Published on Statistics in Medicine (2013), 32 (30), 5222-5240.
Publisher's version available at http://onlinelibrary.wiley.com/doi/10.1002/sim.5989/abstract

An introduction with medical applications to functional data analysis

Helle Sørensen[a*‡], Jeff Goldsmith[b], Laura M. Sangalli[c]

Abstract

Functional data are data that can be represented by suitable functions, such as curves (potentially multi-dimensional) or surfaces. This paper gives an introduction to some basic but important techniques for the analysis of such data, and the techniques are applied to two datasets from biomedicine. One dataset is about white matter structures in the brain in multiple sclerosis patients; the other dataset is about three-dimensional vascular geometries collected for the study of cerebral aneurysms. The techniques described are smoothing, alignment, principal component analysis, and regression.

Keywords: curve alignment; functional principal component analysis; functional regression; smoothing

1 Introduction

Functional data are generated from underlying continuous functions. Each observation consists of discrete measurements $x_1, \ldots, x_n$ taken at time or location points $s_1, \ldots, s_n$, but these data points are assumed to arise from a (smooth) function X such that x_j is an observation of $X(s_j)$. Interest is in the functions as such rather than the individual measurements, and therefore differs from traditional multivariate analysis in both conceptual framework and statistical tools used for analysis. For example, the medical questions to be answered could be "How is the function x associated with the severity of a disease and with survival from the disease?"; "Which subjects exhibit similar patterns of an expression profile and which subjects exhibit different patterns?"; or "What is the general growth pattern of a tumor, and how do we estimate an intrinsic time scale for each individual?" Functional data analysis enables us to answer such questions regarding regression and prediction, clustering, and variation in the time/space direction, as well as many others.

This paper is a review of some of the techniques developed for functional data. It is built around two datasets from medical science: a dataset of three-dimensional vascular geometries collected for the study of cerebral aneurysms, and a dataset on white matter structures in the brain in multiple sclerosis patients.

The illustrating datasets are just two among many examples of functional data in the medical sciences. Technical development over the last few decades has resulted in equipment that produces vast amounts of data that are functional in nature, such as measurements over fine time or space grids and images with many pixels. Examples include data on electrical

[a]Laboratory for Applied Statistics, Department of Mathematical Sciences, University of Copenhagen, Denmark
[b]Department of Biostatistics, Columbia Mailman School of Public Health, Columbia University, USA
[c]MOX — Dipartimento di Matematica, Politecnico di Milano, Italy.
*Correspondance to: Helle Sørensen, Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, DK-2100 Copenhagen East, Denmark
‡Email: helle@math.ku.dk

activity of the heart (electrocardiography or ECG), data on electrical activity along the scalp (electroencephalography or EEG), continuous activity monitoring through accelerometers, motor control after stroke, chromatograms and other spectral analysis, growth curves, data reconstructed from medical imaging, expression profiles in genetics and genomics. Functional data are also important in analytical chemistry, biomechanics, plant science, engineering and many other fields.

Appropriate statistical methods have been developed along with the appearance of functional data. Work was already initiated by the 60's and 70's, but development accelerated in the 80's and 90's as functional data was more widely produced and recognized. The first edition of [1] made the methods available to a larger audience and has had an enormous impact on the spread of functional data analysis. The book mainly covers parametric (and semi-parametric) approaches and explorative methods, and has been accompanied by a more applied version covering roughly the same topics [2]. Other important books are [3] and [4] on non-parametric methods, and [5] with emphasis on hypothesis tests in models for functional data. In addition to these monographs, is a vast quantity of scientific papers ranging from theoretical to applied. Papers that are relevant for our applications will be mentioned in subsequent sections.

The aims of a functional data analysis are generally the same as for other statistical analyses, and there exist functional versions of many standard statistical methods. The challenges with functional data lie in the infinite-dimensional nature of the data, the implicit assumption of smoothness, and the extra variability in the time direction, among others. There are close connections to longitudinal data, since in practice the data consists of repeated measures on each subject. Classical longitudinal data analysis often models the expected values as explicit low-dimensional functions of time (polynomial or non-linear), whereas functional approaches allow for greater flexibility. Recent work has blurred the distinction between longitudinal and functional data settings by considering the former as sparse observations of continuous functions; such analyses are less flexible than for dense observations but emphasize conceptually functional interpretations; see [6] and [7, Part III].

There are also close connections to analysis of high-dimensional data. A naive approach for functional data would be to use functional values at a large number of grid points as input to a multivariate analysis. However this would not take the serial structure (of time or space) into account, and thus may lose power by neglecting inherent structure or result in non-smooth estimates. Hence, methods must be adapted to the functional setting.

The aim of this paper is to make researchers in the medical sciences aware of the challenges of and possibilities for functional data. It is intended for researchers who are familiar with classical statistical disciplines like regression, classification, principal component analysis, but who do not necessarily have experience with functional data. Sections 2–5 cover important themes in functional data analysis. Our selection of topics has been motivated by the two illustrative datasets, and is by no means exhaustive. In each section, the challenges and possible analysis techniques are outlined, and the methods are applied to one or both datasets. There are few mathematical and technical details, but references are given to appropriate papers with more detailed information.

Section 2 is about smoothing, which brings discrete data to functional form. Section 3 is about curve alignment (also called warping or registration), i.e. methods for separating phase and amplitude variation. Section 4 describes the functional version of principal component analysis used to extract the leading components of the data. Functional regression, describing the association between variables some of which are functional, is the topic of Section 5. In the end, there is a short section on software (Section 6) and some concluding remarks (Section 7).

1.1 DTI tract profile data

Our first motivating dataset comes from a neuroimaging study of multiple sclerosis (MS) patients [8, 9]. MS is an immune-mediated inflammatory disease marked by symptomatic attacks that are associated with the demyelnating lesions. These lesions can affect white-matter tracts, which are bundles of myelinated axons (the projections of nerve cells that propagate electrical signals). Damage to the myelin sheath protecting white matter axons disrupts the transmission of signals in the central nervous system and can result in severe patient disability. To noninvasively quantify white matter microstructure we use diffusion tensor imaging (DTI), an MRI-based modality that traces the diffusion of water in the brain [10, 11, 12]. Specific white matter tracts, such as the corticospinal tracts (which connect the motor cortex to the body) and the corpus callosum (which connects the two hemispheres of the brain), can be identified and studied using full-brain DT images.

For several major white-matter tracts, we have one-dimensional functional summaries called tract profiles. These profiles localize white matter properties as measured by water diffusion along the anatomical path of the tract through the brain. In this data set, tract profiles form the basis of our analysis. Figure 1 shows the position of the corticospinal tracts in the brain, as well as the fractional anisotropy tract profiles for the right corticospinal tract with two subjects highlighted. Fractional anisotropy measures how strongly directional is the diffusion of water, with values near one indicating complete anisotropy and values near zero indicating completely isotropic diffusion.

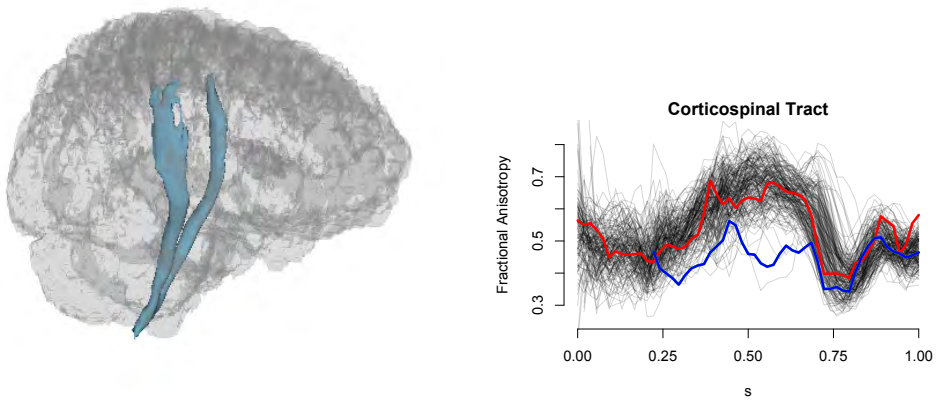

Figure 1: DTI data. Left: Position of the corticospinal tracts (blue) in the brain. Right: Fractional anisotropy tract profiles for the right corticospinal tract, with two subjects highlighted.

Our dataset consists of measurements on 150 MS patients and 36 control subjects, obtained under an institutional review board-approved protocol. For each subject we observe tract profiles as described above and MS case status; additional scalar measures of cognitive performance are observed for MS patients only. Two major problems arise in analyzing this dataset: 1) data compression, keeping in mind potential issues with respect to measurement error and missing data (some subjects are missing up to 20% of the tract profile, as shown in Figure 1); and 2) scalar-on-function regression to understand associations between anatomical structures summarized in tract profiles and both case status and cognitive performance.

1.2 AneuRisk65: three-dimensional vascular geometry data

The second illustrative dataset, AneuRisk65, comes from the AneuRisk project, a scientific endeavor that has gathered together researchers of different scientific fields, ranging from neurosurgery and neuroradiology to statistics, numerical analysis and bio-engineering, with the aim of studying the pathogenesis of cerebral aneurysms. These are deformations of the cerebral arteries (see Figure 2, left panel), rather frequent in the adult population and normally not disrupting. On the other hand, rupture of a cerebral aneurysm is a rare event, but with very high mortality. Cerebral aneurysms of the vessels are believed to be caused by complex interactions between the biomechanical properties of artery walls and effects of hemodynamical forces exerted on the vessel walls, such as wall shear stress and pressure; the hemodynamic forces in turn depend on the vessel morphology itself. In particular, it has been conjectured that the pathogenesis of these deformations is influenced by the morphological shape of cerebral arteries, through the effect that the morphology has on the hemodynamics. For this reason, one of the main goals of the AneuRisk project has been the study of relationships between vessel morphology and aneurysm presence and location. See the AneuRisk webpage http://mox.polimi.it/it/progetti/aneurisk/ and references therein.

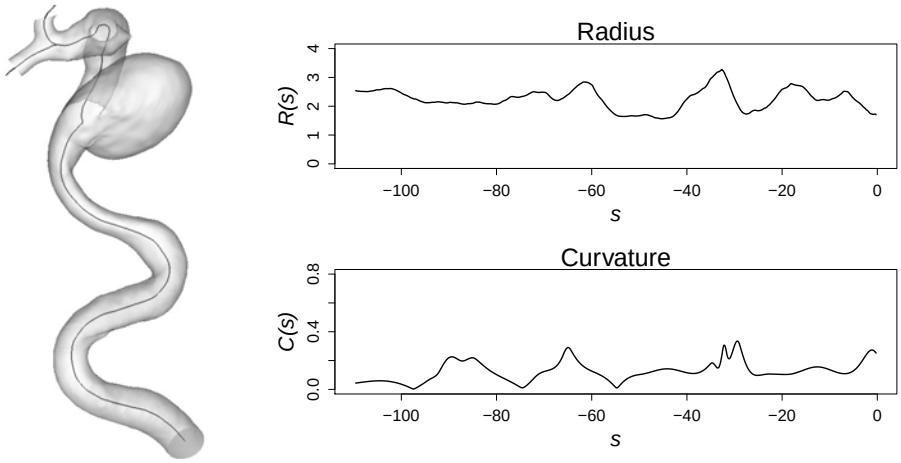

Figure 2: AneuRisk65 data. Left: Three-dimensional image of an Internal Carotid Artery (ICA) with an aneurysm (subject 1), obtained from an angiography via the reconstruction algorithm coded in the Vascular Modeling ToolKit VMTK (see Section 2). The black line inside the vessel is its reconstructed centerline. Right: Radius profile (top) and curvature profile (bottom) of the ICA. The curvature is computed using the smoothing technique described in Section 2.4. Location s along the ICA centerline is measured as a curvilinear abscissa that goes from the terminal bifurcation of the ICA towards the heart (measured in mm).

The deformation of the vessel wall may originate along one of the Internal Carotid Arteries (ICA, in short), two large arteries bringing blood to the brain; or at the terminal bifurcation of the ICA; or after the bifurcation, in the so-called Willis Circle. Each of the two ICAs sits for most of its length outside the skull, along the neck, surrounded by muscle tissues; just before its terminal bifurcation it enters inside the skull, passing through a dural ring (i.e., a hole in the skull bone). Arteries downstream of ICA terminal bifurcation float in

The project involved MOX Laboratory for Modeling and Scientific Computing (Dip. di Matematica, Politecnico di Milano), Laboratory of Biological Structure Mechanics (Dip. di Ingegneria Strutturale, Politecnico di Milano), Istituto Mario Negri (Ranica), Ospedale Niguarda Ca' Granda (Milano) and Ospedale Maggiore Policlinico (Milano), and has been supported by Fondazione Politecnico di Milano and Siemens Medical Solutions Italia.

the brain humor, inside the skull. For this reason, aneurysms located at or after the ICA terminal bifurcation are more life-threatening; the rupture of one such aneurysm is fatal in most cases.

The AneuRisk65 dataset is based on a set of three-dimensional angiographic images taken from 65 subjects who were suspected of being affected by cerebral aneurysms. Most of the subjects were in fact found to have an aneurysm at or after the terminal bifurcation of the ICA, or along the ICA; only a small subgroup had no visible aneurysms. The statistical analyses conducted within the AneuRisk project have focused on the geometrical features of the ICA, which is clearly recognizable in each of the 65 angiographies, in particular on the ICA radius and curvature profiles (see Figure 2, right panel). Radius and curvature are in fact known to highly influence the local hemodynamics, and hence, through this effect, they may influence aneurysms' pathogenesis. See, e.g., [13, 14, 15].

2 From discrete-time measurements to functions

Functional data analysis deals with data where each sample unit is thought of as a function. In practice, however, the data consist of discrete-time measurements or measurements on a grid, and the sequence of observed values are commonly observed with noise. Hence, one step in a functional data analysis often consists of converting the discretely observed data to smooth functions. Moreover, it is often of interest to study also the derivatives of the functional data; in some cases, the derivatives are themselves objects of analysis, while in other cases estimates of curve derivatives may be of help for further processing of the data, such as in curve alignment procedures (see Section 3). In this section we discuss methods that bring the data from discrete to functional form; the resulting curves are used as input to subsequent analyses.

2.1 Smoothing with least squares and penalized least squares

Let us consider data from I subjects. For subject i, we observe W_{ij} with $j = 1, \ldots, n_i$ and W_{ij} corresponding to time/location point s_{ij}. Notice the flexibility in the set-up: the number of measurements as well as the observation points may differ between subjects (not all n_i are the same, and s_{ij} may differ from $s_{i'j}$); the measurements are not necessarily measured at equidistant points; and the data could be sparse or dense at the subject level. Additionally, a common assumption in functional data analysis is that W_{ij} is a noisy observation of $X_i(s_{ij})$:

$$W_{ij} = X_i(s_{ij}) + \epsilon_{ij},$$

most often with the implicit assumption that ϵ_{ij}s are iid $N(0, \sigma^2)$. The domain of X_i will be suppressed in the notation; in particular in definitions of derivatives and in integrals.

Our problem is to estimate or reconstruct the true functions X_i $(i = 1, \ldots, I)$ from the observed data W_{ij} $(i = 1, \ldots, I, j = 1, \ldots, n_i)$. One popular solution described in this section is to use basis expansions and impose smoothness either by restricting the basis or through explicit smoothness constraints. A basis expansion is a linear combination of known basis functions; hence we choose K basis functions $\psi_1, \ldots, \psi_K$ and consider estimates of the form

$$\hat{X}_i(s) = \sum_{k=1}^{K} \hat{c}_{ik} \psi_k(s), \tag{1}$$

The coefficients $\hat{c}_{i1}, \ldots, \hat{c}_{iK}$ are estimated from the data as described below, and the basis functions $\psi_1, \ldots, \psi_K$ can potentially differ between subjects. Notice that $\hat{X}_i$ is defined for all s, not only those where observations are available. The expansion also provides expressions of the derivatives ($\hat{X}_i'$, $\hat{X}_i''$, etc.) in terms if the derivatives of the basis functions; for example,

the first and second order derivatives are

$$\hat{X}_i'(s) = \sum_{k=1}^{K} \hat{c}_{ik}\psi_k'(s), \quad \hat{X}_i''(s) = \sum_{k=1}^{K} \hat{c}_{ik}\psi_k''(s). \tag{2}$$

Common choices of bases are Fourier bases and polynomial spline bases, both illustrated in Figure 3 (see also, e.g., [1]). Notice that Fourier basis functions (left plot) are periodic, which makes them particularly useful for periodic data, and that spline basis functions (right plot) can be chosen to have local support, which makes them computationally very efficient and good at capturing local features of the functional data.

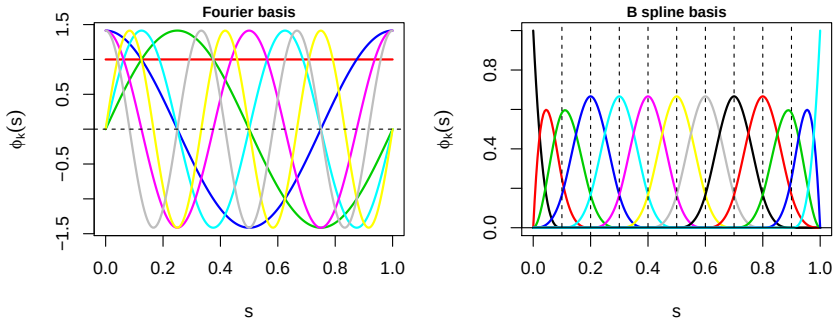

Figure 3: Left: Fourier basis with $K = 7$ basis functions (one of them constant and equal to 1). Right: B-spline basis of order 4 (cubic polynomials) with $K = 13$ basis functions; the vertical lines show the 9 equally spaced interior spline knots.

The estimation problem amounts to choosing the coefficients, c_{ik}. For a fixed i, let $W_i = (W_{i1}, \ldots, W_{in_i})^T$ be the column of observations, and Ψ_i be the $n_i \times K$ matrix with entry (j, k) equal to $\psi_k(s_{ij})$. For a column of coefficients $c_i = (c_{i1}, \ldots, c_{iK})$ the j'th element of $\Psi_i c_i$ is equal to $\hat{X}_i(s_{ij})$ from (1). Hence, the sum of squared errors is

$$\mathrm{SSE}_i = \sum_{j=1}^{n_i} \Big(W_{ij} - \hat{X}_i(s_{ij})\Big)^2 = (W_i - \Psi_i c_i)^T (W_i - \Psi_i c_i),$$

and the least squares solution to the estimation problem is the well-known $\tilde{c}_i = \left(\Psi_i^T \Psi_i\right)^{-1} \Psi_i^T W_i$. If it is not reasonable to assume that the measurement errors ϵ_{ij} are independent, or that they have the same variance, then a weight matrix should be incorporated in the expression for SSE_i in the usual way.

The degree of smoothness in the preceding is implicitly controlled by the number of basis functions, K. If K is too small, then some of the features in $\hat{X}_i$ will be smoothed away, whereas if K is too large, then the least squares prediction will overfit the data, and find features that have only occurred by chance and are thus not reproducible. Thus the least squares prediction is quite sensitive to the choice of K as well as to the type of basis functions used.

In order to address these problems, a roughness penalty may be introduced. A typical choice for this regularizing term involves the curvature of the function. Let $\lambda > 0$, and consider the penalized sum of squared errors

$$\mathrm{SSE}_{i,\lambda} = \mathrm{SSE}_i + \lambda \int \left(\hat{X}_i''(s)\right)^2 ds = \mathrm{SSE}_i + \lambda c_i^T R_\psi c_i. \tag{3}$$

Here, R_ψ is the $K \times K$ matrix with entry (k, l) equal to $\int \psi_k''(s)\psi_l''(s)\, ds$. When estimation of some curve derivative is also of interest, it is common to consider roughness penalties

involving derivatives of order higher than two, and use a basis with higher regularity (for instance, polynomial spline bases with higher order); see Section 2.4. Discrete penalties applied directly to the basis coefficients have also been suggested [16].

For a given λ, the value of $\mathrm{SSE}_{i,\lambda}$ is minimized by

$$\hat{c}_i(\lambda) = \left(\Psi_i^T \Psi_i + \lambda R_\psi\right)^{-1} \Psi_i^T W_i. \tag{4}$$

In particular, notice that the solution is still explicit and depends on λ. For large values of λ wigglyness of $\hat{X}_i$ is heavily penalized so $\hat{X}_i$ is quite smooth; for small values of λ the estimation $\hat{X}_i$ is more similar to unpenalized least squares estimate and thus potentially wiggly if K is large. In other words, penalization imposes smoothness on $\hat{X}_i$ and therefore implicitly reduces the space of admissible functions. The effective number of estimated parameters is $\mathrm{df}(\lambda) = \mathrm{trace}\big(\Psi_i(\Psi_i^T\Psi_i + \lambda R_\psi)^{-1}\Psi_i^T\big)$ which is decreasing in λ and equal to K for $\lambda = 0$ (corresponding to least squares). With an appropriate value of λ, we can use a rich basis — with a number of basis function as large as n_i — without affecting the prediction $\hat{X}_i$ much [17]. Moreover, as long as the bases are chosen rich enough, predictions computed with penalization are quite robust against the type of basis functions (e.g. Fourier and splines), the exact placement of knots in a spline basis, etc. Altogether penalization reduces the impact of the choice of basis.

It remains to select the smoothing parameter λ. Low values of λ result in model overfit while high values result in oversmoothing; our goal is to balance these competing effects. It is most common to rely on goodness-of-fit criteria such as the generalized cross-validation criterion (GCV) or Akaike's criterion (AIC). GCV measures goodness-of-fit by the ability to predict left-out observations, see [18] and [1], whereas AIC measures goodness-of-fit by the log-likelihood (assuming iid Gaussian error terms) penalized by the effective number of estimated parameters, $\mathrm{df}(\lambda)$. Another option is to treat the coefficients c_{ik} as random variables with λ as a variance parameter, and use maximum likelihood (ML) or restricted maximum likelihood (REML) for estimation of λ; see [19] and [20]. Simulation studies in [20] suggest that the ML/REML approach is less prone to overfit the data and has better numerical properties compared to the AIC/GCV approach, but the predicted functions corresponding to the optimal λ from the different criteria are most often comparable. Prediction of $\hat{X}_i$ is rarely the aim of the study, and our experience is that the final results are not very sensitive towards the selection approach.

2.2 Selection of basis functions and estimation approach

So far we have considered known basis functions, selected before the analysis, in combination with penalization. As an alternative, it is possible to select the basis adaptively to the data. Within the spline field, this can for instance be attained by free-knot regression splines; see, e.g., [21] and references therein. An example in this direction is given in Section 2.4. Other basis systems that are naturally locally-adaptive are provided by wavelets, which are particularly well suited for data with spikes or other strongly localized features; see [22] for an overview. Finally, one may use an empirical basis computed from the eigenfunctions of the empirical covariance operator, i.e., the principal components; see Section 4 below and [23]. In this latter case, smoothing becomes in fact a more integral part of the analysis.

While many smoothing methods are available, features of a particular dataset or analysis will influence which approach is most appropriate. For instance, spline-based approaches (including low-dimensional, penalized, and free-knot methods) are very appealing when derivatives are of interest, because derivatives of spline basis functions are often directly available. The flexibility of penalized splines may be detrimental if functional data are sparsely observed, as in traditional longitudinal studies. Instead, low-dimensional splines or parametric forms may be better suited: although they are more restrictive, given only a few data points for each curve these methods can be more stable than a penalized approach. Alternatively,

functional principal components methods (discussed in Section 4) provide a data-driven collection of basis functions. FPCA can be used for dense or sparse data and are well-suited for dimension reduction, but lack easily computed smooth derivatives.

2.3 Illustrating penalized spline smoothing with the DTI data

The DTI data illustrated in Figure 1 exhibit a roughness that is unlikely to represent true underlying anatomical features: fractional anisotropy in the right corticospinal tract is smooth along the axis of the tract. Noise in the tract profiles has several potential sources, ranging from measurement error in image acquisition to registration errors in defining anatomical structures to generate the profiles. Regardless of the source, it is necessary to smooth observed curves prior to further analysis; here we discuss a penalized spline approach for smoothing functional data.

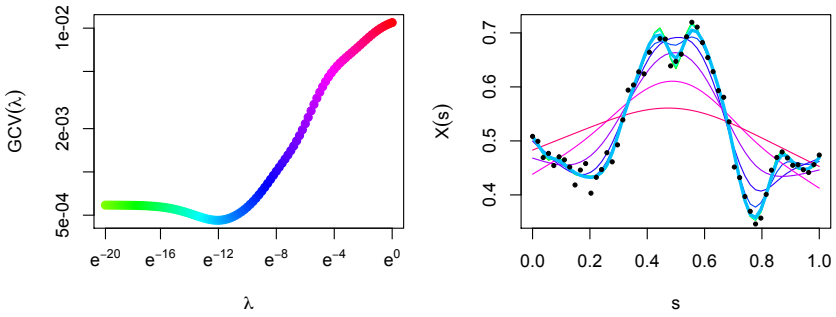

Figure 4: DTI data. Left: The GCV curve over a fine grid of tuning parameters λ. Right: Observed data for a single subject, as well as fitted curves for several tuning parameters. The colors of the fitted curves in the right panel correspond to the tuning parameter values in the GCV curve at left.

Because the placement and number of knots does not greatly affect overall fit when explicit penalization is used, we choose a rich cubic B-spline basis consisting of 35 functions. For a given value of the tuning parameter λ we use equation (4) to explicitly estimate the basis coefficients and, therefore, the expansion of each curve in our dataset. In this example we choose λ using GCV; similar results are obtained using REML. The left panel of Figure 4 evaluates GCV(λ) for a dense grid of potential λ values, and from this we choose the minimum $\lambda \approx e^{-12}$. Estimates for a single tract profile using several values of λ, including that chosen by GCV, are shown in the right panel with colors corresponding to the value of λ in the left panel. The chosen fit preserves the major structures and many details of the tract profile while removing minor deviations that do not represent anatomical features.

2.4 Illustrating curve derivatives estimation with the AneuRisk65 data

We here use the AneuRisk65 data to illustrate curve fitting when estimation of the curve derivatives is also of interest.

As mentioned in the introduction, the analysis of the AneuRisk project focused on the ICA, the artery with aneurysms possibly developing along or downstream of it. The vessel is clearly identifiable for each patient and is reconstructed from the angiographic image using the reconstruction algorithm coded in the "Vascular Modeling ToolKit" VMTK [24, 25]. In particular, for every subject i, VMTK reconstruction includes the three spatial coordinates of the vessel centerline (X_{ij}, Y_{ij}, Z_{ij}) for each point s_{ij} on a fine grid along a curvilinear abscissa that goes from the terminal bifurcation of the ICA towards the heart (measured in

mm). The left panel of Figure 2 shows the reconstruction of the ICA of the first subject in the dataset, with the reconstructed vessel centerline, and the top panels of Figure 5 display the three spatial coordinates of the centerline versus the approximate curvilinear abscissa. The reconstruction also provides, for each grid point s_{ij}, the radius R_{ij} of the vessel lumen section; the reconstructed radius profile for the first patient was shown in Figure 2 (top right panel).

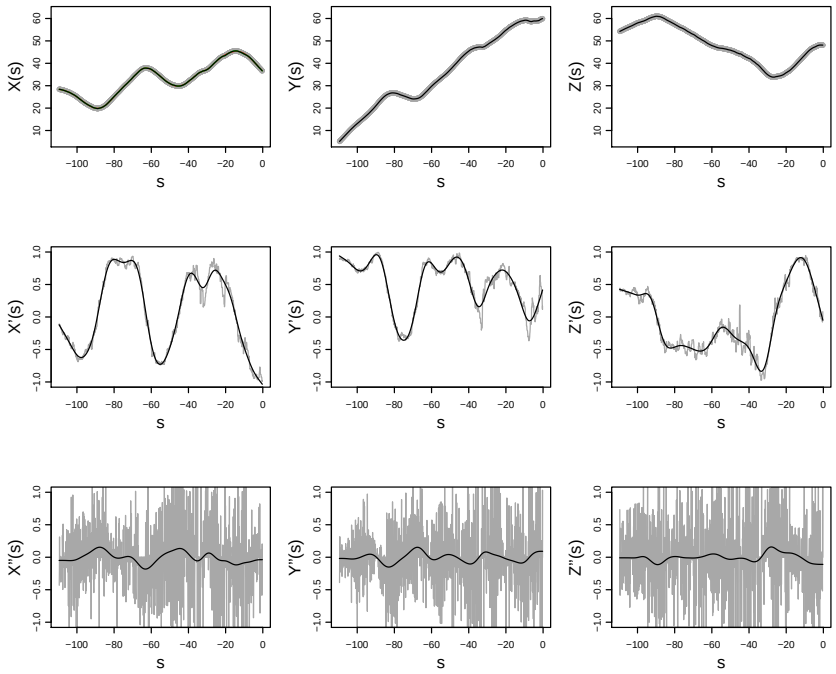

Figure 5: AneuRisk65 data. Top: Penalized spline estimates of ICA centerline for the first patient, i.e., $\hat{X}_1(s), \hat{Y}_1(s), \hat{Z}_1(s)$, superimposed to raw data $(s_{1j}, X_{1j}), (s_{1j}, Y_{1j}), (s_{1j}, X_{1j})$ in grey. Center: Corresponding first derivatives $\hat{X}_1'(s)$, $\hat{Y}_1'(s)$ and $\hat{Z}_1'(s)$ superimposed to first central differences in grey. Bottom: Corresponding second derivatives $\hat{X}_1''(s)$, $\hat{Y}_1''(s)$ and $\hat{Z}_1''(s)$ superimposed to second central differences in grey.

Besides the radius, another geometrical feature that has been studied within the project is the vessel curvature, identified by the curvature of the vessel centerline. Reconstructed ICA centerlines are of course affected by measurement and reconstruction errors. To compute their curvature it is hence necessary to obtain accurate estimates of the centerlines themselves as well as of their first two derivatives. Since we need smooth estimates of second derivatives, we here use splines with an higher order than those used for DTI data; we can for instance employ order five polynomial splines and penalize the third derivative, minimizing an SSE_λ defined analogously to (3). Also in this case we use a rich basis. The estimates of each space coordinate, $\hat{X}_i(s), \hat{Y}_i(s)$ and $\hat{Z}_i(s)$, are shown as solid black lines in the top panels of Figure 5, superimposed to raw data, displayed in grey (almost completely hidden by estimates). The effective number of estimated parameters is 20.

Differentiating these estimates as described in (2), we obtain estimates of first and second derivatives; $\hat{X}_i'(s), \hat{Y}_i'(s), \hat{Z}_i'(s)$ and $\hat{X}_i''(s), \hat{Y}_i''(s), \hat{Z}_i''(s)$, respectively; these estimates are shown as solid black lines in the central and bottom panels of the same figure. The estimates are superimposed to central differences (displayed in grey) that are rough estimates of first and second derivatives computed at each grid point as normalized differences of data values at nearby grid points. The higher order of the polynomial spline basis (order five) and the

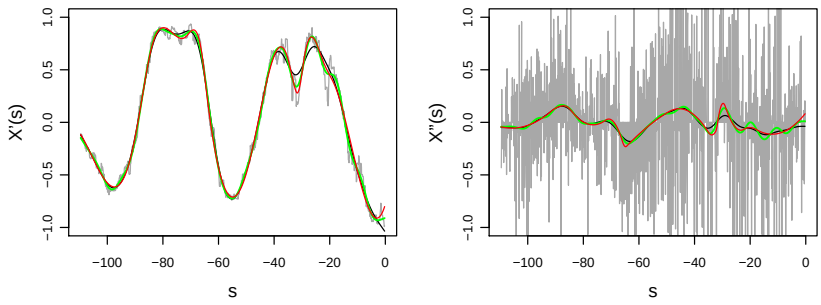

Figure 6: AneuRisk65 data. First and second derivatives (only first space coordinate) of estimated ICA centerline for the first patient, i.e., $\hat{X}_1'$ and $\hat{X}_1''$. The figure display the derivatives of the estimates of the ICA centerline obtained by penalized splines with effective number of estimated parameters equal to 20 (black) and with effective number of estimated parameters equal to 30 (green), and by free-knot splines with 20 basis functions (red). The estimates are superimposed to central differences (grey).

higher order of the derivative in the roughness penalty (derivative of order three), used here to fit the reconstructed centerline, ensure smooth estimates not only of the curve itself but also of its first and second derivatives, that in this example are respectively splines of order four (cubic splines), and splines of order three (quadratic splines). The estimated derivatives can hence be used to compute an estimate of the curvature of the vessel centerline, that is of interest in the study.

Figure 6 compares estimates of the first and second derivative of the first space coordinate, X_i, obtained from different smoothing procedures. First, consider penalized spline estimates with effective number of estimated parameters equal to 20 (shown in black, and identical to the center left and bottom left panels of Figure 5) and with effective number of estimated parameters equal to 30 (shown in green). By using a larger effective number of estimated parameters, penalized spline estimates can better capture the local features of the curve, such as the peaks and troughs in the first and thus subsequent derivatives, but at the price of being more wiggly over the whole range, as is apparent in the second derivative. This must be avoided to obtain sensible estimates of curvature profiles. Figure 6 also displays estimates obtained via the multi-dimensional free-knot spline technique detailed in [15]. Instead of using a rich spline basis system and penalizing some curve derivative, free-knot splines select and use only a few basis functions, chosen adaptively to the data. In particular, the selection of the basis from a rich basis system is driven by the minimization of a penalized sum of squared errors criterion, where the penalty term is now proportional to the number K_i of chosen basis, $\text{SSE}_i + \lambda K_i$; see [21] for details. In [15] this technique is extended to multidimensional curves, taking into account simultaneously all the space coordinates of the multidimensional curve in the basis selection. Thanks to their adaptivity to data, free-knot splines estimates are able to better detect salient localized features. Figure 6 shows the estimated first and second derivatives of X_i obtained from the free-knot spline estimate of order five with 20 basis functions (red). Comparing the different estimates in Figure 6, it is clear that free-knot splines estimates exhibit more clear-cut peaks and troughs; yet smoothness is maintained. The profile of the centerline curvature shown in the bottom right panel in Figure 2 is computed from these estimates.

3 Phase variation and alignment

With replications of functional data there are two important types of variation between curves: amplitude variation and phase variation. Phase variation — or curve misalignment — refers to the phenomenon that different curves exhibit more or less the same features, but that these features occur at different times or space locations for different subjects. Misalignment arises naturally for many reasons, including changes in data observation and recording, differences in the timing of disease onset and progression, and natural variation in the size and shape of anatomical structures. An illustration of misaligned data using the AneuRisk65 dataset is provided in Figure 7. In particular, the left panels of this figure display the first derivatives of the estimated ICA centerlines for the 65 subjects, in the X, Y and Z direction, respectively. It is apparent that the three-dimensional centerlines display a considerable misalignment. This misalignment is the expression of a strong phase variability present among the subjects, largely due to the different dimensions of the ICA across subjects. If not taken properly into account, the misalignment acts as a confounding factor and may blur subsequent analyses. For example, the mean curve will not exhibit features as clearly as the original curves when they occur at different time/location points. This is clearly illustrated in the left panels of Figure 7, where the mean curves are indicated as black solid lines.

It is thus often essential to include alignment (also called registration or warping) of the curves as part of the analysis.

3.1 Landmark alignment and continuous alignment

The idea is to find transformations h_i such that the transformed curves

$$\tilde{X}_i(s) = X_i \circ h_i(s) = X_i(h_i(s))$$

are "as similar as possible". The functions h_i are called warping functions and should be increasing. These warping functions capture the phase variability; amplitude variability, on the other hand, is the remaining variability in vertical direction among the aligned curves. In some cases time or location is merely shifted from curve to curve, for example because the measurements are started at random time points. For these situations it is natural to use $h_i(s) = s + \delta_i$. In other situations phase variation is a matter of dilation, in which case $h_i(s) = a_i s$ is a natural choice of warping function. In yet other situations the time or space deformation is more complex. In any case, the warping functions must be estimated from the data. Alignment, or registration, is a very active field of research and the subject of a large literature; see, e.g., the literature review in [26].

There are two main strategies for aligning observed functions. If the curves exhibit well-defined features or landmarks, such as peaks or valleys, one option is to warp the curves such that the features occur at the same time or location for all curves; this process is referred to as landmark registration; see e.g., [1, 27]. This approach is used in the DTI dataset, in which white matter tracts exhibit unique lengths, shapes, and trajectories across subjects. A trained neuroradiologist identified seven anatomical landmarks using the full-brain DT images; these landmarks were then used to register curves. Thus the tract profiles shown in Figure 1 are observed on the same domain across subjects and major features are aligned. While landmark-based registration can be accurate, it may also require significant user input and can be sensitive to the accuracy of the landmark identification. In some applications it is not possible to identify well-defined features that can be taken as landmarks, and landmark registration is of course not suited to these situations.

An alternative strategy consists in the so-called continuous registration; see, e.g., [1, 28, 29, 30, 31], with examples of registration of children growth curves. Various approaches have been proposed in this context. A common one is based on the definition of a suitable distance (or closeness) measure between curves, that measures dissimilarity (or similarity) between curves.

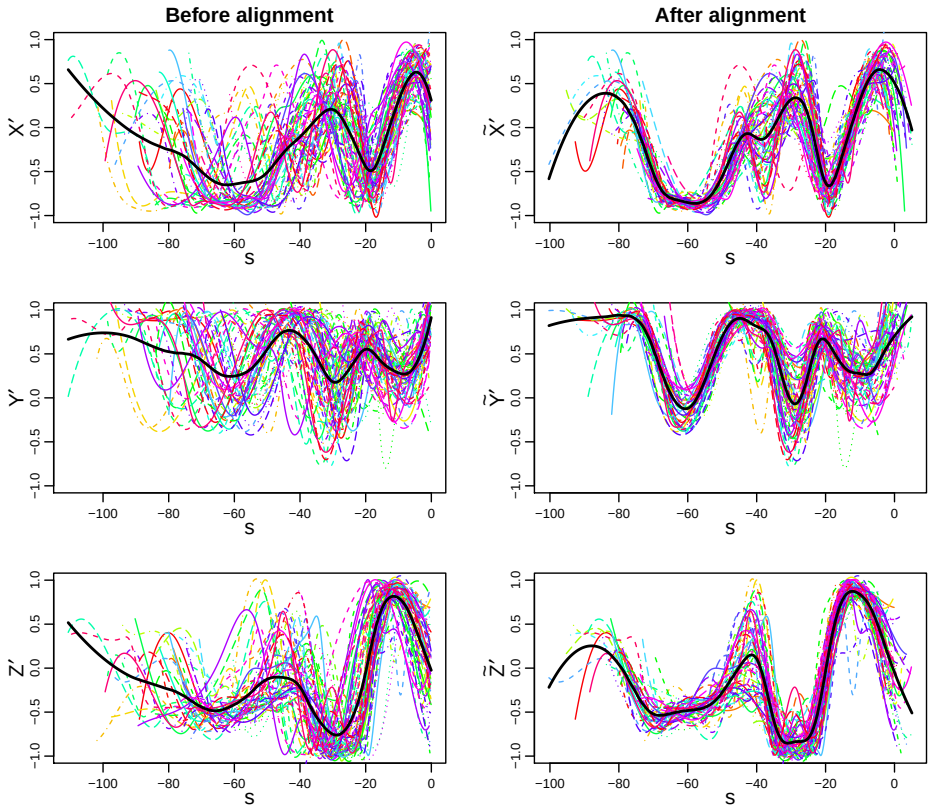

Figure 7: AneuRisk65 data. Left: First derivatives $X_i'(s)$, $Y_i'(s)$ and $Z_i'(s)$ of estimated ICA centerlines, before alignment. The superimposed solid black lines are average curves (estimated as local means). Right: First derivatives $\tilde{X}_i'(s)$, $\tilde{Y}_i'(s)$ and $\tilde{Z}_i'(s)$ of estimated ICA centerlines, after alignment. The superimposed solid black lines are the first derivatives of the reference centerline (estimated as local means).

The curves are thus aligned by warping their time or space abscissa parameters choosing the optimal warping function in some class of admissible warping functions in order to minimize the final distance among the curves or, equivalently, maximize their final similarity. Note that this problem is not univocally defined as different measures of distance or similarity between curves can be considered (see, e.g., [3] for examples of various distance measures), as well as different classes of admissible warping functions (e.g., simple translations or dilations, increasing linear transformations or more complex increasing transformations), leading to different registration results. In fact, the choice of the couple formed by distance measure and admissible warping functions defines the distinction between phase variability and amplitude variability in the specific problem under analysis. This choice must thus be problem specific.

How is alignment performed in practice? If there is a reference curve, X_0, then each of the observed curves $X_1, \ldots, X_n$ could be registered to X_0 by landmark or continuous registration, as described above. Most often, however, there is no such reference curve. Then iterative procrustes procedures may be used, see, e.g., [1]. Such procedures alternate between "reference estimation steps", where a reference curve is estimated using all the curves obtained at the previous iteration, and "alignment steps", where each curve is aligned to the reference curve estimated at the previous step.

It should be noticed that both the phase variation and/or the amplitude variation may be

associated to the phenomenon (for instance, the pathology) under study. It is thus necessary to study both types of variations to see how they relate to the problem being investigated.

3.2 Illustrating alignment with the AneuRisk65 data

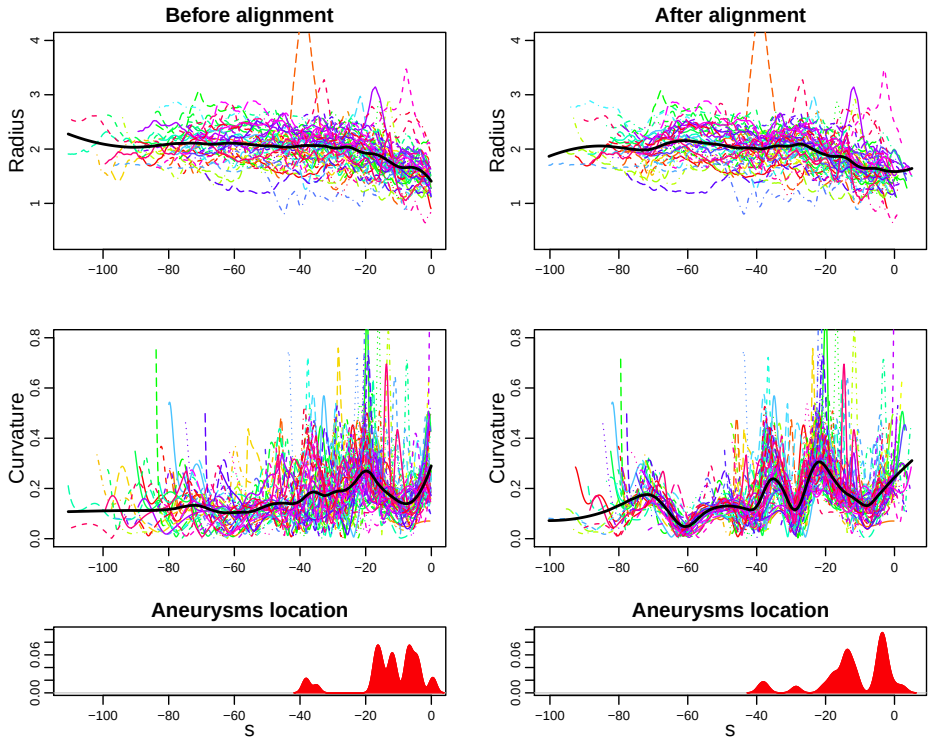

Figure 8: AneuRisk65 data. Top: ICAs radius profiles for all subjects before alignment (left) and after alignment (right); the superimposed solid black line is the average radius profile estimated by local means. Center: ICAs curvature profiles for all subjects before alignment (left) and after alignment (right); the superimposed solid black line is average the curvature profile estimated by local means. Bottom: Estimates of the probability density function of the location of aneurysms along the ICA or at its terminal bifurcation, before alignment (left) and after alignment (right).

To enable meaningful comparisons across subjects in the AneuRisk study, it is necessary to first efficiently decouple the phase and the amplitude variability. In this application phase variation is mainly due to differences in the dimension of carotids among subjects, whereas the amplitude variation is mainly due to differences in the vessel morphological shapes. The right panels of Figure 7 show the aligned first derivatives of the vessel centerlines, obtained by the continuous alignment technique detailed in [14] and [31]. The procedure has removed most of the phase variation, making it easier to compare curves from different subjects. Notice also that the mean curves of aligned data, drawn as black solid lines (right panels), represent the features of the individual curves far better than the mean curves of data that are not aligned (left panels). The variability captured by the optimal warping functions found during this alignment process (not shown here) is analysed in [14] and is found not to be associated to the aneurysmal pathology. Subsequent analysis may hence focus on the aligned data.

The optimal warping functions can be used to correspondingly align the radius and curvature profiles; see Figure 8. After alignment it is possible to start appreciating a common pattern for the curvature profiles (center right panel) that was not visible before alignment (center left panel). The effect of alignment on the radius curves is less prominent; this is because the radius profile displays a smaller variation along the ICA with respect to the curvature profile. The registered radius and curvature profiles highlight many interesting aspects. Figure 8 shows that the vessel gets progressively narrower toward the terminal bifurcation of the ICA; this is the so-called tapering effect. Tapering concerns all arteries, but it is particularly apparent close to the terminal bifurcation of the ICA, where the artery has to enter inside the skull. The figure also shows that most ICAs display two peaks of curvature in the terminal part of the vessel; these peaks of curvature are in correspondence with the carotid syphon. The bottom panels of the same figure display Gaussian kernel density estimates of the aneurysm location along the ICA, before and after alignment (left and right, respectively), based on data from patients having an aneurysm along the ICA or at its terminal bifurcation. The majority of ICA aneurysms are located in the terminal part of the vessel, where tapering is stronger, and after the main curvature peak. These results support the conjecture concerning the influence of the vessel morphology and the aneurysm onset, via the hemodynamics. In fact, the tapering of the vessel and the peak in its curvature determine hemodynamic regimes that may facilitate aneurysm formation and development. The relationship between morphological and hemodynamical features, and their impact on aneurysm pathology, is further explored in [13]. The bottom right panel of Figure 8 shows that, after alignment, the locations of ICA aneurysms cluster in two neatly separated groups, before and after –13 mm from the vessel terminal bifurcation. This fact suggest that this is the average position of the dural ring, i.e., the hole in the skull bone the ICA goes through to enter inside the skull. Notice that this ring cannot be detected directly through angiographies, but indications of the location of aneurysms relative to the dural ring may be of great importance, since aneurysms within the skull are more dangerous, as explained in Section 1.2.

4 Principal component analysis

The aim of functional principal component analysis (FPCA) is to find a few functions that capture most of the amplitude variation between replications of functions. More specifically, we are looking for functions $\xi_1, \ldots, \xi_K$ such that $X_i(s)$ is well-approximated by a linear combination $\sum_{k=1}^{K} c_{ik}\xi_k(s)$, where ideally K is small. This is similar to the smoothing problem considered in Section 2, but for FPCA the basis functions are estimated from the data (rather than being pre-specified), and the goal is dimension reduction by using a few efficient basis functions.

4.1 Eigen functions and pricipal component scores

The starting point is the covariance operator Σ defined as

$$\Sigma(s,t) = \mathrm{Cov}[X_i(s), X_i(t)] = \mathrm{E}\left[(X_i(s) - \mu(s))\,(X_i(t) - \mu(t))\right],$$

where $\mu(s) = \mathrm{E}[X_i(s)]$ is the expected value at time s. Notice that the curves are considered as replications such that μ and Σ does not depend on i. In practice, Σ should be estimated from the data; see [32] and [33]. Commonly the empirical covariance of the observed data is smoothed using a bivariate smoother after the main diagonal has been removed; this simultaneously borrows information across neighboring locations in the covariance operator and removes the "nugget" effect of measurement error. The spectral decomposition of Σ produces eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots$ and eigenfunctions $\xi_1, \xi_2, \ldots$, such that $X_i(s) = \mu(s) +$

$\sum_{k=1}^{\infty} c_{ik}\xi_k(s)$. The coefficient $c_{ik} = \int (X_i(s) - \mu(s))\xi(s)\,ds$ is called the kth score for curve i.

Similarly to standard principal component analysis, in FPCA the first eigenfunction ξ_1 has the property that it maximizes the variance $\mathrm{Var} \int \xi(s)X_i(s)\,ds$ among all functions ξ with $\int \xi^2(s)\,ds = 1$ (normalization). For $k \geq 2$, the kth eigenfunction ξ_k has the property that is maximizes $\mathrm{Var} \int \xi_k(s)X_i(s)\,ds$ among all functions ξ that satisfy $\int \xi(s)^2\,ds = 1$ as well as $\int \xi_j(s)\xi_k(s)\,ds = 0$ for $j = 1, \ldots, k-1$ (orthogonal to all previous eigenfunctions). In this sense ξ_1 captures the most variation, ξ_2 the second-most variation, uncorrelated to the previous, and so forth. In practice the first K principal components are used to express the observed curves so that $X_i(s) \approx \mu(s) + \sum_{k=1}^{K} c_{ik}\xi_k(s)$, and K is chosen so that a high proportion of total variance is explained. The kth eigenvalue λ_k can be interpreted as the contribution of term k to the total variance. The eigenvalues often decrease very quickly: in many applications $K < 10$ suffices to explain a large proportion of overall variability. Principal component scores c_{ik} can in practice be estimated either using a Riemann sum approximation to the integral definition, or as random effects in a mixed model framework. The second approach is common when curves are sparsely observed or subject to measurement error, but in many cases the two approaches are approximately equal [32].

After FPCA the effective dimension of each curve has been reduced to a vector of length K, and the scores $c_{i1}, \ldots, c_{iK}$ can be used as covariates in subsequent analyses regarding regression classification, regression, prediction, etc.; see the next paragraph and Section 5 for examples.

Notice that FPCA is often applied to curves with different distributions (for example curves from different treatment groups), even though the construction implicitly assumes that the curves have same mean function and covariance operator. The idea is that principal component scores can be used to distinguish different types of curves. For instance [14] shows an FPCA of aligned radius and curvature profiles of the ICA of all patients in the AneuRisk65 dataset. A discriminant analysis of the scores of the first two principal components of aligned radius and curvature is used to discriminate two groups of patients: a first group of patients having the aneurysm at or after the terminal bifurcation of the ICA, and a second group of patients having the aneurysms along the ICA, before its terminal bifurcation, or having no apparent aneurysm. In particular, the discriminant analysis is able to correctly identify patients in the first group, that, as mentioned in Section 1.2, are those having the most dangerous aneurysms. This gives further strong statistical evidence in favor of the conjecture explored within the AneuRisk project.

4.2 Illustrating FPCA with the DTI tract profile data

We illustrate FPCA using the DTI data set. From all available data for the right corticospinal tract, we estimate the mean function $\mu(s)$ and the covariance Σ; after smoothing, we obtain eigenfunctions $\xi_k(s)$ and eigenscores λ_k. Figure 9 shows in the left panel the mean function $\mu(s)$ as well as $\mu(s) \pm \sqrt{\lambda_1}\xi_1(s)$ for $k = 1, 2$. The mean function is highest on the interval from 0.3 to 0.7, roughly corresponding to the midbrain and internal capsule; fractional anisotropy is highest in this region and is lower in the medulla and pons (0.0 to 0.3), and in the corona radiata and the subcortical white matter (0.7 to 1.0). The first principal component is roughly a mean shift indicating that the overall level of fractional anisotropy varies across subjects over the full domain of the tract profile; on the other hand the second principal component affects fractional anisotropy only in a limited range. Figure 9 also illustrates the distribution of loadings for the first two principal components, separately for cases and healthy controls. For the first PC, the groups are indistinguishable, meaning that the variability in the overall level of fractional anisotropy is not apparently related to disease status. On the other hand there appears to be a group effect for the second PC loading, which indicates that differences between groups may be localized to particular tract regions.

As expected, increasing K improves the quality of the approximation $X_i(s) \approx \mu(s) +$

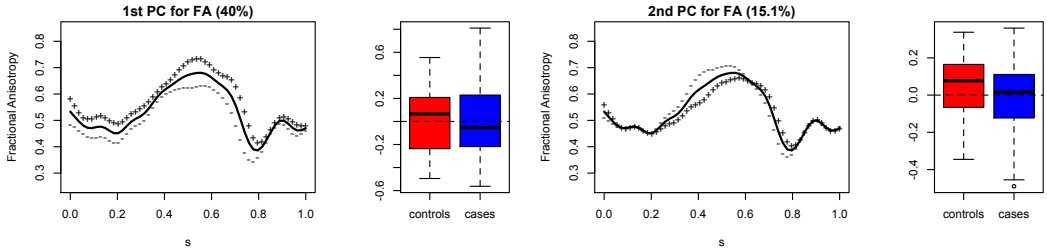

Figure 9: DTI data. The first and third panels (from the left) show $\mu(s)$ and $\mu(s) \pm \sqrt{\lambda_k}\xi_k(s)$ for $k = 1, 2$ respectively, with all quantities estimated using the DTI dataset. Boxplots of the corresponding scores for MS patients and healthy controls are shown in second and fourth panel.

$\sum_{k=1}^{K} c_{ik}\xi_k(s)$. In Figure 10, we show the approximation for two curves when $K = 1, 3, 10$. Later principal components contribute less to the approximation than earlier components, but it is important to choose K large enough to obtain reasonable expansions.

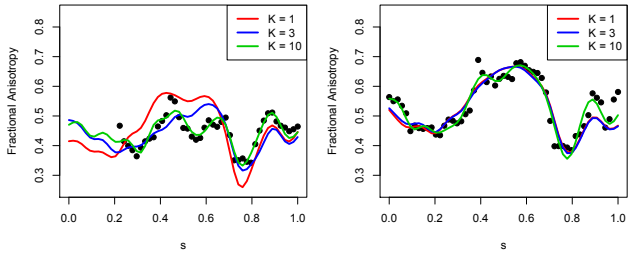

Figure 10: DTI data. The expansions $\mu(s) + \sum_{k=1}^{K} c_{ik}\xi_k(s)$ for two different curves (left and right) and different values of K.

De-noising and interpolation for the DTI data is incorporated in the FPC analysis, rather than being performed as a separate pre-processing step. Because the covariance operator is smoothed, the estimated principal components derived from the covariance are also smooth. Additionally, notice that the left curve in Figure 10 is not observed near the medulla. Such missing data is difficult to address using curve-specific smoothing, but can be handled in an FPCA framework. The mean and principal component basis functions (common to all tract profiles) are estimated using all available data across subjects. Curve-specific loadings are estimated using the data available for that curve, and are combined with mean and basis functions to estimate the curve over the full domain. Thus FPCA addresses both measurement error and missing data in the DTI data context.

5 Functional regression

The term “functional regression" is used for the study of associations between variables, when one or more of them are functional; see [1] and [6] for overviews. In the following, emphasis will be on the situation with a scalar response and functional covariates; these situations often naturally arise in real data analysis, as in the DTI example where the association between cognitive function and anatomy as measured by tract profiles is of interest. Although our focus is on scalar-on-function regression, we briefly mention function-on-scalar and function-on-function regression towards the end of the section. In the following, we assume that the functional observations have been sufficiently pre-processed using techniques described in the

previous sections to allow their use as covariates in a regression model. For ease of notation we additionally assume that the mean curve has been subtracted from each X_i; this changes the population intercept but not other aspects of the model.

5.1 Functional linear model

The simplest model for a scalar response is the following: for subject i a one-dimensional continuous response Y_i is observed, as well as a function $X_i(s)$, and we are interested in the conditional distribution of Y_i given $X_i(s)$. Assume that, conditionally on $X_1, \ldots, X_n$, the response variables $Y_1, \ldots, Y_n$ are independent and Gaussian with mean

$$\mathrm{E}[Y_i] = \alpha + \int X_i(s)\beta(s)\, ds \tag{5}$$

and variance σ^2; this is referred to as the functional linear model. To build intuition for the interpretation of this model, assume for the moment that curves are observed on $[0, T]$. Consider a partitioning of this interval into N subintervals of length T/N, and let $s_j = jT/N$. Then $\mathrm{E}[Y_i] \approx \alpha + \sum_{j=1}^{N} \frac{T}{N}\beta(s_j)X_i(s_j)$. Thus, the functional model given by (5) is a natural extension of the ordinary multiple regression model with covariates $\tilde{X}_{ij} = \frac{T}{N}X_i(s_j)$. In contrast to ordinary multiple regression where the regression coefficients are allowed to vary freely, in the functional context it is natural to require some amount of smoothness in the coefficient function $\beta(s)$.

Our main interest is in inference about the coefficient function, an infinite-dimensional parameter. Two broadly defined approaches have been pursued: 1) projection of $\beta(s)$ onto a low-dimensional basis and using standard regression models for inference; 2) use of a rich, flexible basis for $\beta(s)$ and imposing smoothness constraints and other explicit penalties. We note that this dichotomy is not perfect, but argue that it is useful for discussing general approaches.

In the low-dimensional approach, it is common to project both the estimates $X_i(s)$ and the coefficient function $\beta(s)$ using the functional principal components basis derived from the X_is (see Section 4), so that $\beta(s) = \sum_{k=1}^{K} \xi_k(s)\beta_k$. Called functional principal components regression (FPCR), this approach has been widely studied and is a common starting point for functional regression analysis [34, 1]. The coefficient function is thus restricted to the space spanned by the first K PCs; the truncation lag K implicitly controls the smoothness of the coefficient function and can be considered a tuning parameter. This approach builds on the intuition that the first K components explain a high proportion of the predictor variability, and may therefore be useful for modeling associations with the outcome. After expanding in terms of the orthonormal PC basis, we have that $\int X_i(s)\beta(s)\, ds = \sum_{k=1}^{K} c_{ik}\beta_k$, so the functional model can be expressed as a multiple linear regression with subject-specific PC loadings as scalar covariates and regression parameters $\beta_1, \ldots, \beta_K$, and inference proceeds as in standard multiple linear regression.

For the penalized approach, we express $\beta(s)$ in terms of the spline basis $\psi_1, \ldots, \psi_M$ so that

$$\beta(s) = \sum_{j=1}^{M} b_j\psi_j(s) = \psi(s)b$$

where $\psi(s) = (\psi_1(s), \ldots, \psi_M(s))$ is a row and $b = (b_1, \ldots, b_M)^T$ is a column. Rather than estimating α and b via maximum likelihood, we impose smoothness via a penalty term added

to the log-likelihood [35, 36, 33]. For fixed smoothing parameter λ the expression

$$f_\lambda(\alpha, b) = \sum_{i=1}^{n} \left(Y_i - \alpha - \int \beta(s) X_i(s)\, ds \right)^2 + \lambda \int \left(\beta''(s)\right)^2 ds$$
$$= \sum_{i=1}^{n} \left(Y_i - \alpha - \int \psi(s) b X_i(s)\, ds \right)^2 + \lambda b^T R_\psi b \tag{6}$$

should be minimized with respect to $\alpha \in \mathbb{R}$ and $b \in \mathbb{R}^M$. Here R_ψ is the $M \times M$ penalty matrix associated with the basis ψ, typically with entry (k, l) equal to $\int \psi_k''(s)\psi_l''(s)\, ds$. Our use of the notation $\psi(s)$ for the spline basis expansion of $\beta(s)$ recalls the smoothing problem in Section 2, although one is not constrained to use the same basis for both problems.

Many conceptual aspects of penalized functional regression are similar to those for smoothing. The tuning parameter, λ, reflects the trade-off between model fit and smoothness — too large values impose a large degree of smoothness at the cost of model fit, whereas too small values result in model overfit — and the parameter is selected along the same principles that were used for smoothing in Section 2. For example, generalized cross-validation (GCV) aims at minimizing prediction errors of left-out observations, whereas the REML approach treats the coefficients as random effects ([20, 33]).

Recent work has sought to extend these approaches in several directions. A criticism of the FPCR approach is the implicit assumption that first few major directions of variability in the predictors are related to the outcome, and no other directions are. To address this, a combination of FPCR and penalized regression is considered in [37] to allow larger numbers of FPC basis functions to be included while maintaining overall smoothness in $\beta(s)$. In [38] the authors propose a shrinkage-inducing lasso penalty to select only useful FPC basis functions. In the penalized spline context, [39] introduced penalties geared toward "interpretability": coefficient functions are induced to be linear, constant, or zero.

The methods developed or mentioned above are useful approaches for the functional linear model, but a complete literature review of all methods is outside the scope of this introductory paper. Although each method assumes the same outcome structure in equation (5), the modeling approaches can result in very different coefficient estimates and fitted values due to differing assumptions about the structure of $\beta(s)$, basis functions, and penalization. In practice, we highly recommend that multiple methods are applied to any specific dataset to explore the effect of these assumptions on model fit.

5.2 Generalized functional linear models and functional outcomes

The basic model above given by (5) can be straightforwardly generalized to allow non-Gaussian outcomes. For example, if the response is binary, then a logistic regression version of the model assumes that, conditionally on $X_1, \ldots, X_n$, the binary responses $Y_1, \ldots, Y_n$ are independent and that

$$\log \frac{P(Y_i = 1)}{P(Y_i = 0)} = \alpha + \int \beta(s) X_i(s)\, ds.$$

This model (put in a more general exponential family framework) was described in detail in [40, 41, 23, 33], and allows for multiple functional covariates as well as ordinary covariates. The ideas for estimation and inference carry over from the Gaussian case: coefficient functions can be expressed using low-dimensional basis expansions or penalized splines. Different estimation methods for the same basic model may give different results.

Finally, some comments on models for functional responses [42, 43]. Let $Y_i(s)$ be a functional response and Z_i an $1 \times p + 1$ vector of ordinary covariates for subject i (including an intercept term), and consider a function-on-scalar linear model:

$$\mathrm{E}\left[Y_i(s)\right] = Z_i \boldsymbol{\beta}(s)$$

where $\boldsymbol{\beta}(s)$ row-stacks functional coefficients $\beta_0(s), \beta_1(s), \ldots, \beta_p(s)$. Coefficients $\beta_j(s)$ can be expanded using an M-dimensional spline basis as above so that $\beta_j(s) = b_j\psi(s)$ with $b_j \in \mathbb{R}^M$. Letting $\boldsymbol{b}$ represent the matrix of row-stacked vectors b_j, we seek to minimize

$$f(\boldsymbol{b}) = \sum_{i=1}^{n} \left[\int \left(Y_i(s) - Z_i \boldsymbol{b} \psi(s) \right)^2 \, ds \right]$$

over $\boldsymbol{b}$ in $\in \mathbb{R}^{p+1 \times M}$. As elsewhere, we can add penalization terms to the rows of $\boldsymbol{b}$ and thereby explicitly induce smoothness in estimated coefficient functions; see [1, Chapter 13] and [44] for details. If the response $Y_i(s)$ as well as the covariate $X_i(u)$ are functional, a commonly used function-on-function model is

$$\mathrm{E}\left[Y_i(s)\right] = \mu(s) + \int X_i(u)\beta(u,s)\, du.$$

Notice that $X_i(u)$ and $Y_i(s)$ can be observed over distinct domains. Once again, it is typical to expand coefficient functions using spline bases and estimate parameters based on minimizing a (penalized) sum of squares [1, Chapters 14, 16].

5.3 Illustrating scalar-on-function regression with the DTI tract profile data

For the DTI dataset we study the association between tract profiles and cognitive performance for the MS patients. Cognitive performance is measured using the paced auditory serial addition test, or PASAT score, and takes on values between 0 and 60 with 60 being a perfect score. This test measures a variety of cognitive processes, including auditory processing, short term memory, and arithmetic calculation [45]. In this subsection, we are interested in the association between performance on the PASAT and anatomical structure measured using the fractional anisotropy tract profile of the right corticospinal tract.

We consider three methods for estimating model (5) using the DTI data set: (1) FPCR with the number of basis functions chosen via cross-validation; (2) penalized spline regression with tuning parameter λ selected using restricted maximum likelihood; and (3) the "flrti" method for interpretable coefficient of [39], with zeroth and first derivative penalties, with all tuning parameters chosen via crosss-validation. Coefficient function estimates are shown in the left panel of Figure 11, and a scatterplot matrix of observed outcomes and fitted values from each method is given in the right panel.

For these data, the three estimation approaches result in similar inferences about the location and direction of association for sections of the tract profile that are related to the outcome, and yield similar fitted values. Points near 0.2 and near 0.8 (near the pons and the corona radiata, respectively) are most strongly related to the cognitive outcome. For the region near 0.2, the coefficient function shows that above-average fractional anisotropy is related to higher cognitive performance, while the reverse is true near the corona radiata. Despite the general agreement, there are important differences. The FPCR estimate is the least smooth and may be difficult to interpret; the penalized spline estimate is smooth across the full range of s, but is not zero anywhere and has strange behavior in the right tail; the "flrti" estimate contains a few regions with constant effects and is identically equal to zero elsewhere. The fitted values from each method are highly correlated (near .9 for all pairwise comparisons) but not identical. The regressions explain 10–14% of the total variation of the cognitive performance (highest for the penalized method).

Although these numbers are not large, the analysis indicates an interesting association between the corticospinal tract and cognitive function. Anatomically, the tract primarily transmits signals from the motor cortex and should not necessarily be associated with cognitive function, and it is likely that the corticospinal tract is useful as an indicator of overall

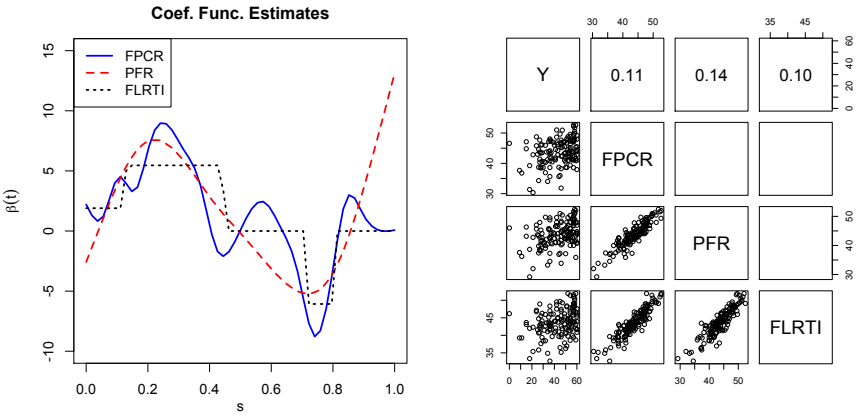

Figure 11: DTI data. Left: estimated coefficient functions for the functional linear model using FPCR, a penalized approach, and "flrti". Right: scatterplot matrix of observed outcomes and fitted values from each of the three methods; the numbers in the top row are the ratios of total variance of cognitive performance described by the regressions.

disease burden rather than as a direct effector of cognitive performance. Indeed, additional analyses have indicated that the relationship between the corticospinal tract and cognitive performance is diminished after adjusting for anatomical information more directly related to cognition in the corpus callosum [46].

6 Software

There are a few R packages directed towards functional data analysis. The `fda` package, [47], has functions for smoothing, alignment, principal component analysis, functional regression with continuous scalar response (but not with non-continuous response), and functional regression with functional response. All functions are also implemented in Matlab, and there is an accompanying book which covers both programs [48]. The R package `fda.usc`, [49], extends some of the facilities in `fda` and furthermore incorporates some of the non-parametric methods proposed in [3] as well as bootstrap techniques. It also provides an overview of existing R packages for functional data. The R package `refund`, see [50] and [33], is designed for functional regression with scalar response and covers the generalized linear model set-up from Section 5 with several functional covariates as well as ordinary covariates. It relies on functions from the `mgcv` package, in particular the function `gam`. The `mgcv` package, see [51], is designed for generalized additive models, i.e. models that include smooth terms, and is very useful for smoothing and regression in the functional data setting.

In Matlab, PACE is a comprehensive package for functional data analysis, which implements methods developed and discussed in [6, 32, 23, 43], among others. Based on functional principal component analysis (FPCA), it is particularly useful for a versatile collection of smoothing and regression problems; dense as well as sparse data, scalar as well as functional response, longitudinal data, bootstrap, etc.

7 Discussion and conclusion

In this manuscript we have attempted to provide an introduction to several basic techniques in functional data analysis, and to emphasize that functional data naturally arise in many biomedical applications. The functional perspective is useful in providing a framework to understand complex data observed within and across subjects. Tools for analyzing functional

data inherently respect the structure and order found in these data, and therefore can have intuitive interpretations that increase subject-area knowledge.

We have focused on major techniques used for functional data: smoothing, alignment, principal component analysis, and scalar-on-function regression. Each of these areas is the subject of ongoing research, and our exposition was intended as a partial overview rather than as a definitive description. Moreover, there are many important topics that we have chosen not to address, including ordering and robust statistics, data depth, multidimensional functional data (images, surfaces, spatial data), clustering, classification, functional outcome regression, hypothesis tests. Interested readers may find the monographs by Ramsay and Silverman [1], and by Ferraty and Vieu [3] to be useful starting points for further investigation.

One important message is that there are many exciting possibilities with functional data, and although we have given some guidance along the way, there are obviously many choices and decisions to be made during the analysis. We encourage scientists to play with different tools and thus explore the robustness of results against different sets of assumptions and computational approaches.

As described in the introduction, functional approaches are currently used in many areas, but functional data will be even more prominent as the conceptual framework becomes more well-known and as tools for analysis become easier to implement on new data. The impact of functional data will extend to other scientific domains through the dissemination of statistical methods and through increased collaboration between statisticians and scientists.

Acknowledgement

The AneuRisk65 dataset is a product the AneuRisk project; more information about the project can be found at the AneuRisk webpage http://mox.polimi.it/it/progetti/aneurisk/ where the data are also made available. An increasing data warehouse concerning aneurysm pathology can be accessed from the AneuRisk Web Repository http://ecm2.mathcs.emory.edu/aneurisk/ managed by Emory University and Orobix. L. Sangalli acknowledges funding by MIUR Ministero dell'Istruzione dell'Università e della Ricerca, *FIRB Futuro in Ricerca* research project "Advanced statistical and numerical methods for the analysis of high dimensional functional data in life sciences and engineering" (http://mox.polimi.it/users/sangalli/firbSNAPLE.html), and by the program Dote Ricercatore Politecnico di Milano - Regione Lombardia, research project "Functional data analysis for life sciences".

References

[1] Ramsay JO, Silverman BW. *Functional Data Analysis*. Second edn., Springer, 2005.

[2] Ramsay JO, Silverman BW. *Applied Functional Data Analysis*. Springer, 2002.

[3] Ferraty F, Vieu P. *Nonparametric Functional Data Analysis: Theory and Practice*. Springer, 2006.

[4] Ferraty F, Romain Y ((eds.)). *The Oxford Handbook of Functional Data Analysis*. Oxford University Press, 2011.

[5] Horvath L, Kokoszka P. *Inference for Functional Data with Applications*. Springer, 2012.

[6] Müller HG. Functional modelling and classification of longitudinal data. *Scandinavian Journal of Statistics* 2005; **32**:223–240.

[7] Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G. *Longitudinal Data Analysis*. Handbook of Modern Statistical Methods, Chapman & Hall/CRC, 2009.

[8] Reich DS, Smith SA, Zackowski KM, Gordon-Lipkin E, Jones CK, Farrell JAD, Mori S, van Zijl PCM, Calabresi PA. Multiparametric magnetic resonance imaging analysis of the corticospinal tract in multiple sclerosis. *NeuroImage* 2007; **38**:271–279.

[9] Reich DS, Ozturk A, Calabresi PA, Mori S. Automated vs. conventional tractography in multiple sclerosis: variability and correlation with disability. *NeuroImage* 2010; **49**:3047–3056.

[10] Basser P, Mattiello J, LeBihan D. MR diffusion tensor spectroscopy and imaging. *Biophysical Journal* 1994; **66**:259–267.

[11] LeBihan D, Mangin J, Poupon C, Clark C. Diffusion tensor imaging: Concepts and applications. *Journal of Magnetic Resonance Imaging* 2001; **13**:534–546.

[12] Mori S, Barker P. Diffusion magnetic resonance imaging: Its principle and applications. *The Anatomical Record* 1999; **257**:102–109.

[13] Passerini T, Sangalli LM, Vantini S, Piccinelli M, Bacigaluppi S, Antiga L, Boccardi E, Secchi P, Veneziani A. An integrated CFD-statistical investigation of parent vasculature of cerebral aneurysms. *Cardiovascular Engineering and Technology* 2012; **3**:26–40.

[14] Sangalli LM, Secchi P, Vantini S, Veneziani A. A case study in exploratory functional data analysis: Geometrical features of the internal carotid artery. *Journal of the American Statistical Association* 2009; **104**:37–48.

[15] Sangalli LM, Secchi P, Vantini S, Veneziani A. Efficient estimation of three-dimensional curves and their derivatives by free-knot regression splines, applied to the analysis of inner carotid artery centrelines. *Journal of the Royal Statistical Society, C* 2009; **58**:285–306.

[16] Eilers PHC, Marx BD. Flexible smoothing with b-splines and penalties. *Statistical Science* 1996; **11**:89–121.

[17] Ruppert D. Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics* 2002; **11**:735–757.

[18] Craven P, Wahba G. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross validation. *Numerische Mathematik* 1979; **31**:377–403.

[19] Ruppert D, Wand M, Carroll R. *Semiparametric Regression*. Cambridge University Press, 2003.

[20] Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society, B* 2011; **73**:3–36.

[21] Zhou S, Shen X. Spatially adaptive regression splines and accurate knot selection schemes. *Journal of the American Statistical Association* 2001; **96**:247–259.

[22] Ogden RT. *Essential Wavelets for Statistical Applications and Data Analysis*. Birkhauser (Boston), 1997.

[23] Müller HG, Stadtmüller U. Generalized functional linear models. *The Annals of Statistics* 2005; **33**:774–805.

[24] Piccinelli M, Bacigaluppi S, Boccardi E, Ene-Iordache B, Remuzzi A, Veneziani A, Antiga L. Geometry of the internal carotid artery and recurrent patterns in location, orientation and rupture status of lateral aneurysms: an image-based computational study. *Neurosurgery* 2011; **68**:1270–1285.

[25] Piccinelli M, Veneziani A, Steinman DA, Remuzzi A, Antiga L. A framework for geometric analysis of 852 vascular structures: applications to cerebral aneurysms. *IEEE Transactions on Medical Imaging* 2009; **28**:1141–1155.

[26] Vantini S. On the definition of phase and amplitude variability in functional data analysis. *TEST* 2012; **21**:676–696.

[27] Gasser T, Kneip A. Searching for structure in curve samples. *Journal of the American Statistical Association* 1995; **90**:1179–1188.

[28] Ramsay JO, Li X. Curve registration. *Journal of the Royal Statistical Society, Series B.* 1998; **60**:351–363.

[29] Wang K, Gasser T. Synchronizing sample curves nonparametrically. *The Annals of Statistics* 1999; **27**:439–460.

[30] Gervini D, Gasser T. Self-modelling warping functions. *Journal of the Royal Statistical Society, Series B* 2004; **66**:959–971.

[31] Sangalli LM, Secchi P, Vantini S, Vitellli V. k-mean alignment for curve clustering. *Computational Statistics and Data Analysis* 2010; **54**:1219–1233.

[32] Yao F, Müller HG, Wang J. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* 2005; **100**(470):577–590.

[33] Goldsmith J, Bobb J, Crainiceanu CM, Caffo B, Reich D. Penalized functional regression. *Journal of Computational and Graphical Statistics* 2011; **20**:830–851.

[34] Cardot H, Ferraty F, Sarda P. Functional linear model. *Statistics and Probability Letters* 1999; **45**:11–22.

[35] Marx BD, Eilers PHC. Generalized linear regression on sampled signals and curves: a P-spline approach. *Technometrics* 1999; **41**:1–13.

[36] Cardot H, Ferraty F, Sarda P. Spline estimators for the functional linear model. *Statistica Sinica* 2003; **13**:571–591.

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

[38] Lee ER, Park BU. Sparse estimation in functional linear regression. *Journal of Multivariate Analysis* 2011; **105**:1–17.

[39] James GM, Wang J, Zhu J. Functional linear regression that's interpretable. *Annals of Statistics* 2009; **37**:2083–2108.

[40] James GM. Generalized linear models with functional predictors. *Journal Of The Royal Statistical Society Series B* 2002; **64**:411–432.

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

[42] Ramsay J, Dalzell C. Some tools for functional data analysis. with discussion and a reply by the authors. *Journal of the Royal Statistical Society: Series B* 1991; **53**:539–572.

[43] Yao F, Müller HG, Wang JL. Functional linear regression analysis for longitudinal data. *Annals of Statistics* 2005; **33**:2873–2903.

[44] Reiss PT, Huang L. Fast function-on-scalar regression with penalized basis expansions. *International Journal of Biostatistics* 2010; **6**:Article 28.

[45] Gronwall DMA. Paced auditory serial-addition task: A measure of recovery from concussion. *Perceptual and Motor Skills* 1977; **44**:367–373.

[46] Swihart B, Goldsmith J, Crainiceanu CM. Testing for functional effects. *Under Review* 2012; .

[47] Ramsay JO, Wickham H, Graves S, Hooker G. *fda: Functional Data Analysis* 2012. URL http://CRAN.R-project.org/package=fda, R package version 2.2.8.

[48] Ramsay JO, Hooker G, Graves S. *Functional Data Analysis with R and MATLAB*. Springer, 2009.

[49] Febrero-Bande M, de la Fuente MO. Statistical computing in functional data analysis: The R package fda.usc. *Journal of Statistical Software* 2012; **51**(4):1–28. URL http://www.jstatsoft.org/v51/i04/.

[50] Crainiceanu C, Reiss P. *refund: Regression with Functional Data* 2011. URL http://CRAN.R-project.org/package=refund, R package version 0.1-5.

[51] Wood SN. *Generalized Additive Models: An Introduction with R*. Chapman & Hall, 2006.