28	2	LD	29	1	LD	29	1
RD	RD	2	30	RD	2	30	RD	1	31
NO	True Group	NO	LD	True Group	NO	LD	True Group	NO	LD
vs	NO	21	2	NO	21	2	NO	21	2
LD	LD	4	26	LD	1	29	LD	3	27
NO	True Group	NO	RD	True Group	NO	RD	True Group	NO	RD
vs	NO	19	4	NO	20	3	NO	20	3
RD	RD	3	29	RD	2	30	RD	2	30

Table 4: Misclassification rates based on leave-one-curve-out classification.

Scenario	FLRFPCA	FPR	FLRWLASSO
LD vs RD	6.45	4.84	3.23
NO vs LD	11.32	5.66	9.43
NO vs RD	12.73	9.09	9.09

half of the interval and negative on the other, suggesting that the *difference* between the two halves of a signal is associated to the lameness status. This is not surprising. Notice that, due to symmetry of trot, and because all signals start with stance on the right diagonal, we would expect the behaviour of $\hat{\beta}(t)$ for NO vs. LD on the interval $(0, 0.5)$ to be similar to the behaviour of $\hat{\beta}(t)$ for NO vs. RD on the interval $(0.5, 1)$, and vice versa. This is indeed the case for FRP when it comes to shape, but not scale, and to some extent for FLRWLASSO, but not FLRFPCA.

In order to examine the stability of the estimates we used the bootstrap approach discussed in Section **??**. Figure 7 shows the result for FLRWLASSO with $K_B = 300$ bootstrap samples. The blue line represents the estimated parameter function for the original data, the gray curves are the estimated parameter function for each bootstrap sample after by using L^2-norm, and the red dashed line represents the 95% pointwise confidence band for the parameter function. It is not surprising that the pointwise confidence bands are more narrow than the range of the gray curves, since the latter takes into account the whole shape of the curves. However, we would have expected the blue line to be closer to the center of the bootstrap distribution as it is the case for functional linear regression with continuous outcome (Febrero-Bande and Oviedo de la Fuente, 2012). The extremely large of $\hat{\beta}(t)$ for FLRFPCA and FPR led to numerical problems in the bootstrap computations, which were therefore not carried out for these two methods.

Altogether the study of the estimated parameter functions confirm our impression from the simulation studies, that the parameter estimates are unstable and not all that reliable. They can

at best give an impression about the associations between the binary outcome and the covariate curves.

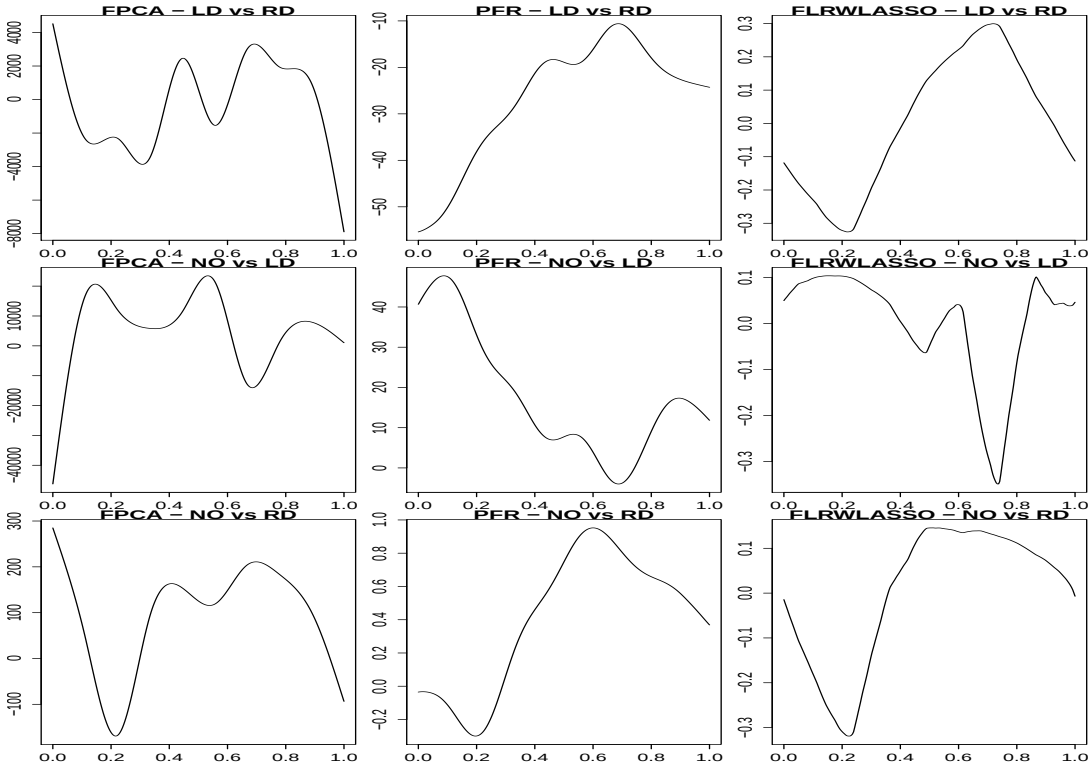

Figure 6: Estimated parameter function $\hat{\beta}(t)$ for the three different approaches and three scenarios. All the available data have been used in each scanario.

Finally, some comments on execution time. The computation time in seconds for fitting the model with all available data for each approach and scenario has been measured and is reported in Table 5. The R-implementation has been executed on a i3-core Pentium processor with 4 GM of RAM. As the table shows, FLRWLASSO is far more computationally efficient than the other two approaches, at least a factor 25. Such a difference is important when performing many modelfits such as in a leave-one-curve-out study or with bootstrap computations.

Table 5: Execution time in seconds for the different approaches and scenarios.

Scenario	FLRFPCA	FPR	FLRWLASSO
LD vs RD	6.62	8.75	0.26
NO vs LD	6.92	7.66	0.26
NO vs RD	6.48	7.16	0.24

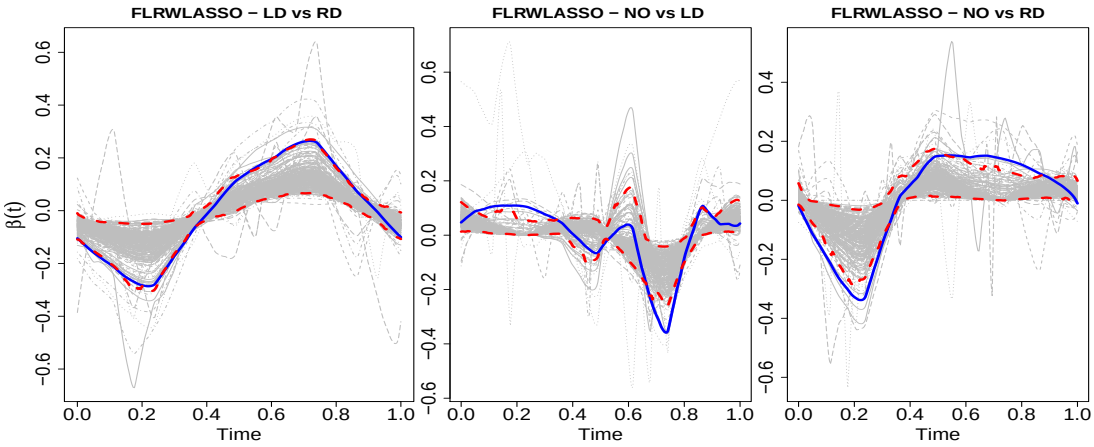

Figure 7: Estimates of the parameter function $\beta(t)$ for the FLRWLASSO method. The blue curve represents the estimate of parameter function for the original data, the gray curves are the estimates from 300 bootstrap samples after taking away 5% of the outlier estimated parameters, and the red dashed curves display pointwise 95% confidence band for the parameter function based on the bootstrap samples.

7 Discussion

In this paper, functional logistic regression was reviewed and three approaches were compared: functional logistic regression with wavelets and LASSO penalization (FLRWLASOO), functional logistic regression via functional principle component analysis (FLRFPCA), and penalized functional regression (PFR). The performance of an approach in functional logistic regression can be assessed in two directions: classification and estimation of the parameter function. Based on our simulation study and the application to the lamess data, we conclude the following:

- Misclassification rates are similar for the three approaches with PFR performing slightly better in the simulation study, and PFR and FLRWLASSO performing better than FLRFPCA in the application.
- None of the three methods does a good job in estimating the parameter function. The estimates are not reliable, mainly due to scale problems. The scale problems can lead to numerical problem in connection to bootstrap or other ressampling methods.
- FLRWLASSO gives the most stable estimates of the parameter function, and despite the above-mentioned problems, the estimated parameter function is still interesting in applications as it contains information about which parts of the functional covariates that hold information related to the binary response.
- The R-implementation of the three methods showed that FLRWLASSO is far more computational efficient than the other two methods, at least by a factor 25.

In summary, PFR and FLRWLASSO seems to be preferable to FLRFPCA from a classification point, and FLRWLASSO is much faster and more stable regarding estimation of the parameter function. For these reasons, and in particular if bootstrap or other resampling methods are used as part of the analysis, we recommend to use FLRWLASSO for functional logistic regression.

References

Aguilera, A., Aguilera-Morillo, M., Escabias, M., and Valderrama, M. (2011). Penalized spline approaches for functional principal component logit regression. In *Recent Advances in Functional Data Analysis and Related Topics*, pages 1–7. Springer.

Aguilera, A. M., Escabias, M., and Valderrama, M. J. (2006). Using principal components for estimating logistic regression with high-dimensional multicollinear data. *Computational Statistics & Data Analysis*, 50(8):1905–1924.

Aguilera, A. M., Escabias, M., and Valderrama, M. J. (2008). Discussion of different logistic models with functional data. application to systemic lupus erythematosus. *Computational Statistics & Data Analysis*, 53(1):151–163.

Capra, W. B. and Müller, H.-G. (1997). An accelerated-time model for response curves. *Journal of the American Statistical Association*, 92(437):72–83.

Cardot, H., Ferraty, F., and Sarda, P. (1999). Functional linear model. *Statistics & Probability Letters*, 45(1):11–22.

Cardot, H., Ferraty, F., and Sarda, P. (2003). Spline estimators for the functional linear model. *Statistica Sinica*, 13(3):571–592.

Crainiceanu, C., Reiss, P., Goldsmith, J., Huang, L., Huo, L., Scheipl, F., Swihart, B., and Huang, M. L. (2014). Package 'refund'.

Delaigle, A., Hall, P., and Apanasovich, T. V. (2009). Weighted least squares methods for prediction in the functional data linear model. *Electronic Journal of Statistics*, 3:865–885.

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

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science*, pages 89–102.

Eilers, P. H. and Marx, B. D. (2010). Splines, knots, and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(6):637–653.

Escabias, M., Aguilera, A. M., and Valderrama, M. J. (2004). Principal component estimation of functional logistic regression: discussion of two different approaches. *Journal of Nonparametric Statistics*, 16(3-4):365–384.

Febrero-Bande, M., Galeano, P., and González-Manteiga, W. (2010). Measures of influence for the functional linear model with scalar response. *Journal of Multivariate Analysis*, 101(2):327–339.

Febrero-Bande, M. and Oviedo de la Fuente, M. (2012). Statistical computing in functional data analysis: the r package fda. usc. *Journal of Statistical Software*, 51(4):1–28.

Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1.

Goldsmith, J., Bobb, J., Crainiceanu, C. M., Caffo, B., and Reich, D. (2011a). Penalized functional regression. *Computational and Graphical Statistics*, 20(4):830–851.

Goldsmith, J., Crainiceanu, C. M., Caffo, B. S., and Reich, D. S. (2011b). Penalized functional regression analysis of white-matter tract profiles in multiple sclerosis. *NeuroImage*, 57(2):431–439.

Goldsmith, J., Greven, S., and Crainiceanu, C. (2013). Corrected confidence bands for functional data using principal components. *Biometrics*, 69(1):41–51.

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

James, G. M., Wang, J., and Zhu, J. (2009). Functional linear regression that's interpretable. *The Annals of Statistics*, 37(5A):2083–2108.

Marx, B. D. and Eilers, P. H. (1999). Generalized linear regression on sampled signals and curves: a p-spline approach. *Technometrics*, 41(1):1–13.

Mousavi, S. N. and Sørensen, H. (2015). Multinomial functional regression with wavelets and lasso penalization. manuscript.

Müller, H.-G. and Yao, F. (2008). Functional additive models. *Journal of the American Statistical Association*, 103(484):1534–1544.

Ramsay, J. O. and Dalzell, C. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 539–572.

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

Ratcliffe, S. J., Heller, G. Z., and Leader, L. R. (2002). Functional data analysis with application to periodically stimulated foetal heart rate data. ii: Functional logistic regression. *Statistics in Medicine*, 21(8):1115–1127.

Rice, J. A. (2004). Functional and longitudinal data analysis: perspectives on smoothing. *Statistica Sinica*, 14(3):631–648.

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 233–243.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Number 12. Cambridge university press.

Ryan, T. P. (2008). *Modern Regression Methods*, volume 655. John Wiley & Sons.

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

Stashak, T. (2002). Examination for lameness. *Adams' Lameness in Horses*, 5:113–183.

Thomsen, M. H., Jensen, A. T., Sørensen, H., Lindegaard, C., and Andersen, P. H. (2010). Symmetry indices based on accelerometric data in trotting horses. *Journal of Biomechanics*, 43(13):2608–2612.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Wei, P., Tang, H., and Li, D. (2014). Functional logistic regression approach to detecting gene by longitudinal environmental exposure interaction in a case-control study. *Genetic Epidemiology*, 38(7):638–651.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):3–36.

Wu, T. T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, pages 224–244.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590.

Zhao, Y., Ogden, R. T., and Reiss, P. T. (2012). Wavelet-based lasso in functional linear regression. *Journal of Computational and Graphical Statistics*, 21(3):600–617.

A Principal analysis via conditional expectation

The PACE algorithm includes estimation of the eigenvalues, eigenfunctions, and scores and consists of the following steps:

Step 1 Estimate the mean function $\mu(t)$ using univariate smoothing to the pooled observations under working independence.

Step 2 Use a method-of-moment approach to construct a raw estimate of the covariance matrix, and then use bivariate smoothing to smooth the off-diagonal elements in the covariance matrix. Then use the smoother to also recover the covariance function along the diagonal.

Step 3 Spectral decomposition of the smooth covariance function gives eigenvalues and eigenfunction, $\{\hat{\lambda_k}, \hat{\phi_k}(t), k = 1, 2, \cdots, N\}$.

Step 4 Estimate the truncation lag K, i.e., by the cumulative percent variance method or cross validation.

Step 5 Estimate the measurement error variance by considering the difference between the diagonal elements of the smooth covariance matrix and the raw estimate of the covariance matrix.

Step 6 Use the "best linear unbiased predictor" (BLUP) approach to estimate the FPC scores ξ_{ik} in the following mixed effect model:

$$X_i(t_{ij}) = \mu(t_{ij}) + \sum_{k=1}^{K} \phi_k(t_{ij})\xi_{ik} + \varepsilon_{ij}$$

where $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$, $\xi_{ik} \overset{iid}{\sim} N(0, \lambda_k)$ and in addition $\varepsilon_{\mathbf{i}} = (\varepsilon_{i1}, \varepsilon_{i2}, \cdots, \varepsilon_{iN})$ and $\xi_{\mathbf{i}} = \{\xi_{ik}, k = 1, 2, \cdots, K\}$ are independent. The mean function $\mu(t)$, the spectral decomposition $\{\lambda_k, \phi_k(t), k = 1, 2, \cdots, K\}$, the residual varians σ^2, and K are considered known and equal to their estimated versions. Hence, the goal is to predict ξ_{ik} given $X_i(t_{ij}), \mu(t)$, $\{\lambda_k, \phi_k(t), k = 1, 2, \cdots, K\}$, K, and σ^2.

Step 7 Estimate the smooth curve X_i as the Karhunen-Loève expansion with K terms, $\hat{X}_i(t) = \hat{\mu}(t) + \sum_{k=1}^{K} \hat{\xi}_{ik}\hat{\phi}_k(t)$.

III

Generalized time-varying regression of multilevel functional data

Seyed Nourollah Mousavi
Department of Mathematical Sciences
University of Copenhagen

Ana-Maria Staicu
Department of Statistics
North Carolina State University

Damla Şentürk
Department of Biostatistics
UCLA School of Public Health

Publication details

Preparing to submit (2015).

Generalized Time-Varying Regression of Multilevel Functional Data

October 30, 2015

Seyed Nourollah Mousavi

Department of Mathematical Sciences,

Copenhagen, Denmark

email: nourollah@math.ku.dk

Ana-Maria Staicu

Department of Statistics, North Carolina State University, 2311 Stinson Drive,

Raleigh, NC 27695-8203, USA

email: staicu@stat.ncsu.edu

Damla Şentürk

Department of Biostatistics, UCLA School of Public Health,

Los Angeles, CA 90095, USA

email: dsenturk@stat.psu.edu

Abstract

We provide a concurrent functional model for the multilevel functional data of the type subject-unit when both covariate and response with the unit-level being functional. We suggest a generalized regression model for this multilevel functional setting to relate the functional response to the structural component of the functional covariate which enables an easy interpretation and new insights into the building blocks of the regression model. The proposed estimation approach is based on method of moment techniques which leads to a fast computation. The proposed approach performs well in simulations.

Key words: Multilevel functional data, Principle component analysis, Concurrent functional model, Generalized time-varying regression, Taylor expansion, Bivariate smoothing.

1 Introduction

Functional data analysis is a fast growing area of statistical research with increasingly diverse range of applications from economics, medicine, agriculture, chemometrics and so on. In this paper, we consider functional data which have a multilevel structure, of the type subject-unit, with the unit-level data being functional observations. Our focus is to develop association models when both the outcome and the predictors have this multilevel structure.

Regression models for functional responses and functional covariates, when there is a single response and a single predictor per subject, have been under intense development. Let $X(\cdot)$ be the predictor function of $s, s \in [0,S]$ and $Y(\cdot)$ be the response function of $t, t \in [0,T]$. To associate the response Y to the predictor X, depending upon the problem at hand, three possible models have been considered and developed.

1) When there is evidence that the response at current time t relates to the entire profile of the predictor, this model is well-knows as Functional linear model. For instance, Ramsay and Dalzell (1991); Ramsay and Silverman (2005); Scheipl et al. (2014); Wang (2014); Yao et al. (2005b) considered this relationship through $\int_0^S X(s)\beta(s,t)\,ds$ where $\beta(s,t)$ is the bivariate coefficient function and assumed to be smooth and square integrable. Among these studies, several approach proposed to estimate $\beta(s,t)$ from the data. For example, Ramsay and Dalzell (1991) used piecewise Fourier bases for $\beta(s,t)$, Ramsay and Silverman (2005) used

two series of basis functions such as splines to expand $\beta(s,t)$ and then using a functional version of normal equation to estimate $\hat{\beta}$, while Ferraty et al. (2012) proposed a functional version of Nadaraya-Watson estimate of the regression operator. Yao et al. (2005b) used functional principle component approach to expand functional response and functional covariate. They considered iid measurement errors and functional principle scores for $X(t)$ and $Y(t)$ were computed using Principal Analysis by Conditional Estimation (PACE) method (Yao et al., 2005a). More recently, Ivanescu et al. (2014) proposed a penalized regression when there are more than one covariate function and also additional scalar covariates in the model by applying a quadratic roughness penalties to avoid overfitting which is an extension of penalized functional regression (Goldsmith et al., 2012a), and Wang (2014) developed a linear mixed model using Expectation/Conditional Maximization Either (ECME) algorithm to maximize the log likelihood function.

2) Historical functional linear models relate the response at current time t to the covariate function observed on time-window with length Δ prior to t, say, $\int_{\max\{0,t-\Delta\}}^{t} X(s)\beta(s,t)ds$ (Malfait and Ramsay, 2003). Some regularization techniques such as basis truncation, roughness penalty, and LASSO were investigated in Harezlak et al. (2007) by using B-spline basis functions. Kim et al. (2011) suggested using functional principle component for both functional response and covariate and a preset basis function for the parameter function.

3) When the response at current time t relates to the predictor at time t and furthermore both response and predictor have been observed at same domain, one possible model that can build this relationship is called concurrent functional model or varying coefficient models (Ramsay and Silverman, 2005, Ch. 14) which we consider to our work in this paper. In this case $\beta(s,t) = \beta(t)$ and is a special case of the varying coefficient model introduced by Hastie and Tibshirani (1993). Fan and Zhang (2000a) suggested a two-step procedure for the parameter function which in the first step estimation the parameter of a pointwise regression and in the next step smoothing the pointwise estimations in order to estimate the parameter function. Huang et al. (2004) suggested to use B-spline bases with the regularization by the truncation in knot selection to represent the functional coefficient. In order to use functional varying coefficient model for longitudinal

data, Şentürk and Müller (2010) introduce a history index and assume that the value of functional response at time t, $Y(t)$ is predicted by the recent past of the predicted process (but not future and distance past). Şentürk and Nguyen (2011) develop a new estimation procedure for varying coefficient model based on covariance representation which appropriate for highly spare longitudinal data.

In contrast, regression models when both the response and predictors have the multilevel functional structure of the type described here, have received less attention. Crainiceanu et al. (2009) considered association models the case when there is scalar response per subject and the predictor is multilevel functional. Gertheiss et al. (2013); Goldsmith et al. (2012b) considered multiple scalar responses per subject and multilevel functional covariates, of the type described here.

In recent longitudinal studies data consists of a collection of functions/images for each subject and both the response and the covariate of interest are functional (Pomann et al., 2015). Furthermore the response values may be binary 0/1, or not necessarily normally distributed. In this paper we discuss association models when both the response and the predictors have a multilevel functional structure and are defined in the same domain. There are several sources of novelty of this framework: (1) proposal of generalized regression models for this multilevel functional setting; (2) relating the response to the structural components of the covariate. Gertheiss et al. (2013) discussed a regression model for functional predictor and scalar response that both are observed at multiple visit in a longitudinal case and relates the response to the structural component of the response.

Formally, consider the setting where for each subject $i = 1,\ldots,n$ we observe data of the form $[\{(Y_{ij}(t_{ijl}) : l\}, \ \{(X_{ij}(s_{ijp}) : p\}, \ \ j = 1,\ldots m_i]$ with $t_{ijl}, s_{ijp}\ in \mathscr{T}$; without loss of generality it is assumed that $\mathscr{T} = [0,1]$. For presentation simplicity we assume the time at which the response and the covariate are observed coincide and are the same across the subjects.

We assume that $X_{ij}(t)$ is a noisy measurement of the following underlying subject-specific and unit within subject-specific functional signals that $X_{ij}(t) = \mu(t) + Z_i(t) + U_{ij}(t) + \varepsilon_{ij}(t)$. Here, $\mu(\cdot)$ is the mean function, $Z_i(\cdot)$ is the subject specific deviation from the mean, $U_{ij}(\cdot)$ is the unit-specific deviation from the subject-

specific mean, and $\varepsilon_{ij}(\cdot)$ is noise. It is assumed that Z_i and U_{ij} are square integrable random curves in $\mathscr{T} = [0,1]$. To ensure identifiability we assume that the random processes Z_i, U_{ij} and ε_{ij} are uncorrelated, have mean zero and that ε_{ij} is a white process with covariance function $\text{cov}\{\varepsilon_{ij}(t), \varepsilon_{ij}(t')\} = \sigma_\varepsilon^2$, if $t = t'$ and 0 otherwise.

We are interested in association models that relate the current observation of the response at time t of the jth unit for the ith subject to the current value of the jth predictor at the same time t within the ith subject. Assume that given $Z_i(t)$ and $U_{ij}(t)$, the distribution of the response $Y_{ij}(t)$ follows the exponential family with smooth linear predictor $\eta(t)$ and dispersion parameter ϕ, i.e. $Y_{ij}(t) \sim EF(\eta(t), \phi)$, where

$$E[Y_{ij}(t)|Z_i(t), U_{ij}(t)] = g(\eta(t)) \quad \texttt{and} \quad \eta(t) = \beta_0(t) + \beta_1(t)Z_i(t + \beta_2(t)U_{ij}(t) \tag{1}$$

for known increasing link function $h = g^{-1}$.

The proposed regression model above, allows a concurrent relationship between the time point t of the two functional measurements. Concurrent relations between functional processes as a function of a third variable, which may be time, have been modeled via the varying coefficient model, first introduced by Cleveland et al. (1992); Hastie and Tibshirani (1993). Varying coefficient models have been widely used in the analysis of time dependent processes in the past decade (e.g., see Chiang et al., 2001; Fan and Zhang, 2000b, 2008; Hoover et al., 1998; Huang et al., 2004; Şentürk and Nguyen, 2011).

Nevertheless, our paper is different from others in the literature in the following aspects. 1) Even though there have been multiple proposals for the analysis of multilevel functional data (Baladandayuthapani et al., 2008; Crainiceanu et al., 2009; Di et al., 2009; Gertheiss et al., 2013; Guo, 2002; Li et al., 2007; Morris and Carroll, 2006; Morris et al., 2003, 2001; Staicu et al., 2010), a regression model relating multilevel functional response and predictors has not been proposed in the literature to the best of our knowledge. The proposed model relates multilevel functional variables via a regression model for a generalized response. 2) The proposed regression model provides a simple platform to separate the effects of different levels of the multilevel predictor on the response for the first time in literature, enabling easy interpretation and new insights into the building blocks of the regression model. 3) The proposed estimation algorithm based on method of moments

approaches allows for fast computation. Due partly to the simple forms proposed in separating the effects in regression and partly to the specific estimation procedure proposed, the algorithm remains fast and easy to implement. 4) Subject-specific predictions based on the proposed generalized time-varying regression are proposed utilizing functional principle components and best linear unbiased prediction (BLUP).

The rest of the paper is organized as follows. We begin by considering the model and the interpretation of the coefficients in Section 2. Estimation and Prediction procedure and possible options will be discussed in Section 3. We proceed with simulation studies in sections 4. Section 5 summarizes our conclusions and discussion.

2 Generalized time-varying regression for multilevel functional data

We propose the generalized time-varying linear models, a statistical framework to relate a generalized multilevel functional response to a multilevel functional predictor as model (1). The simple and practical decomposition of $X_{ij}(t)$ enables us to separate the different types of effects of the multilevel predictor on the response. This model is determined by parameter functions $\beta_0(t)$, $\beta_1(t)$ and $\beta_2(t)$, which are assumed to be square integrable on $\mathscr{T}$, in addition to the link function $g = h^{-1}$ which we assume that the link function $h(\cdot)$ is a monotone and twice continuously differentiable function with bounded derivatives and is thus invertible. Note that $\beta_0(t)$ is an intercept function which captures the variation in the response that does not depend on any of the covariate functions, $\beta_1(\cdot)$ is the time-varying effect of the subject-specific deviation Z_i, and $\beta_2(\cdot)$ is the effect of the unit-specific deviation U_{ij}. In the following we will try to make a road map for the estimation procedure.

2.1 Further model specification

Assuming that X_{ij} has small variation around its mean, as assumed similarly in Hall et al. (2008), implies that

$$Z_i(t) + U_{ij}(t) = \delta\{Z_i^*(t) + U_{ij}^*(t)\},$$

for a small constant δ and in addition, Z^* and U^* are Gaussian processes with zero mean and bounded covariance. By Taylor expansion of the function $g(\cdot)$ about $\beta_0(t)$ we have

$$g\Big(\beta_0(t)+\beta_1(t)Z_i(t)+\beta_2(t)U_{ij}(t)\Big)=g\Big(\beta_0(t)\Big)+\delta\Big\{\beta_1(t)Z_i^*(t)+\beta_2(t)U_{ij}^*(t)\Big\}g'\Big(\beta_0(t)\Big)+O_p(\delta^2). \quad (2)$$

It follows from (2) and also iterated expectation that $E\big\{Y_{ij}(t)\big\}=g\big(\beta_0(t)\big)+O(\delta^2)\equiv\mu_Y(t)$. In addition, it follows that for $j\neq j'$,

$$\begin{aligned} R(t,t') &\equiv cov\big\{Y_{ij}(t),X_{ij'}(t')\big\}=g'\big(\beta_0(t)\big)\beta_1(t)K_Z(t,t')+O(\delta^3), \quad \text{and} & (3)\\ Q(t,t') &\equiv E\Big[Y_{ij}(t)\big\{X_{ij}(t')-X_{ij'}(t')\big\}\Big]=g'\big(\beta_0(t)\big)\beta_2(t)K_U(t,t')+O(\delta^3), & (4)\end{aligned}$$

where $K_Z(t,t')=cov\big\{Z_i(t),Z_i(t')\big\}$ and $K_U(t,t')=cov\big\{U_{ij}(t),U_{ij}(t')\big\}$ denote the covariance functions of the processes Z_i and U_{ij}, respectively. We assume that consistent estimators for these covariance functions are available. For example, Di et al. (2009) could be used to obtain the consistent estimators of $K_Z(t,t')$ and $K_U(t,t')$; let $\widehat{K}_Z(t,t')$ and $\widehat{K}_U(t,t')$ denote such estimators for the two covariance functions. The equations (2)-(4) provide the intuition behind the estimation procedure. Specifically, we first find consistent estimators for $\mu_Y(t)$, $R(t,t')$ and $Q(t,t')$, say, $\widehat{\mu}_Y(t)$, $\widehat{R}(t,t')$ and $\widehat{Q}(t,t')$ respectively, and then obtain estimators for the coefficients functions as

$$\widehat{\beta}_0(t)=g^{-1}\{\widehat{\mu}_Y(t)\},\qquad \widehat{\beta}_1(t)=\frac{\widehat{R}(t,t)}{g'\{\widehat{\beta}_0(t)\}\widehat{K}_Z(t,t)}\ ,\quad \text{and}\quad \widehat{\beta}_2(t)=\frac{\widehat{Q}(t,t)}{g'\{\widehat{\beta}_0(t)\}\widehat{K}_U(t,t)}. \quad (5)$$

Our approach is based on method-of-moment estimators of $\widehat{\mu}_Y(t)$, $\widehat{R}(t,t')$ and $\widehat{Q}(t,t')$. More precisely, $\widehat{\mu}_Y(t)$ can be obtained using local linear smoothing of the aggregated data $\{(t,Y_{ij}(t)),i=1,\ldots,n;j=1,\ldots,m_i\}$. We use a penalized spline to the pooled data under independence to estimate the overall mean function of response curves and the smoothing parameter is selected via restricted maximum likelihood (REML). For the covariance of the R and Q functions we use a two-step method, to account for the smoothness of these functions. Firstly, let

$$\begin{aligned} \widetilde{R}(t,t') &= \left\{\sum_{j\neq j'}\Big(Y_{ij}(t)-\bar{Y}_{\cdot\cdot}(t)\Big)\Big(X_{ij'}(t')-\bar{X}_{\cdot\cdot}(t')\Big)\right\}\Big/\Big|\{(j\neq j')\}\Big|\ ,\\ \widetilde{Q}(t,t') &= \left\{\sum_{j\neq j'}Y_{ij}(t)\Big(X_{ij}(t')-X_{ij'}(t')\Big)\right\}\Big/\Big|\{(j\neq j')\}\Big| \end{aligned}$$

the method-of-moment estimators of R and Q, where $|S|$ denotes the cardinality of set S. In addition, $\bar{X}_{\cdot\cdot}(t)$ and $\bar{Y}_{\cdot\cdot}(t)$ are the raw mean functions for predictors and responses curves respectively and can be computed as follows:

$$\bar{X}_{\cdot\cdot}(t) = \sum_{i=1}^{n}\sum_{j=1}^{m_i} X_{ij}(t) / \sum_{i=1}^{n} m_i \ , \ \ \bar{Y}_{\cdot\cdot}(t) = \sum_{i=1}^{n}\sum_{j=1}^{m_i} Y_{ij}(t) / \sum_{i=1}^{n} m_i \ .$$

In step two, the smooth estimates of $\widetilde{R}(t,t')$ and $\widetilde{Q}(t,t')$ are obtained by applying a two-dimensional smoothing to the off-diagonal elements of the first step estimates (Yao et al., 2003); let $\widehat{R}(t,t')$ and $\widehat{Q}(t,t')$ denote the refined smooth estimates.

Next we focus on estimating the covariance functions of the processes Z_i and U_{ij}, say, $K_Z(t,t')$ and $K_U(t,t')$. For achieving a consistent estimator of these covariance functions, the approach used by Di et al. (2009); Staicu et al. (2010) is utilized. Let $K_T(s,t) = cov\{X_{ij}(s), X_{ij}(t)\}$ be the total covariance function, $K_B(s,t) = cov\{X_{ij}(s), X_{ij'}(t)\}$ be the between covariance function and $K_W(s,t) = \frac{1}{2}\Big\{cov\{[X_{ij}(s) - X_{ij'}(s)], [X_{ij}(t) - X_{ij'}(t)]\}\Big\}$ be the within covariance function, then we have

$$K_T(s,t) = K_Z(s,t) + K_U(s,t) + \sigma^2\delta_{st} \ , \ K_B(s,t) = K_Z(s,t) \ , \ \ K_W(s,t) = K_U(s,t) + \sigma^2\delta_{st} \tag{6}$$

where δ_{st} is the Kronecker delta that is equal to 1 if $t = s$, and 0 otherwise.

3 Estimation approach

The equations (6) provide the algorithm to achieve the estimates of $K_Z(s,t)$ and $K_U(s,t)$. After using the method-of-moment approach to find the raw estimates of $K_T(s,t)$ and $K_W(s,t)$, say, $\tilde{K}_T(s,t)$ and $\tilde{K}_W(s,t)$, a bivariate smoothing is applied to the off-diagonal elements of the raw estimates, denoted by $\hat{K}_T(s,t)$ and $\hat{K}_W(s,t)$. The smoothing parameters are selected via REML. Therefore, the estimation of $K_B(s,t)$ is given by

$$\hat{K}_B(s,t) = \hat{K}_T(s,t) - \hat{K}_W(s,t)$$

Keep in mind that a covariance function should be positive definite. So we need to check this property for all estimations of covariance functions especially for $\hat{K}_B(s,t)$ that is estimated as a difference of two

covariance functions. This problem can be solved utilizing the methods proposed by Hall et al. (2008); Yao et al. (2005a).This approach is based on trimming eigenvalues-eigenfunctions pairs where eigenvalues are negative.

Now we need to select the number of eigenfunctions which is an important practical problem in FPCA. Several alternative approaches have been investigated in the literature. Rice and Silverman (1991) used cross validation approach, Yao et al. (2005a) proposed using Akaik's information criterion(AIC), Di et al. (2009); Staicu et al. (2010) used a combination of the cumulate percent variance(CPV) and size of variance of principle components (SVPC), Greven et al. (2011) proposed likelihood ratio criteria, and Goldsmith et al. (2013) considered the CVP approach. We use the approach of the combination of CPV and SVPC which is used as follows:

$$K = \min\left\{k : \frac{\sum_{j=1}^{k}\lambda_j}{\sum_{j=1}^{N}\lambda_j} \geq P_1\ , \lambda_k < P_2\right\}$$

Where $P1$ represents the percent explained variance and P_2 indicates the minimum variance that could be discarded. These two thresholds can be chosen via simulation.

Once consist estimates of $\mu_Y(t)$, $R(s,t)$, $Q(s,t)$, $K_Z(s,t)$, and $K_U(s,t)$ are available, the coefficients $\beta_0(t)$, $\beta_1(t)$, and $\beta_2(t)$ will be estimated through formulas (5). Based on our experience from the simulation study, when there is no variation on Z_i and U_{ij} processes which usually this could be happened at the endpoints of the interval or rarely for a few time points at the interval, then the magnitude of $K_Z(t,t)$ and $K_U(t,t)$ in these time points would be very small or even zero. In that case, the estimates of $\beta_1(t)$ and $\beta_2(t)$ might be biased. To overcome this issue, a possible solution would be using an interpolation to find an appropriate approximate of $\beta_1(t)$ and $\beta_2(t)$ at these time points. This can be done using thresholds, say, c_Z and c_U that should be defined with respect to the magnitude of K_Z and K_U on the diagonal of the covariance matrices. For a threshold c_Z, let $\mathscr{T}'$ shows the set of these time points where $K_Z(t,t) < c_Z$ while $\mathscr{T}$ shows the all time points of t. A linear interpolation such as splines with interior knots $\mathscr{T}\setminus\mathscr{T}'$ of $\beta_1(t), t \in \mathscr{T}\setminus\mathscr{T}'$ is done and then this interpolation will be used to approximate the value of $\beta_1(t)$ for $t \in \mathscr{T}'$. This should be done in the same way for $\beta_2(t)$ with the threshold c_U and the covariance matrix $K_U(s,t)$.

We are of course investigating how the Taylor expansion of function $g(\cdot)$ about $\beta_0(t)$ in (2) depends on a small constant δ. In simulation study, we consider three different values of δ, namely, 0.5, 1 and 2. The results from the simulation are shown that the estimations are not depend on the value of δ. Also, the consistency of the estimators has been done through the simulation study.

3.1 Prediction

In this section we focus on prediction of the response profile based on the observed value of covariate variables. More precisely, for the our proposed approach, we are particularly interested in predicting about the functional response $Y_{ij}(t)$ for a new given signal $X_{ij}(t)$ with the multilevel structure. In this chapter we will discuss about the prediction and we will investigate the cases that response has a Gaussian and binary distribution.

Our approach is based on the decomposition of the observed curves as $X_{ij}(t) = \mu(t) + Z_i(t) + U_{ij}(t) + \varepsilon_{ij}(t)$, for some arbitrary time point $t \in \mathscr{T}$. Let $\widehat{K}_Z$ and $\widehat{K}_U$ be consistent estimators of the covariance functions of Z and U processes, and denote by $\widehat{\sigma}^2_\varepsilon$ the estimated of measurement noise variance. As mentioned earlier, such consistent estimators can be obtained by employing the methods proposed by Di et al. (2009). By having these estimates, we can use Mercer's theorem (Indritz, 1963) which provide the spectral decompositions of $\widehat{K}_Z(t,t')$ and $\widehat{K}_U(t,t')$ as follows:

$$\widehat{K}_Z(t,t') = \sum_{\ell \geq 1} \widehat{\lambda}_{Z,\ell} \widehat{\Phi}_{Z,\ell}(t) \widehat{\Phi}_{Z,\ell}(t') \quad , \quad \widehat{K}_U(t,t') = \sum_{\ell \geq 1} \widehat{\lambda}_{U,\ell} \widehat{\Phi}_{U,\ell}(t) \widehat{\Phi}_{U,\ell}(t'). \tag{7}$$

Where $\lambda_{Z,1} \geq \lambda_{Z,2} \geq \ldots > 0$, $\lambda_{W,1} \geq \lambda_{W,2} \geq \ldots > 0$ are eigenvalues and $\{\widehat{\Phi}_{Z,\ell}(t)\}_{\ell \geq 1}$ and $\{\widehat{\Phi}_{W,\ell}(t)\}_{\ell \geq 1}$ are orthogonal bases of eigenfunctions but are not mutually orthogonal. Due to dealing with infinite expansion, spectral decomposition of $\widehat{K}_Z(t,t')$ and $\widehat{K}_U(t,t')$ in (7) are impractical. So, we need to consider a finite dimensional approximation for these decompositions. Denote by N_Z and N_U the truncations of eigenfunctions for Z and U processes respectively.

In the next step we need to estimate the FPC scores or loadings. Several approaches have been discussed to estimate FPC scores. For instance, numerical integration or shrinkage estimator of PC scores (Yao et al.,

2003) can be used for standard functional data without and with measurement error, respectively. For sparse functional data, Goldsmith et al. (2013); Greven et al. (2010); Yao et al. (2005a) used conditional expectation to find the best linear unbiased predictions (BLUPs) of FPC scores. The latter approach also has been used by Di et al. (2009); Greven et al. (2010) in multilevel functional setting, although Di et al. (2009) also proposed Markov Chain Monte Carlo (MCMC). BLUPs have been used to estimate the FPC scores $\xi_{i\ell}$ and $\zeta_{ij\ell}$. It follows that the prediction for the ith subject-specific and jth unit-specific trajectories are $\widehat{Z}_i^{N_Z}(t) = \sum_{\ell=1}^{N_Z} \widehat{\xi}_{i\ell}\widehat{\Phi}_{Z,\ell}(t)$ and $\widehat{W}_{ij}^{N_U}(t) = \sum_{\ell=1}^{N_U} \widehat{\zeta}_{ij\ell}\widehat{\Phi}_{W,\ell}(t)$, respectively.

Once the coefficients functions $\beta_0(t)$, $\beta_1(t)$, $\beta_2(t)$, the ith subject-specific $Z_i(t)$, and the jth unit-specific $U_{ij}(t)$ are estimated, using plug-in estimates for (1), one can predict individual mean response trajectories in the generalized multilevel functional model by

$$\widehat{Y}_{ij}^{N}(t) = g\{\widehat{\beta}_0(t) + \widehat{\beta}_1(t)\widehat{Z}_i^{N_Z}(t) + \widehat{\beta}_2(t)\widehat{U}_{ij}^{N_U}(t)\}, \tag{8}$$

where $N = (N_Z, N_U)^T$ denotes the vector of truncations for processes Z and U.

4 Simulations

In this section, the simulation study was conducted to evaluate the proposed methodology with extensive simulations. Simulated data sets are constructed from the following model:

$$X_{ij}(t_{ijl}) = \mu(t_{ijl}) + Z_i(t_{ijl}) + U_{ij}(t_{ijl}) + \varepsilon_{ij}(t_{ijl})$$

and

$$E\left[Y_{ij}(t)|Z_i(t), U_{ij}(t)\right] = g\Big(\beta_0(t) + \beta_1(t)Z_i(t) + \beta_2(t)U_{ij}(t)\Big)$$

Where $\{i = 1, 2, \cdots, n\}$, $\{j = 1, 2, \cdots, m\}$ and $\{l = 1, 2, \cdots, N\}$ index the subject, the unit, and the number of measurements per the unit within the subject.

Let $K_Z(s,t) = \sum_{k=1}^{K_Z} \lambda_{Z,k}\phi_{Z,k}(s)\phi_{Z,k}(t)$ and $K_U(s,t) = \sum_{k=1}^{K_U} \lambda_{U,k}\phi_{U,k}(s)\phi_{U,k}(t)$ where $\phi_{Z,1}(t) = \sqrt{3}(2t-1)$, $\phi_{Z,2}(t) = \sqrt{5}(6t^2 - 6t + 1)$, $\phi_{U,1}(t) = \sqrt{2}\cos(\pi(2-.5)t)$, and $\phi_{U,2}(t) = \sqrt{2}\cos(\pi(3-.5)t$. We are of course

investigating the variation of FPC scores for both covariance functions by choosing three different values of λ_Z and λ_U as follows:

- $\lambda_{Z,k} = \lambda_{U,k} = (k-0.5)^{-2}\pi^{-2}$ for $k = 1,2$ and $\lambda_{Z,k} = \lambda_{U,k} = 0$ for $k \geq 3$.
- $\lambda_{Z,k} = \lambda_{U,k} = (k)^{-2}$ for $k = 1,2$ and $\lambda_{Z,k} = \lambda_{U,k} = 0$ for $k \geq 3$.
- $\lambda_{Z,k} = \lambda_{U,k} = \exp(-k)$ for $k = 1,2$ and $\lambda_{Z,k} = \lambda_{U,k} = 0$ for $k \geq 3$.

In this case, note that the level 1 eigenfunctions, $\{\phi_{Z,k}\}_{k\geq 1}$, and the level 2 eigenfunctions, $\{\phi_{U,k}\}_{k\geq 1}$, are not mutually orthogonal. We generated the FPC scores at levels 1, $\xi_{i,k}$ from $N(0,\lambda_{Z,k})$ for $k = 1,2$, so $Z_i(t) = \sum_{k=1}^{K_Z} \xi_{i,k}\phi_{Z,k}(t)$ and the FPC scores at level 2, $\zeta_{ij,k}$ from $N(0,\lambda_{U,k})$ for $k = 1,2$, so $W_{ji}(t) = \sum_{k=1}^{K_U} \zeta_{ij,k}\phi_{U,k}(t)$.

Now all requirements is ready to generate X from $X_{ij}(t) = \mu_X(t) + Z_i(t) + U_{ij}(t) + \varepsilon_{ij,X}(t)$, where $\varepsilon_{ij,X}(t)$ are independent and identically distributed from $N(0,\sigma)$. We consider that the true mean function as $\mu_X(t) = 2\sin(2\pi t)$ and so for each $i = 1,\ldots,30$ and $j = 1,\ldots,20$ the curves $X_{ij}(t)$ are observed at $N = 31$ equally spaced time points in $[0,1]$. As we called the title of the work as generalized time-varying regression for multilevel functional data, the procedure should be applied for all responses belong to the exponential family. In this simulation study we consider two responses, namely, Gaussian and Bernoulli variables. In Gaussian case, $Y_{ij}(t) = g\{\beta_0(t) + \beta_1(t)Z_i(t) + \beta_2(t)U_{ij}(t)\} + \varepsilon_{ij,Y}(t)$ where the inverse canonical link function is $g(x) = x$ and $\varepsilon_{ij,Y}(t) \overset{i.i.d}{\sim} N(0,\sigma)$ while in Bernoulli case with probability $g\{\beta_0(t)+\beta_1(t)Z_i(t)+\beta_2(t)U_{ij}(t)\}$, the inverse canonical link function would be $g(x) = \exp(x)/(1+\exp(x))$. Finally, the true functional coefficients are taken as follows: $\beta_0(t) = \sin(2\pi t)$, $\beta_1(t) = -\sin(1.5\pi t)$ and $\beta_2(t) = \sin(3\pi t)$.

We present numerical results for these noise levels: $\sigma = 0.05, 0.25$ and 0.5. So, by having three different values for eigenvalues λ_Z, λ_U, and two kind of response, Gaussian and Bernoulli, this gives a total of 18 designs. For each design, 100 datasets are generated. Figure 1 shows a random sample with size 80 of a dataset. Left panel shows covariate signals and right panel shows the corresponding response curves of the dataset. Two thresholds P_1 and P_2 discussed in Section 2, were taken 0.99 and $\frac{1}{N} = \frac{1}{31}$ respectively. We used the proposed method to estimate the functional coefficients $\beta_0(t), \beta_1(t)$, and $\beta_2(t)$ and other desire criteria

that will be discussed in the next section.

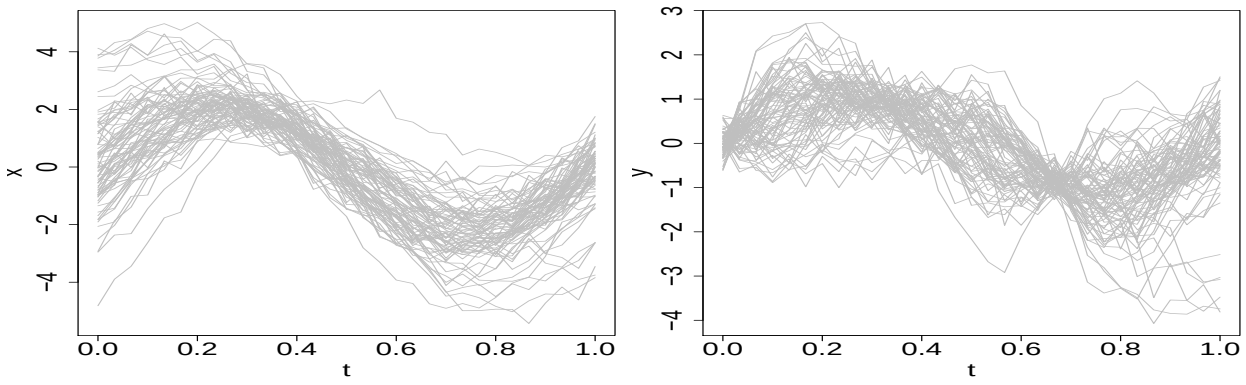

Figure 1: A random sample with size 80 of the simulated data in Gaussian case. Left: covariate signals, Right: the response profile.

4.1 Results

Here we discuss the results of the simulation study for both cases, Gaussian and Bernoulli. First, we are interested in getting an overview of the variability in estimates of the coefficient functions. Figure 2 illustrates the estimation of the coefficients functions in model (1) based on 100 fitted the proposed approach in Gaussian case, $g(x) = x$, with $n = 30$, $m = 20$ and $N = 31$ for three different amount of eigenvalues but for fixed values of measurement error, namely, $\sigma = .25$. Shown in the left panel of this figure is the estimation of the coefficient function $\beta_0(t)$ for each simulation in gray color, the true function $\beta_0(t) = \sin(2\pi t)$ in blue color, and the mean function of all estimations in red color. Displayed in the middle and right panel are the same as the left panel but for the coefficient functions $\beta_1(t)$ and $\beta_2(t)$, respectively. The realization show the same patterns to the corresponding true coefficient functions. The variability of $\beta_2(t)$ is smaller than two other coefficient function, i.e., $\beta_0(t)$ and $\beta_1(t)$. As it was expected estimates of $\beta_1(t)$ and $\beta_2(t)$ are biased around time points that $\hat{K}_Z(t,t)$ and $\hat{K}_U(t,t)$ are close to zero or even zero. This indeed was can be seen for the estimates of $\beta_2(t)$ closed to the end of interval [0,1]. In that case, the possible proposed approach for this problem discussed in the end of Section 2 was conducted to overcome this issue and the result of it cab be seen on estimate of $\beta_2(t)$ for one or two time points at the end of the interval $[0,1]$. Increasing eigenvalues

$\lambda_{Z,k}$ and $\lambda_{U,k}$ amounts to large variability in the estimates but not much fo $\beta_2(t)$. The estimates of the coefficients functions for the binary response based on 100 fitted the proposed approach for the simulated data is shown in Figure 3.

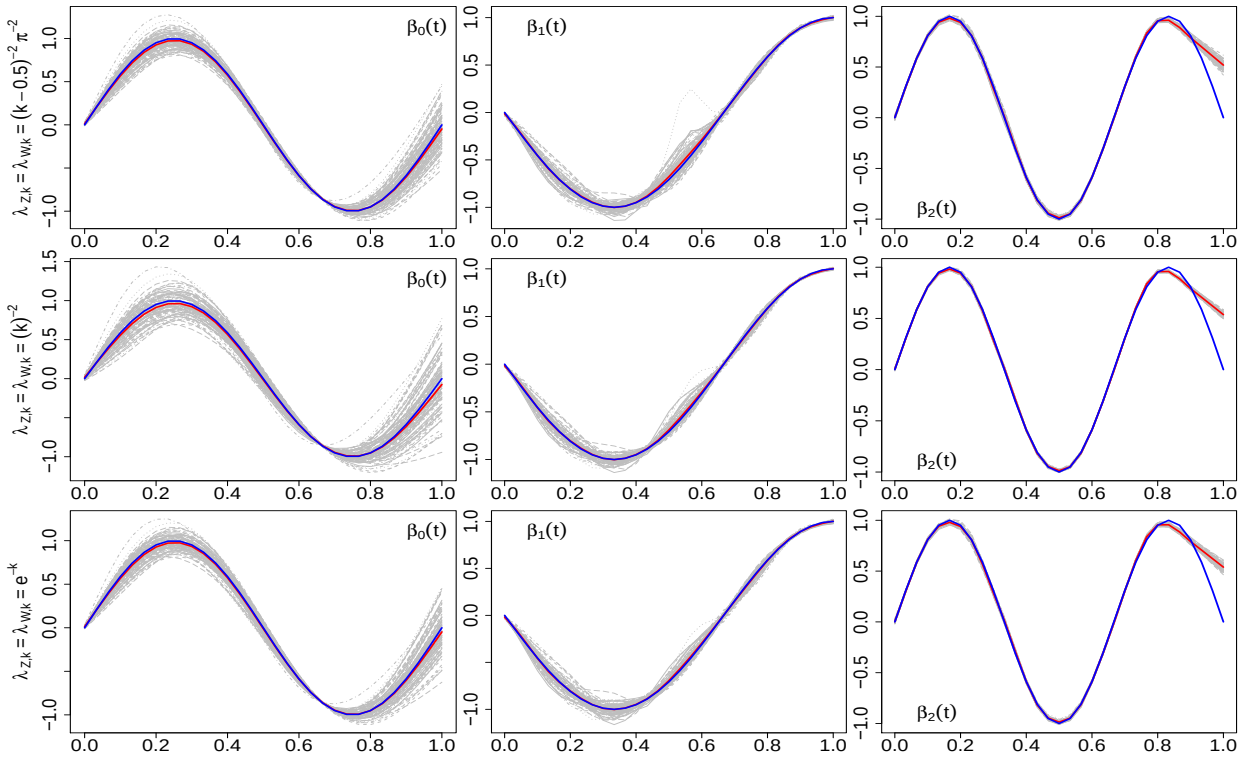

Figure 2: Estimated regression functions of simulated dataset for Gaussian distribution of the response : $\beta_0(t)$ (left), $\beta_2(t)$ (center) and $\beta_2(t)$ (right). Red curves correspond to that means of estimated coefficients and blue ones are the true coefficient functions.

Then, let us to turn to other results of the simulation study. For evaluating the proposed approach we need to use some criteria. Corresponding the approach and the quantities used in the model, we indeed focus at these intuitive deviation criteria, namely, Integrated Mean Squared Error (IMSE), Mean Error (ME), and Integrated prediction Error (IPE):

$$\text{IMSE}(\beta_i(t)) = \frac{1}{N.simu}\int\{\beta_i(t)-\hat{\beta}_i(t)\}^2dt, \qquad \text{ME}_i = \frac{\int\{\beta_i(t)-\hat{\beta}_i(t)\}^2dt}{\int\beta_i^2(t)dt}, \qquad i=1,2,3.$$

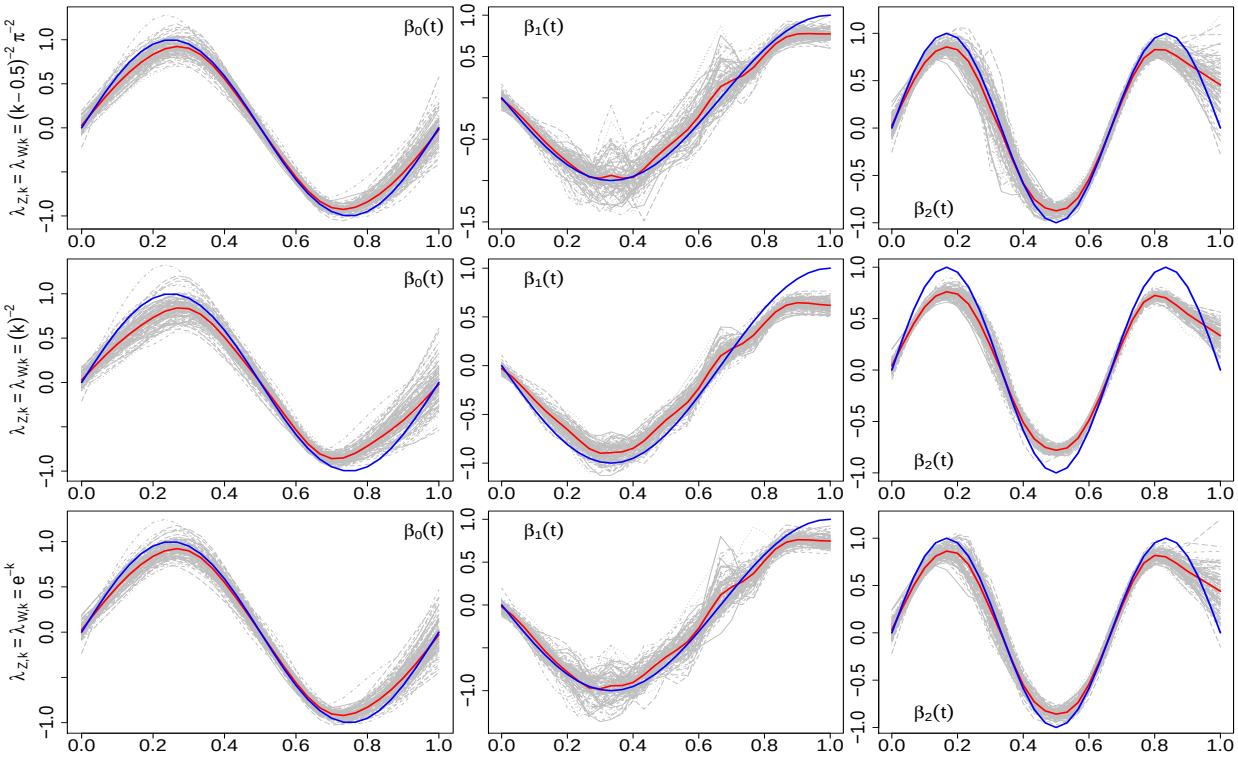

Figure 3: Estimated regression functions of simulated dataset for Bernoulli distribution of the response : $\beta_0(t)$ (left), $\beta_2(t)$ (center) and $\beta_2(t)$ (right). Red curves correspond to that means of estimated coefficients and blue ones are the true coefficient functions.

$$\text{IPE}(Y) = \frac{1}{N.simu}\int \{y_{ij}(t) - \widehat{y}_{ij}(t)\}^2 dt, \qquad \text{ME}_{\mu_Y} = \frac{\int [g^{-1}\{\mu_Y(t)\} - g^{-1}\{\hat{\mu}_Y(t)\}]dt}{\int [g^{-1}\{\mu_Y(t)\}]^2 dt}$$

$$\text{ME}_{K_Z} = \frac{\int \{K_Z(t,t) - \widehat{K}_Z(t,t)\}^2 dt}{\int K_Z^2(t,t)dt}, \qquad \text{ME}_{K_U} = \frac{\int \{K_W(t,t) - \widehat{K}_W(t,t)\}^2 dt}{\int K_W^2(t,t)dt},$$

$$\text{ME}_R = \frac{\int \{R(t,t) - \widehat{R}(t,t)\}^2 dt}{\int R^2(t,t)dt}, \qquad \text{ME}_Q = \frac{\int \{Q(t,t) - \widehat{Q}(t,t)\}^2 dt}{\int Q^2(t,t)dt}.$$

Percentiles of the deviation measures are explained in Table 1 and 2 for Gaussian and Bernoulli distribution of the response, respectively. Table 3 displays the deviation criteria computed from the simulation study for both Gaussian and Bernoulli responses but for the same scenario, namely, $\lambda_{Z,k} = \lambda_{U,k} = \exp(-k)$ and $\sigma = 0.5$. As it was expected, the criteria for the continuous case is smaller than binary.

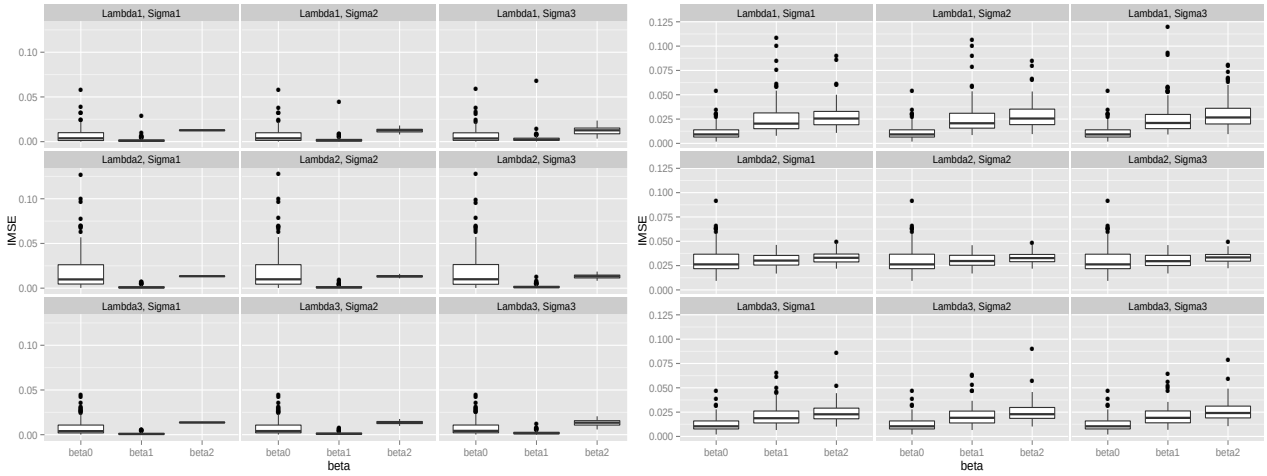

Figure 4: Box plot of IMSE criterion computed for the estimates of coefficients functions for all possible scenarios. Left panel corresponds to the Gaussian response while right panel relates to the Binary response.

5 Discussion

In this paper, we introduced a statistical framework in order to make an association model in multilevel functional data of the type subject-unit when both response and covariate have functional form. The conditional distribution is supposed to follow the exponential family and in this paper, simulation study has been conducted for Gaussian and binary cases. To make the association, a simple practical decomposition of the functional covariate based on subject-specific and unit within subject-specific is used and then different type of effects of the multilevel functional covariate on multilevel functional response were estimated. From computational point of view, the methods for estimating the within and between covariance have been modified by using optimized functions in R packages.

In multilevel functional data when there is dependency between the unit within the subject, for instance the spatial dependency, then it would be expedient to consider this dependency in decomposition of the covariate as well as the model should be extended for these kind of data in an appropriate way .

References

Baladandayuthapani, V., Mallick, B. K., Young Hong, M., Lupton, J. R., Turner, N. D., and Carroll, R. J. (2008). Bayesian hierarchical spatially correlated functional data analysis with application to colon car-

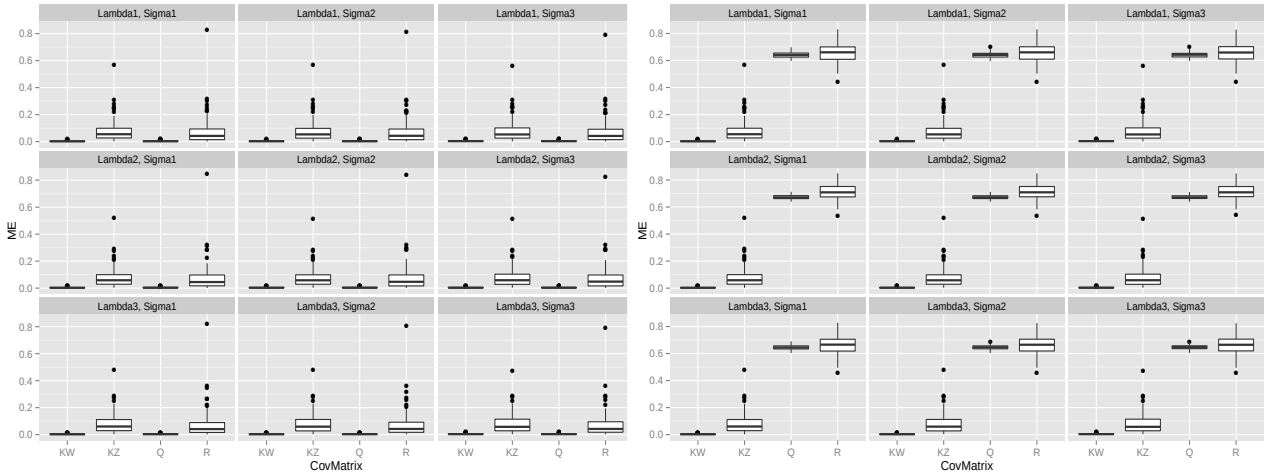

Figure 5: Box plot of ME criterion computed for the estimates of some quantities used in the proposed approach for all possible scenarios. Left panel corresponds to the Gaussian response while right panel relates to the Binary response.

cinogenesis. *Biometrics*, 64(1):64–73.

Chiang, C.-T., Rice, J. A., and Wu, C. O. (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *Journal of the American Statistical Association*, 96(454):605–619.

Cleveland, W. S., Grosse, E., and Shyu, W. M. (1992). Local regression models. *Statistical models in S*, pages 309–376.

Crainiceanu, C. M., Staicu, A.-M., and Di, C.-Z. (2009). Generalized multilevel functional regression. *Journal of the American Statistical Association*, 104(488):1550–1561.

Di, C.-Z., Crainiceanu, C. M., Caffo, B. S., and Punjabi, N. M. (2009). Multilevel functional principal component analysis. *The annals of applied statistics*, 3(1):458.

Fan, J. and Zhang, J.-T. (2000a). Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(2):303–322.

Fan, J. and Zhang, W. (2000b). Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scandinavian Journal of Statistics*, 27(4):715–731.

Table 1: Percentiles of the deviation measures presented are estimated from 100 simulations where $n = 30$, $m = 20$ and $g(x) = x$. Here σ_1, σ_2 and σ_3 stand for 0.05 , 0.25 and 0.5 respectively. While λ_1, λ_2 and λ_3 stand for $\lambda_{Z,k} = \lambda_{U,k} = (k-.5)^{-2}\pi^{-2}$, $\lambda_{Z,k} = \lambda_{U,k} = k^{-2}$ and $\lambda_{Z,k} = \lambda_{U,k} = e^{-k}$ respectively.

(σ, λ)	ME_0			ME_1			ME_2			ME_{μ_Y}	ME_{K_Z}	ME_{K_U}	ME_R	ME_Q
	Med	25%	75%	Med	25%	75%	Med	25%	75%	Med	Med	Med	Med	Med
(σ_1, λ_1)	.008	.003	.021	.002	.001	.004	.026	.026	.027	.008	.055	.002	.042	.003
(σ_1, λ_2)	.020	.010	.054	.001	.001	.003	.028	.027	.028	.020	.058	.002	.045	.003
(σ_1, λ_3)	.008	.004	.022	.001	.001	.003	.028	.028	.029	.008	.060	.003	.041	.003
(σ_2, λ_1)	.008	.003	.021	.003	.001	.005	.026	.022	.029	.008	.053	.003	.043	.003
(σ_2, λ_2)	.020	.009	.054	.002	.001	.003	.027	.026	.029	.020	.058	.003	.047	.003
(σ_2, λ_3)	.008	.004	.022	.002	.001	.004	.028	.026	.030	.008	.059	.003	.042	.003
(σ_3, λ_1)	.007	.003	.020	.004	.003	.008	.026	.018	.031	.007	.053	.004	.042	.003
(σ_3, λ_2)	.021	.009	.055	.002	.001	.004	.027	.024	.030	.021	.058	.003	.048	.003
(σ_3, λ_3)	.009	.004	.022	.003	.002	.005	.028	.022	.032	.009	.058	.004	.042	.004

Fan, J. and Zhang, W. (2008). Statistical methods with varying coefficient models. *Statistics and its Interface*, 1(1):179.

Ferraty, F., Van Keilegom, I., and Vieu, P. (2012). Regression when both response and predictor are functions. *Journal of Multivariate Analysis*, 109:10–28.

Gertheiss, J., Goldsmith, J., Crainiceanu, C., and Greven, S. (2013). Longitudinal scalar-on-functions regression with application to tractography data. *Biostatistics*, page kxs051.

Goldsmith, J., Bobb, J., Crainiceanu, C. M., Caffo, B., and Reich, D. (2012a). Penalized functional regression. *Journal of Computational and Graphical Statistics*.

Goldsmith, J., Crainiceanu, C. M., Caffo, B., and Reich, D. (2012b). Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(3):453–469.

Goldsmith, J., Greven, S., and Crainiceanu, C. (2013). Corrected confidence bands for functional data using principal components. *Biometrics*, 69(1):41–51.

Greven, S., Crainiceanu, C., Caffo, B., and Reich, D. (2011). Longitudinal functional principal component analysis. In *Recent Advances in Functional Data Analysis and Related Topics*, pages 149–154. Springer.

Table 2: Percentiles of the deviation measures presented are estimated from 100 simulations where $n = 30$, $m = 20$ and $g(x) = \exp(x)/(1+\exp(x))$. Here σ_1, σ_2 and σ_3 stand for 0.05 , 0.25 and 0.5 respectively. While λ_1, λ_2 and λ_3 stand for $\lambda_{Z,k} = \lambda_{U,k} = (k-.5)^{-2}\pi^{-2}$, $\lambda_{Z,k} = \lambda_{U,k} = k^{-2}$ and $\lambda_{Z,k} = \lambda_{U,k} = e^{-k}$ respectively.

(σ, λ)	ME_0			ME_1			ME_2			ME_{μ_Y}	ME_{K_Z}	ME_{K_U}	ME_R	ME_Q
	Med	25%	75%	Med	25%	75%	Med	25%	75%	Med	Med	Med	Med	Med
(σ_1, λ_1)	.019	.014	.029	.041	.030	.063	.053	.040	.068	.019	.055	.002	.660	.640
(σ_1, λ_2)	.054	.045	.076	.060	.051	.071	.068	.059	.076	.054	.058	.002	.708	.671
(σ_1, λ_3)	.022	.016	.033	.037	.028	.052	.047	.038	.060	.022	.060	.003	.666	.645
(σ_2, λ_1)	.019	.014	.029	.041	.031	.062	.053	.040	.073	.019	.053	.003	.660	.642
(σ_2, λ_2)	.054	.045	.076	.059	.051	.071	.068	.060	.075	.054	.058	.003	.709	.671
(σ_2, λ_3)	.022	.016	.033	.038	.028	.052	.047	.038	.062	.022	.059	.003	.665	.646
(σ_3, λ_1)	.019	.014	.029	.042	.030	.060	.055	.041	.075	.019	.053	.004	.659	.643
(σ_3, λ_2)	.054	.045	.076	.059	.050	.071	.069	.061	.075	.054	.058	.003	.709	.671
(σ_3, λ_3)	.022	.016	.033	.038	.028	.052	.050	.039	.064	.022	.058	.004	.665	.647

Table 3: Deviation criteria computed from the for simulated data sets with the same value for $\lambda_{Z,k} = \exp(-k)$ and $\sigma = 0.5$ for both cases.

	$IMSE_{\beta_0}$	$IMSE_{\beta_1}$	$IMSE_{\beta_2}$	ME_{β_0}	ME_{β_1}	ME_{β_2}	ME_{μ_Y}	ME_{K_Z}	ME_{K_U}	ME_R	ME_Q
Gaussian	0.008	0.002	0.013	0.017	0.004	0.028	0.017	0.080	0.005	0.073	0.005
Bernoulli	0.012	0.021	0.026	0.026	0.043	0.054	0.026	0.080	0.005	0.663	0.647

Greven, S., Crainiceanu, C., Caffo, B., Reich, D., et al. (2010). Longitudinal functional principal component analysis. *Electronic Journal of Statistics*, 4:1022–1054.

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

Hall, P., Müller, H.-G., and Yao, F. (2008). Modelling sparse generalized longitudinal observations with latent gaussian processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):703–723.

Harezlak, J., Coull, B. A., Laird, N. M., Magari, S. R., and Christiani, D. C. (2007). Penalized solutions to functional regression problems. *Computational statistics & data analysis*, 51(10):4911–4925.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 757–796.

Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85(4):809–822.

Huang, J. Z., Wu, C. O., and Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica*, 14(3):763–788.

Indritz, J. (1963). *Methods in analysis*. Macmillan New York:.

Ivanescu, A. E., Staicu, A.-M., Scheipl, F., and Greven, S. (2014). Penalized function-on-function regression. *Computational Statistics*, pages 1–30.

Kim, K., Şentürk, D., and Li, R. (2011). Recent history functional linear models for sparse longitudinal data. *Journal of statistical planning and inference*, 141(4):1554–1566.

Li, Y., Wang, N., Hong, M., Turner, N. D., Lupton, J. R., and Carroll, R. J. (2007). Nonparametric estimation of correlation functions in longitudinal and spatial data, with application to colon carcinogenesis experiments. *The Annals of Statistics*, pages 1608–1643.

Malfait, N. and Ramsay, J. O. (2003). The historical functional linear model. *Canadian Journal of Statistics*, 31(2):115–128.

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

Morris, J. S., Vannucci, M., Brown, P. J., and Carroll, R. J. (2003). Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis. *Journal of the American Statistical Association*, 98(463):573–583.

Morris, J. S., Wang, N., Lupton, J. R., Chapkin, R. S., Turner, N. D., Young Hong, M., and Carroll, R. J. (2001). Parametric and nonparametric methods for understanding the relationship between carcinogen-induced dna adduct levels in distal and proximal regions of the colon. *Journal of the American Statistical Association*, 96(455):816–826.

Pomann, G.-M., Sweeney, E. M., Reich, D. S., Staicu, A.-M., and Shinohara, R. T. (2015). Scan-stratified case-control sampling for modeling blood–brain barrier integrity in multiple sclerosis. *Statistics in medicine*.

Ramsay, J. O. and Dalzell, C. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 539–572.

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

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 233–243.

Scheipl, F., Staicu, A.-M., and Greven, S. (2014). Functional additive mixed models. *Journal of Computational and Graphical Statistics*, (just-accepted):00–00.

Şentürk, D. and Müller, H.-G. (2010). Functional varying coefficient models for longitudinal data. *Journal of the American Statistical Association*, 105(491):1256–1264.

Şentürk, D. and Nguyen, D. V. (2011). Varying coefficient models for sparse noise-contaminated longitudinal data. *Statistica Sinica*, 21(4):1831.

Staicu, A.-M., Crainiceanu, C. M., and Carroll, R. J. (2010). Fast methods for spatially correlated multilevel functional data. *Biostatistics*, 11(2):177–194.

Wang, W. (2014). Linear mixed function-on-function regression models. *Biometrics*, 70(4):794–801.

Yao, F., Müller, H.-G., Clifford, A. J., Dueker, S. R., Follett, J., Lin, Y., Buchholz, B. A., and Vogel, J. S. (2003). Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate. *Biometrics*, 59(3):676–685.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005a). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590.

Yao, F., Müller, H.-G., Wang, J.-L., et al. (2005b). Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, 33(6):2873–2903.

IV

Analysis of juggling data: Registration subject to biomechanical constraints

Anders Tolver
Department of Mathematical Sciences
University of Copenhagen

Helle Sørensen
Department of Mathematical Sciences
University of Copenhagen

Martha MullerŞentürk
Department of Mathematical Sciences
University of Copenhagen

Seyed Nourollah Mousavi
Department of Mathematical Sciences
University of Copenhagen

Publication details

Published (2014).

Electronic Journal of Statistics
Vol. 8 (2014) 1856–1864
ISSN: 1935-7524
DOI: 10.1214/14-EJS937F

Analysis of juggling data: Registration subject to biomechanical constraints*

Anders Tolver, Helle Sørensen, Martha Muller and Seyed Nourollah Mousavi

Department of Mathematical Sciences
University of Copenhagen
e-mail: tolver@math.ku.dk; helle@math.ku.dk; m.muller@math.ku.dk; nourollah@math.ku.dk

Abstract: We illustrate how physical constraints of a biomechanical system can be taken into account when registering functional data from juggling trials. We define an idealized model of juggling, based on a periodic joint movement in a low-dimensional space and a periodic position vector (from an undefined joint to the finger tip) of approximately constant length along the observed trajectory. Our registration procedure first warps the cycles in the trial to each other and computes a periodic average, and then estimates the joint movement and the position vector of the abovementioned model.

Keywords and phrases: Biomechanical constraints, decomposition, functional data analysis, juggling trajectories, periodic average, registration, warping.

Received August 2013.

1. Introduction

Functional data are often unsynchronized in their raw form, either due to the sampling process or due to random phase variation (or both). This makes analysis on the raw data problematic since, for example, cross-sectional sample statistics can be misleading. Registration is the process of mapping unsynchronized curves into a synchronized class of functions, with the purpose of effectively filtering out noise before subsequent statistical analyses [1].

At best, registration should use any knowledge of the data generating system, in particular the shape of the underlying signal as well as the nature of possible pertubations. In this paper we discuss registration for functional data from juggling, taking into account simple biomechanical considerations.

Ideally, biomechanics of juggling may be described mathematically by nonlinear dynamical systems, but feedback and feedforward motor control mechanisms are necessary to overrule any disturbed dynamics and thereby impose desired movements or dynamics. We consider data from juggling cycles within in trial as pertubated versions of an idealized periodic movement. The periodic curve represents the average dynamics of the juggling process, whereas the deviations

*Main article 10.1214/14-EJS937.

between the observed data and the idealized signal reflect the complex feedback mechanism between the brain and the motor control system [4].

In conceptualizing an appropriate idealized mathematical model of human juggling, we consider the creation of an electromechanical juggling robot. How would we build and program such a robot? As a minimum, we would construct a rotating finger or hand limb and attach it with a joint to a fixed bar (representing an arm). We could conveniently label the two ends of the hand limb as 'finger tip' and 'joint'.

As a first attempt, we keep the position of the joint fixed and let the position vector from joint to finger tip be periodic. Regarded from a fixed external coordinate frame the position of the finger tip of the robot would trace a trajectory described by

$$f(t) = f_0(t) + c_0$$

where $c_0 \in \mathbb{R}^3$ corresponds to the fixed position of the joint and $f_0 : I \to \mathbb{R}^3$ is the periodic position vector function. Assuming that the robot is a rigid body introduces the geometric constraint that f_0 has constant length, d, such that $|f_0(t)| = d$ for all $t \in I$.

The juggling robot can be improved by allowing the position of the joint to follow a periodic curve. This gives a decomposition of the form

$$f(t) = f_0(t) + c_0(t), \tag{1}$$

where $c_0 : I \to \mathbb{R}^3$ is the trajectory of the joint, while f_0 still describes the vector from joint to finger tip and satisfies $|f_0(t)| = d$ for all $t \in I$ for some d. For identification purposes we assume that c_0 has a simple structure meaning that it belongs to a lower dimensional function space.

In this paper, decompositions of the type (1) will be regarded as idealized juggling signals, and we will demonstrate how to register the observed data towards such idealized signals, i.e. demonstrate that is it is possible to warp and filter the juggling trials such that the resulting curves allow a decomposition of the form (1).

Sections 2 and 3 give a complete description of the registration procedure and details about implementation. In Section 4 we display the results of applying the procedure to the ten trials from the juggling data. Finally, in Section 5 we evaluate the perspectives of combining phase registration and biomechanical constraints.

2. Data and registration procedure

The pre-processed data [2] (lightly smoothed, centered, rotated and trimmed) is the starting point of our analysis, and is referred to as "observed data" or "raw data" in the remainder of the paper. The data indicate the position of the right index finger during juggling and is thus composed of three coordinates. We write $f(t) = (f_1(t), f_2(t), f_3(t))$, and let n denote the number of cycles. There are 10 signals/trials, all collected from the same person. The number of cycles per trial varies from 11 to 13.

The suggested registration procedure is applied to each trial separately, but on all three dimensions and all cycles simultaneously. The implementation details are described in Section 3, but, in short, the complete procedure is split into three steps:

1. **Warping** The observed signal consisting of several cycles is converted into a warped version $f \circ h$, where cycles are warped towards each other using a periodic average function as target for the registration procedure.
2. **Averaging** Based on the warped signal, $f \circ h$, a periodic average, denoted by $\mathcal{P}f$, is computed as a projection onto the (high-dimensional) space of periodic functions.
3. **Decomposition** The periodic average $\mathcal{P}f$ is decomposed into two periodic terms: a joint movement $\mathcal{J}$ belonging to a low-dimensional space, V, and a remainder $\mathcal{P}f - \mathcal{J}f$ with approximately constant length along the trajectory.

The complete procedure involves estimation of a warping function h, a periodic average, and a joint movement $\mathcal{J}f$. Notice that $\mathcal{P}f$ and $\mathcal{J}f$ are periodic per construction, and thus have no between-cycle variation. In particular, we only need to plot the curves on the interval corresponding to one cycle. On the other hand, the warped, but not averaged, curve $f \circ h$ may potentially show amplitude variation between cycles, but presumably only little phase variation, since that has been diminished by warping.

The second step involves projection onto a space of periodic three-dimensional functions. If this projection is denoted by Q_{per}, then $\mathcal{P}f = Q_{per}(f \circ h)$. If $\|\cdot\|$ is the standard L^2-norm and g is a three-dimensional curve, then

$$\frac{\|Q_{per}g\|}{\|g\|} = \sqrt{\frac{\|g\|^2 - \|g - Q_{per}g\|^2}{\|g\|^2}} = \sqrt{1 - \frac{\|g - Q_{per}g\|^2}{\|g\|^2}} \tag{2}$$

takes values in $[0, 1]$ and is a natural measure of the degree of periodicity in g. When data from different cycles are warped against each other as in step 1, we would expect a larger degree of periodicity compared to the raw data. Hence, comparison of $\frac{\|Q_{per}f\|}{\|f\|}$ and $\frac{\|Q_{per}(f\circ h)\|}{\|(f\circ h)\|}$ can be used to quantify the effect of warping on periodicity (see Section 4).

3. Implementation

This section describes technical details of the implementation of our registration procedure. The emphasis is on the decomposition step, since warping and averaging rely on existing techniques and software.

Let f denote a signal consisting of n complete juggling cycles. The duration of each cycle within a trial is rescaled to $[0, 1]$, then the same implementation can be used for all trials, even though the number of cycles are different.

Warping First, we expressed f in terms of 201 Fourier basis functions, and computed the orthogonal projection f_{per} on the space of periodic functions $L_{per,n}$ containing n replications of the same signal. Due to the Fourier basis

representation this amounts to keeping coefficients corresponding to harmonics of order $n, 2n, 3n, \ldots, Kn$ (where K is the largest K such that $Kn \leq 100$). Second, a time warping function h maximizing the coherence between $f \circ h$ and f_{per} was estimated. We used the minimal eigenvalue of a cross-product matrix with a roughness penalty on curvature of h as estimation criterion, see [3, Section 7.6]. In order to ensure a sufficient degree of smoothness of the warped signal $f \circ h$ we restricted h to the space spanned by 101 B-splines of order 5 with equally spaced break points. The roughness of the warping functions were controlled by penalizing the squared integral of second order derivatives. The robostness to the value of the penalty parameter λ was examined and for the results presented below we used $\lambda = 10^{-11}$ based on visual inspection.

Averaging The warped function $f \circ h$ was projected onto $L_{per,n}$ (see the paragraph on the warping step above). Hence, we obtain a periodic average of $f \circ h$, denoted $\mathcal{P}f$ and spanned by periodic harmonics.

Decomposition To implement the estimation of $\mathcal{J}f$ in step 3 it was convenient to expand all functions in terms of orthogonal complex exponentials. Denoting by a_k and $b_k, k = 1, 2, 3$, the three coordinate functions of the periodic average $\mathcal{P}f$ (known) and joint movement $\mathcal{J}f$ (to be estimated), we have expansions

$$a_k(t) = \sum_{j=-m}^{m} a_{k,j} \exp(i\omega jt), \quad b_k(t) = \sum_{j=-l}^{l} b_{k,j} \exp(i\omega jt)$$

and hence

$$a_k'(t) = \sum_{j=-m}^{m} i\omega j a_{k,j} \exp(i\omega jt), \quad b_k'(t) = \sum_{j=-l}^{l} i\omega j b_{k,j} \exp(i\omega jt).$$

Here $\omega = 2\pi n$ where n is the number of cycles.

We emphasize that $\mathcal{P}f$ has already been expressed in a finite Fourier basis, thus m and $a_{k,j}$ are all fixed and known at this point of the analysis, whereas the coefficients b_k should be estimated. For $l < m$ fixed, we collect the unknown parameters in θ:

$$\theta = \{b_{k,j} | k = 1, 2, 3, j = -l, \ldots, l\}$$

Some comments on the choice of l: The regularization assumption $l < m$ is necessary for identification, i.e., for the decomposition (1) to be unique since otherwise we could just let $\mathcal{J}f = \mathcal{P}f - c_0$ with $c_0 \in \mathbb{R}^3$ any fixed vector. For $l < m$ the joint movement $\mathcal{J}f$ belongs to a subspace of lower dimension than $\mathcal{P}f$, and the idea is to choose a small l, such that the joint movement is simple.

Recall that we aim at finding $\mathcal{J}f$ such that $\mathcal{P}f - \mathcal{J}f$ has approximately constant length; hence we want the derivative of the squared length to be approximately zero for all t:

$$D|\mathcal{P}f(t) - \mathcal{J}f(t)|^2 \approx 0.$$

This leads to the following criterion function to be minimized:

$$C(\theta) = \int_0^1 \left[D|\mathcal{P}f(t) - \mathcal{J}f(t)|^2\right]^2 dt \tag{3}$$
$$= \int_0^1 \left[D\sum_{k=1}^{3}(a_k(t) - b_k(t))^2\right]^2 dt$$
$$= 4\int_0^1 \left[\sum_{k=1}^{3} D(a_k(t) - b_k(t)) \cdot (a_k(t) - b_k(t))\right]^2 dt.$$

If we introduce the notation $e_{k,j} = a_{k,j} - b_{k,j}$ (with $b_{k,j} = 0, |k| > l$) for the Fourier coefficients of the difference $\mathcal{P}f - \mathcal{J}f$, and furthermore $c_{j_1,j_2} = \{\sum_{k=1}^{3} j_2 e_{k,j_1} e_{k,j_2}\}$ and let $j \in I_s$ if $j, s-j \in \{-m, \ldots, m\}$, then

$$C(\theta) = \int_0^1 \left[\sum_{s=-2m}^{2m} i\omega \sum_{j\in I_s} c_{s-j,j} \exp(i\omega st)\right]^2 dt.$$

Finally, if we let $d_s = \sum_{j\in I_s} c_{s-j,j}$ and use that $d_{-s} = -\overline{d_s}$ (complex conjugate), then we end up with the following simple formula for the criterion function

$$C(\theta) = -4\omega^2 \sum_{s=-2m}^{2m} d_s d_{-s} = 4\omega^2 \left\{|d_0|^2 + 2\sum_{s=1}^{2m} |d_s|^2\right\}. \tag{4}$$

The representation (4) makes it feasible to compute numerically the value and the gradient of the objective function as a function of θ to be used for the minimization algorithm. Since we are looking for a real valued estimate of the joint movement $\mathcal{J}f$, we found it convenient to reparameterize the problem in terms of a basis of sines and cosines. For the results below we used $l = 1$ corresponding to the joint movement being expressed in terms of first order harmonics only.

4. Results

We applied the registration procedure described above to each of the ten juggling trials. We will use trial 8 for detailed illustration, because the effect of the warping step was largest for this trial.

Warping and averaging Figure 1 shows the effect of steps 1 and 2 (warping and averaging) on trial 8. The vertical coordinate (z)of the raw data (dashed) is shown together with vertical coordinate of the periodic signal $\mathcal{P}f$ (solid). The raw signal does not exhibit much misalignment but the signal is indeed warped slightly. Notice how the warping is more pronounced towards the ends of the trial. The average curve $\mathcal{P}f$ for trial 8 is shown for each coordinate separately in the left part of Figure 2, and as a 3d-curve in the right part of the figure (solid curve).

For the raw data the degree of periodicity, cf. definition (2), was 88.0%, whereas for the warped data this number increased to 98.6%. All other trials had degrees of periodicity of 94.3% to 97.2% before warping and between 97.5%

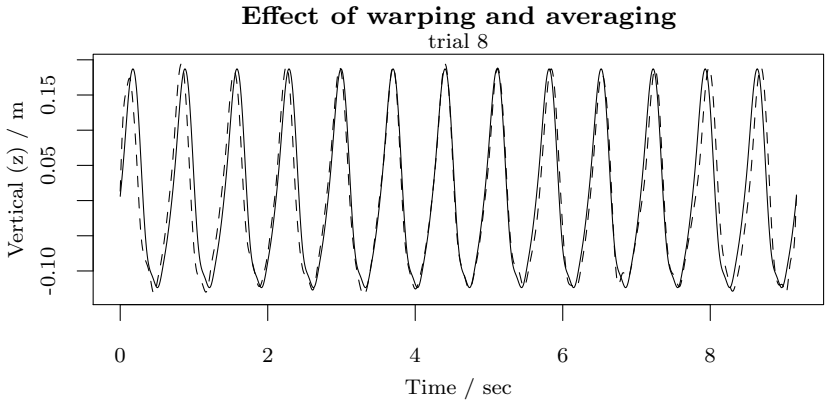

FIG 1. *Warping and averaging for trial 8. The dashed curve shows the z coordinate of the observed data, while the solid curve shows the z coordinate of* $\mathcal{P}f$.

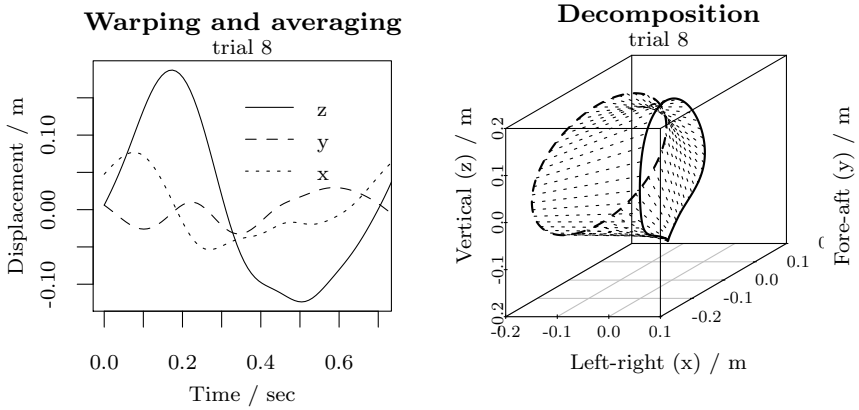

FIG 2. *Left: The three directions of the warped and averaged curve* $\mathcal{P}f$ *for trial 8. Right: 3d-illustration of the decomposition for trial 8. The solid curve shows the average* $\mathcal{P}f$*, the dashed curve shows the estimated joint movement curve* $\mathcal{J}f$*, and the dotted lines illustrate the trajectory of the difference* $\mathcal{P}f - \mathcal{J}f$ *(each dotted line correspond to a specific time point.)*

and 99.2% after warping. Hence, in general, only a limited amount of warping towards the periodic template was necessary. Visually, the raw and averaged trials were almost indistinguishable, except for trial 8 (see Figure 1).

The upper left, upper right and lower left plots of Figure 3 show the three coordinates of the warped curves $f \circ h$ for all ten trials, split into cycles and rescaled to the unit interval. The curves are coloured according to trial (but note that curves from different trials have not been aligned). In general, cycles within a trial are well aligned. Therefore the projection onto $L_{per,n}$ is a good representation of a trial. Note that the projections are similar across trials (-see the lower right part of Figure 3). The warping criterion gives less weight to coordinates with lower amplitude variation. This may explain why most misalignment is present in the y direction.

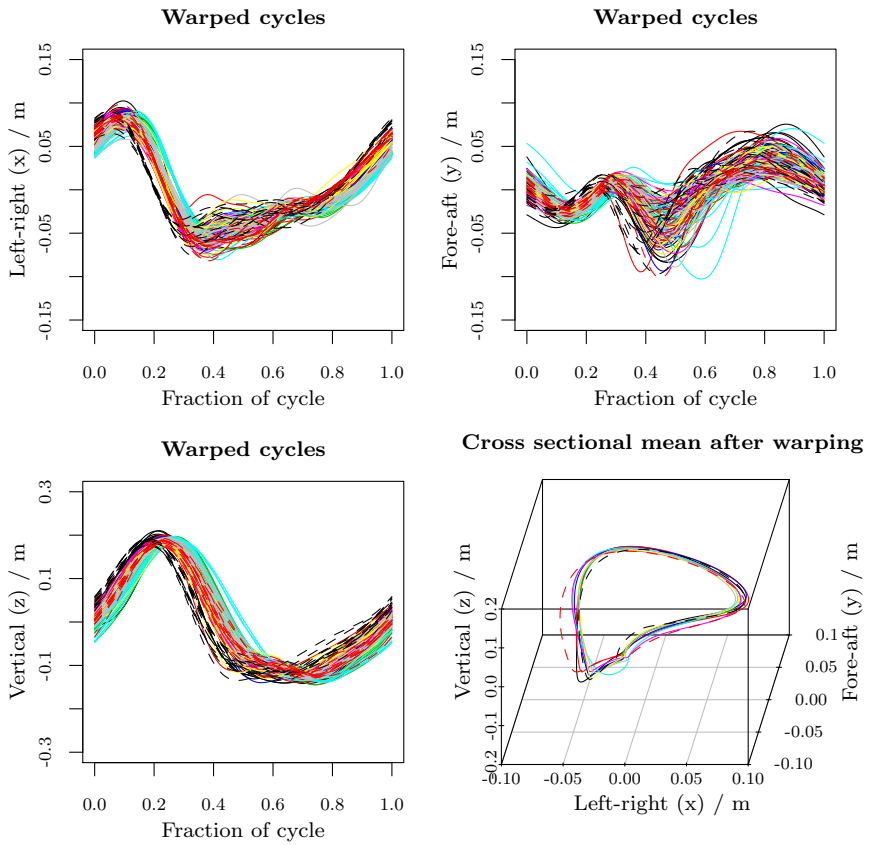

FIG 3. *Upper left, upper right and lower left: The three coordinates of the warped curves $f \circ h$ cut into individual cycles for each trial. For a trial with n cyc les, the complete curve was simply divided into n pieces of the same length, which was then rescaled to the unit interval. Cycles of the same colour and line type stem from the same trial. Lower right: 3d-scatterplot of the periodic average $\mathcal{P}f$ for all trials.*

Decomposition The estimated joint movement $\mathcal{J}f$ for trial 8 is shown as a dashed curve in the right part of Figure 2. Recall that the estimation procedure seeks the curve $\mathcal{J}f$ such that the vector $\mathcal{P}f - \mathcal{J}f$ has approximately constant length over the trajectory. This vector is illustrated by the dotted lines between the two curves, and its length varies from 0.179 m to 0.182 m for trial 8.

The decompositions for all curves are illustrated in Figure 4. The left part shows the length $|\mathcal{P}f - \mathcal{J}f|$ over the trajectories (scaled to the unit interval), and the right part shows the joint movements $\mathcal{J}f$. We make the following immediate observations from Figure 4: First, for all ten trials it was possible to obtain a function $\mathcal{J}f \in V$ such that the distance $|\mathcal{P}f - \mathcal{J}f|$ is approximately constant over time. This indicates that our simplistic biomechanical considerations leading to equation (1) characterizes some of the main features of the data generating mechanism. Second, the estimated length varies from 0.077 m

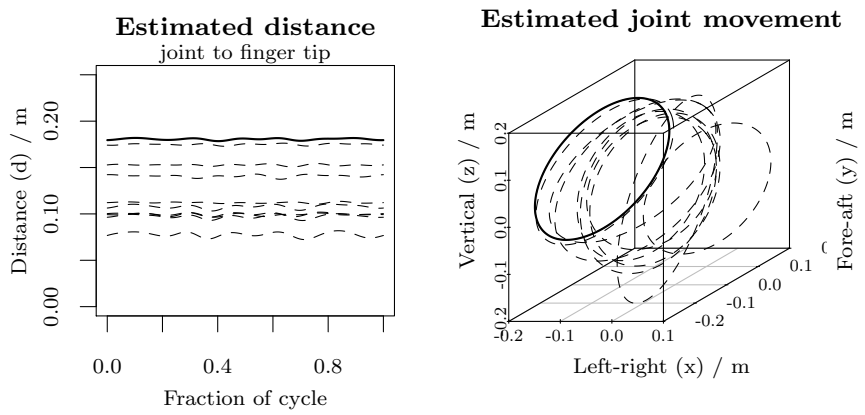

FIG 4. *Left: Estimated trajectory of distances, $|\mathcal{P}f - \mathcal{J}f|$, for all 10 trials. Right: Estimated joint movement, $\mathcal{J}f$, for all ten trials. In both plots the estimate correponding to trial 8 is shown as a solid curve.*

.

to 0.181 m across the ten trials. This is somewhat disappointing as we had hoped for an interpretation of this length as the length of a part of the hand or arm of the juggler. Third, the variation between the estimated joint movement curves is substantial. The decomposition restricts $\mathcal{J}f$ to be spanned by first order harmonics in all three directions. Allthough the curves are approximately elliptic they are different regarding angle and position.

5. Discussion

The purpose of the paper was to illustrate how the physical nature of a biomechanical system could be taken into account when removing phase variation of functional data from juggling. We have demonstrated that it is possible to warp all ten juggling trials such that the resulting structural mean over all cycles allows a decomposition as in (1).

The most striking observation is that the estimated distance from finger tip to joint, which should be an internal constant of the body anatomy, varies substantially across the ten trials. This complicates the physical interpretation of the estimated decomposition. Looking more carefully at the curves in the left part of Figure 4, there seems to be some common patterns in the deviations from constancy. Curves with low values of d seem to have peaks and valleys at the same time points (for example around 0.38 and 0.82), i.e. at the same time points of the juggling cycle. This indicates that our simple model might not have captured all features in the data.

A possible extension of the model would be to allow for more flexibility in the space V for the joint movement, i.e. by introducing harmonics of higher order in the basis for $\mathcal{J}f$. However, it seems more likely that adjustments from the idealized set-up given by (1) is taking place around the finger tip (far from the corpus) rather than at joints closer to the corpus. This suggest to relax the

focus on constant length of $\mathcal{P}f - \mathcal{J}f$. For example, the criterion function $C(\theta)$ in the decomposition step, see (3) and (4), could be adjusted to have a time-varying penalty on deviations from constancy. This would, however, complicate the optimization problem substantially.

In this connection, it should be mentioned that the numerical optimization problem for estimating the decomposition was more challenging than expected. The algorithm we used produced reliable estimates but was slow. This part of the implementation could be improved.

It is important to realize that amplitude and phase variation are bound to be intertwined, as an adjustment via a change in speed (phase) will most likely also change the amplitude. In relation to this, the complicated interplay between the estimation the warping function (step 1) and the estimation of the joint movement (step 3) should also be noticed. In particular, the space V for the joint movement is not invariant to warping (i.e. $g \in V$ does not imply that $g \circ h \in V$ for a warping function h). Too much warping of f may destroy the interpretation of the decomposition. This could be avoided by simultaneously estimating the warping function and the decomposition, i.e. to incorporate the warping (and averaging) step into the decomposition step.

Apart from the suggestions mentioned above, it would be interesting to examine the robustness of the registration. Simulations could clarify the importance of the explicit form of the underlying signal on the performance of the registration procedure. Moreover, it would be interesting to fit a common joint movement curve to all s, and see the effect on the corresponding position vectors $\mathcal{P}f - \mathcal{J}f$ and their lengths.

Acknowledgements

We acknowledge the Mathematical Biosciences Institute, Ohio, for supporting our participation in the workshop on "Statistics of Time Warpings and Phase Variations".

References

[1] Kneip, A. and Ramsay, J. O. (2008). Combining registration and fitting for functional models. *Journal of the American Statistical Association* **103**, 483, 1155–1165. http://amstat.tandfonline.com/doi/abs/10.1198/016214508000000517. MR2528838

[2] Ramsay, J. O., Gribble, P., and Kurtek, S. (2014). Description and processing of functional data arising from juggling trajectories. *Electron. J. Statist.* **8**, 1811–1816, Special Section on Statistics of Time Warpings and Phase Variations.

[3] Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*, Second ed. Springer, New York. MR2168993

[4] Schaal, S., Atkeson, C. G., and Sternad, D. (1996). One-handed juggling: A dynamical approach to a rhythmic movement task. *Journal of Motor Behavior* **28**, 2, 165–183.