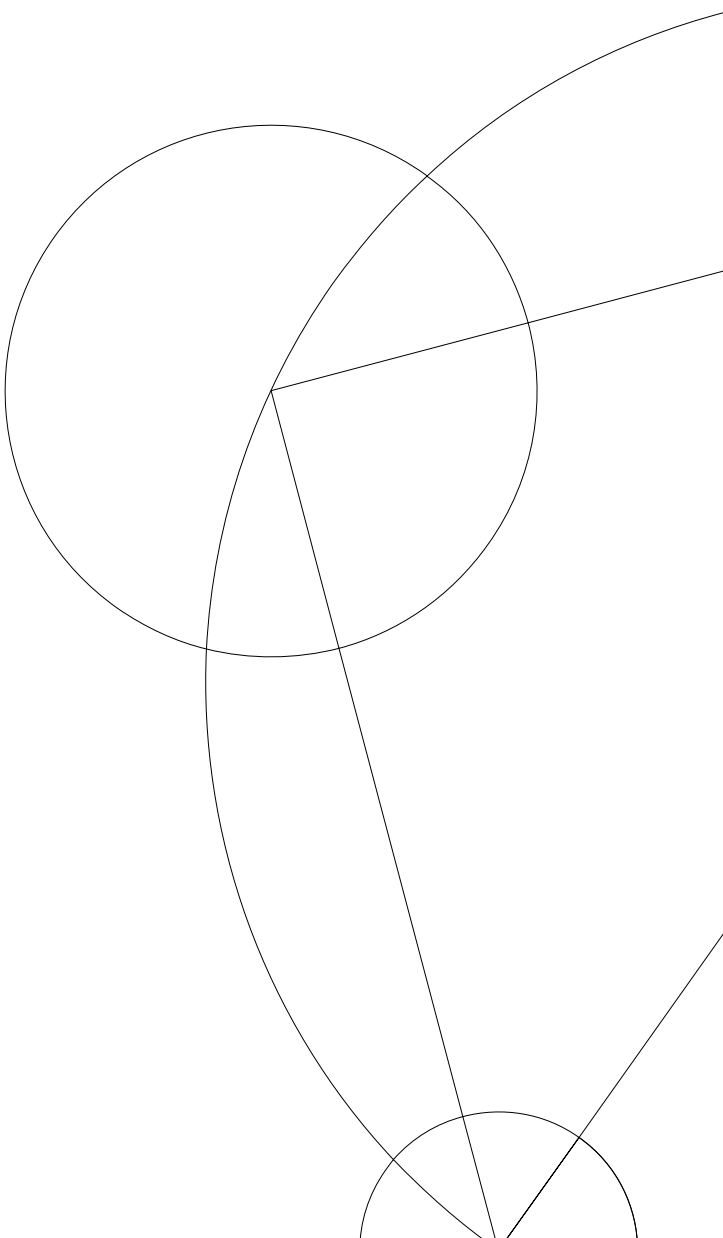
DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF COPENHAGEN

Functional Object Analysis

Toward Statistical Analysis of Functional Objects



PhD thesis

Lars Lau Raket

Thesis overview

This document constitutes my PhD thesis at Department of Computer Science, University of Copenhagen. The subject of the thesis is statistical analysis of functional data, and the main goal of the enclosed work is to go beyond the typical simple analyses of curve data, and open up the possibility of doing model-based statistical analysis of complex functional objects. The functional aspect of data naturally complicate statistical analysis; as opposed to conventional data analysis, geometric information has to be taken into account, and potentially enormous data sizes has to be handled. The contributions of this thesis are both theoretical, computational, and practical, and it offers solutions to some of the mentioned problems in various relevant cases.

During my PhD studies I have worked on a variety of topics. In particular, I have done much work on optical flow estimation and distributed video coding. Rather than writing an overview of my research output during my PhD studies, and the spurious relations between the different work, I have chosen to write a thesis on the area which I find most interesting, and on which I will continue my future research. Thus, optical flow estimation and video coding will not be considered in much detail in this thesis.

Lars Lau Raket
26 March 2014

Included papers

The main contribution of this thesis are the three papers

L.L. Rakêt and B. Markussen, "Approximate inference for spatial functional data on massively parallel processors," *Computational Statistics & Data Analysis*, vol. 72, pp. 227-240, 2014.

L.L. Rakêt, S. Sommer, and B. Markussen, "A nonlinear mixed-effects model for simultaneous smoothing and registration of functional data" *Pattern Recognition Letters*, vol. 38, pp. 1-7, 2014.

L.L. Rakêt, B. Grimme, B. Markussen, G. Schöner, and C. Igel, "Statistical analysis

of human arm movements using timing and motion separation," submitted to Neural Information Processing Systems (NIPS), 2014.

Finally, Appendix A includes a description of an image registration algorithm, which is based on results from the works

L.L. Rakêt, *Duality based optical flow algorithms with applications,* University of Copenhagen prize thesis in Computer Science, Copenhagen University Library, 2013.

L.L. Rakêt, L. Roholm, M. Nielsen, and F. Lauze, "TV-L^1 optical flow for vector valued images," in *Energy Minimazation Methods in Computer Vision and Pattern Recognition* (Y. Boykov, F. Kahl, V. Lempitsky, and F. Schmidt, eds.), vol. 6819 of *Lecture Notes in Computer Science*, pp. 329-341, Springer, 2011.

Excluded papers

While I have found that statistical analysis of functional data has been my most exciting contribution, I have spent the majority of my PhD working in other areas. I have made a number contributions to the fields of image and video processing, where I have mainly focused on variational formulations. Furthermore, I have done a considerable amount of work on integrating such variational methods in video codecs in the field of distributed video coding. These endeavors have been quite successful—for example the MORE codec (Luong et al. 2014), that I have recently co-developed, is the best performing single-view distributed video codec to date. The papers describing this work do unfortunately not fit into the topic of the present thesis, and have therefore been excluded.
The list of excluded publications are as follows:

M. Salmistraro, **L.L. Rakêt**, C. Brites, J. Ascenso, and S. Forchhammer, "Joint disparity and motion estimation using optical flow for multiview distributed video coding", *European Signal Processing Conference (EUSIPCO)* (**accepted**), 2014.

J. Petersen, M.M.W. Wille, **L.L. Rakêt**, A. Feragen, J.H. Pedersen, M. Nielsen, A. Dirksen, M. de Bruijne, "Effect of inspiration on airway dimensions measured in maximal inspiration CT images", *European Radiology* (**in press**), 2014.

H.V. Luong, **L.L. Rakêt**, and S. Forchhammer, "Re-estimation of motion and reconstruction for distributed video coding," *Image Processing, IEEE Transaction on*, vol. 23, pp. 2804-2819 2014.

M. Salmistraro, **L.L. Rakêt**, M. Zamarin, A. Ukhanova, and S. Forchhammer, "Texture side information generation for distributed coding of video-plus-depth," in *Image Pro-*

cessing (ICIP), 2013 20th IEEE International Conference on, pp. 1699-1703, 2013.

M. Salmistraro, M. Zamarin, **L.L. Rakêt**, and S. Forchhammer, "Distributed multi-hypothesis coding of depth maps using texture motion information and optical flow," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on,* pp. 1685-1689, 2013.

H.V. Luong, **L.L. Rakêt**, X. Huang, and S. Forchhammer, "Side information and noise learning for distributed video coding using optical flow and clustering," *Image Processing, IEEE Transaction on*, vol. 21, pp. 4782-4796, 2012.

L.L. Rakêt and M. Nielsen, "A splitting algorithm for directional regularization and sparsification," in *Pattern Recognition (ICPR), 2012 21st International Conference on,* pp. 3094-3098, 2012.

L.L. Rakêt, "Local smoothness for global optical flow," in *Image Processing (ICIP), 2012 19th IEEE International Conference on,* pp. 1-4, 2012.

L.L. Rakêt, J. Søgaard, M. Salmistraro, H.V. Luong, and S. Forchhammer, "Exploiting the error-correcting capabilities of low density parity check codes in distributed video coding using optical flow," in *Proceedings of SPIE, the International Society for Optical Engineering*, vol. 8499 SPIE – International Society for Optical Engineering, 2012.

L.L. Rakêt, L. Roholm, A. Bruhn, and J. Weickert, "Motion compensated frame interpolation with a symmetric optical flow constraint," in *Advances in Visual Computing* (G. Bebis et al., eds.), vol. 7431 of *Lecture Notes in Computer Science*, pp. 329-341, Springer, 2012.

X. Huang, **L.L. Rakêt**, H.V. Luong, M. Nielsen, F. Lauze, and S. Forchhammer, "Multi-hypothesis transform domain Wyner-Ziv video coding including optical flow," in *Multimedia Signal Processing (MMSP), 2011 IEEE 13th International Workshop on*, pp. 1-6, 2011.

Summary

I propose a direction it the field of statistics which I denote *functional object analysis*. This subfields considers the analysis of functional objects defined on continuous domains. In this setting I will focus on model-based statistics, with a particularly emphasis on mixed-effect formulations, where the observed functional signal is assumed to consist of both fixed and random functional effects. This thesis takes the initial steps toward the development of likelihood-based methodology for functional objects. I first consider analysis of functional data defined on high-dimensional Euclidean spaces under the effect of additive spatially correlated effects, and then move on to consider how to include data alignment in the statistical model as a nonlinear effect under additive correlated noise. In both cases, I will give directions on how to generalize the methodology to more complex data setups. Finally, I consider extensions and future directions.

Contributions

The main methodological contributions of this thesis are as follows:

- An operator approximation framework for doing inference in linear functional mixed-effects models (Rakêt & Markussen 2014).
- A description, and source code for doing efficient massively parallel inference in functional mixed-effects models (Rakêt & Markussen 2014).
- A model for doing likelihood inference for functional data under the effect of, possibly random, alignment variation (Rakêt, Sommer & Markussen 2014).

Contents

Chapter 1

Functional data typologies

1.1 Introduction

The objective of functional data analysis is to analyze data of a functional nature—data that are discrete observations of functional objects. Traditionally, the focus of functional data analysis has been on curve data (Ramsay & Silverman 2005), but the toolbox that has been developed for curve data has only had a limited success for more complex functional data. As a result of this focus on curves, some authors have come to regard functional data analysis the study of curve data, and nothing else. In this process of distantiation, analysis of more complex functional objects are often assigned to other subdisciplines with fundamentally different methodology.

In the spirit of unification, Wang & Marron (2007) propose the term object-oriented data analysis which is "the statistical analysis of populations of complex objects". This concept gives a valuable system for talking about analysis of complex objects and comparing methodology for different types of objects. Its generality, however, also means that the unifying goal is mostly achieved in terms of high-level issues, since the considered complex objects can be of very different nature. For example, both smooth curves and tree structures are considered as objects, even though they are fundamentally different. Taking, for example, a stochastic point of view, the random variability in such data types is probably not comparable. The statistical focus of object-oriented data analysis has primarily been on exploratory analysis, with a particular focus on alternatives to principal component analysis for complex data types (Marron & Alonso 2014). While principal component analysis is surely one of the most valuable tools in the data analysis toolbox, proper model-based statistical methodology for functional objects is needed for the field to reach a mature state.

In the following chapters we will propose a line of research along a subfield of object-oriented data analysis; a field which we will denote *functional object analysis*. This subfield considers the analysis of functional objects defined on continuous domains. In

this setting we will particularly focus on mixed-effect formulations, where the observed functional signal is assumed to consist of both fixed and random functional effects.

The move from curves to more complex functional objects represents a major increase in mathematical complexity. But not only that; with the increasing sophistication of measuring devices, data of functional objects today represents some some of the largest datasets available. Neuroimaging devices are capable of producing image volumes with hundreds of millions of voxels, and from microscopes to space telescopes, image resolutions in the gigapixel range are seeing the light of day. Thus, the question of efficient computation is of key importance in functional object analysis.

In the remainder of this chapter we will review different types of functional data and corresponding models.

1.2 Hierarchies of functional data

The broadly defined objective of functional data analysis—to analyze functional data– means that the field encompasses a large number of subfields. Our focus is on statistical methods, so we assume that data contains some degree of variation that, from a suitable viewpoint, can be considered stochastic. Furthermore, we will mainly consider data where a fixed effect of interest is available, and thus we will pay less attention to for example reconstruction of high-dimensional objects from low-dimensional projections. Finally, the main focus will be on functional data analyses where one of the goals is to infer a functional signal of interest.

In this section we will propose different hierarchies of functional data based on: the geometric and topological structure of the underlying functional object; the modeling of functional effects; and the models of uncertainty in the data.

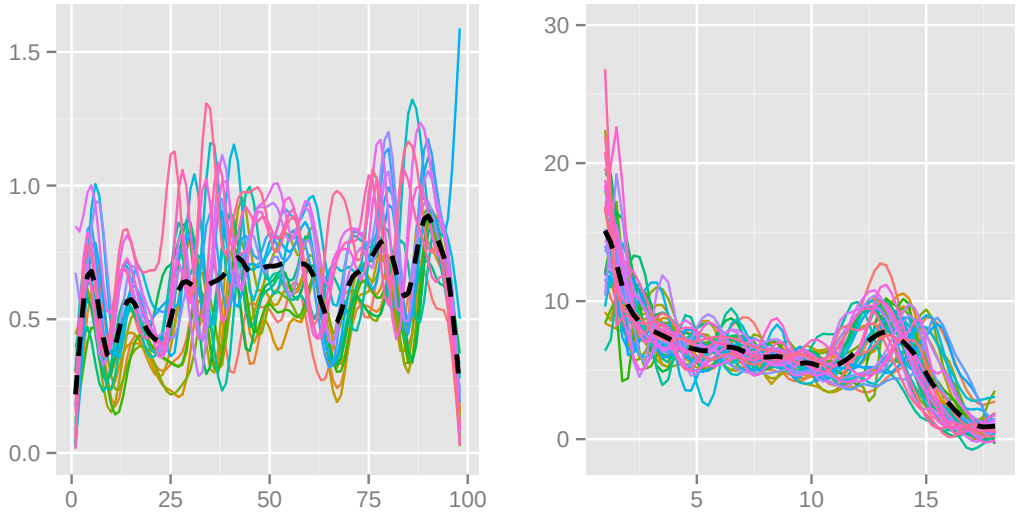
1.2.1 Geometric and topological data hierarchies

Let $\theta : \mathcal{D} \to \mathcal{V}$ be a functional object of interest. The geometric and topological properties of domain $\mathcal{D}$ and codomain $\mathcal{V}$ provide a natural hierarchy of functional data.

The main focus of functional data analysis has been on analyzing curve data, which also yields the simplest mathematical analysis. The typical examples of such data are temporal acceleration signals (Kneip & Ramsay 2008, Srivastava et al. 2011) such as the ones in Figure 1.1 (a), and (derivatives of) growth curves (Ramsay & Silverman 2005), an example of which is given in Figure 1.1 (b).

From the perspective of the functional effect of interest θ, the data complexity can be extended both in terms of the structure of the domain and the codomain of the underlying functional effect.

The differences in data complexity has resulted in a division of the types of functional objects that are analyzed in different branches of science. While an extensive literature on



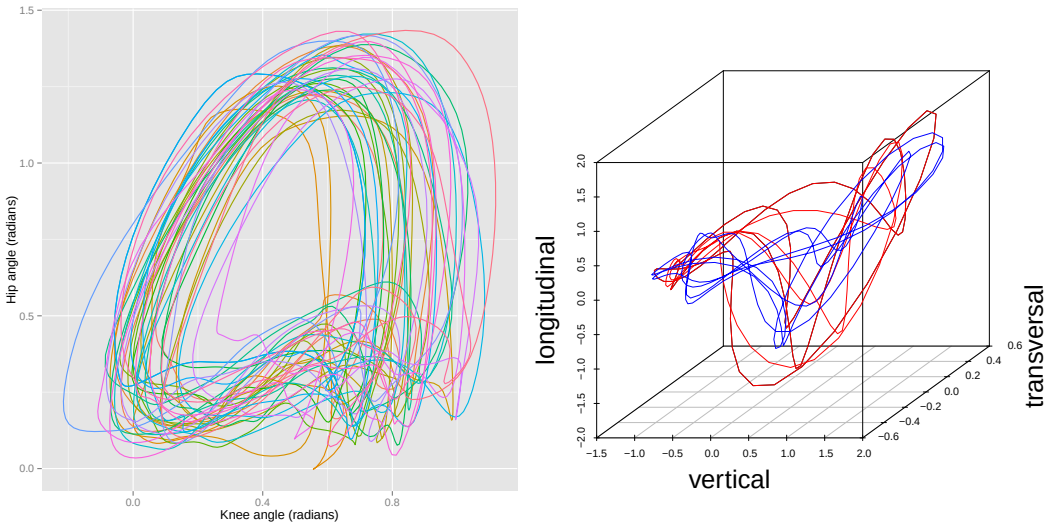
(a) Signature data (b) Growth velocity data

Figure 1.1: (a) 20 replications of the acceleration signal measured at the tip of a pen writing a signature, and (b) growth velocities for 39 male subjects in the Berkeley growth study as a function of age. The dashed black curve shows the pointwise average.

classical statistical analysis of curve data has been generated (Ramsay & Silverman 2005, Ferraty & Vieu 2006, Horváth & Kokoszka 2012), the literature on statistical models for spatial and volumetric data is much sparser, and an abundance of open problems for such data exists (Ferraty & Vieu 2006, Chapter 14). On the other hand, functional objects such as image volumes, and shapes have received considerable attention in computational anatomy and the mathematical branches of computer vision (Grenander & Miller 1998, Paragios et al. 2006). The methodology used in different fields are, however, often not directly compatible. Where statistical analysis is typically focused on the discrete set of observations, the approaches used to analyze complex functional objects often focus on the structure of the underlying function spaces. These differences, and the benefits of the different approaches will be discussed in Section 1.2.4.

In the following paragraphs we will review different types of functional objects.

Space curve data The perhaps simplest extension of curve data consists in extending the dimension of the value space, for example by considering effects $\theta : \mathbb{R} \to \mathbb{R}^q$. This type of data typically arises when one measures several quantities at the same time. A classical example is the San Diego Children's Hospital gait data, which consists of knee and hip angle measurements from 39 children during a gait cycle (Olshen et al. 1989). This data is shown in Figure 1.2 (a). Another typical example of this type of data is



(a) $\theta : \mathbb{R} \to [0, 2\pi]^2$ (b) $\theta : \mathbb{R} \to \mathbb{R}^3$

Figure 1.2: (a) Knee hip gait data, and (b) Repeated 3D acceleration signals from two different trotting horses.

spatial acceleration curves $\theta : \mathbb{R} \to \mathbb{R}^3$ measured using an accelerometer. Figure 1.2 (b) shows an example of such data obtained from two different horses, that has been used for classifying equine lameness (Sørensen et al. 2012).

The availability of space curve data from accelerometers and gyrometers have increased tremendously during recent years due to the presence of such devices in a wide variety of consumer electronics. As a result, this type of data is no longer restricted to expensive controlled experiments. An example of this is the action dataset generated by McCall et al. (2012) using smartphones attached to the belt of participants performing specific actions. This data is used by Tucker et al. (2013) who classify actions based on the observed functional signal, using Fisher-Rao distance between the samples.

A further complication of space curve data happens if the data naturally takes values in a smooth manifold $\mathcal{M}$, which is for example the case for data describing the position of a GPS tracker on the globe over time, in which case $\mathcal{M} = \mathbb{S}^2$. More elaborate manifolds may arise naturally when data is constrained in space. Consider for example human motion analysis using skeletal models (Figure 1.3). If data is given as joint-position in end-effector space (Hauberg et al. 2012), the constraint that bone lengths remain fixed over the timespan of the data acquisition, makes the space of possible values a smooth manifold.

Spatial data Extending the functional effect to a multidimensional domain presents a significant leap in complexity, analogously to the increased complexity encountered when

$\theta : \mathbb{R} \to \mathcal{M}$

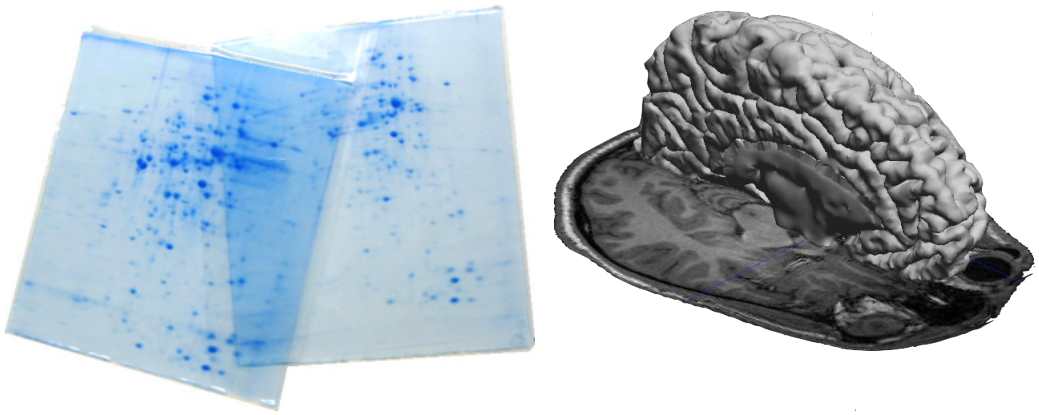
Figure 1.3: Skeletal model for human motion modeling in end-effector space (from Hauberg et al. 2012).

going from ordinary to partial differential equations. Common examples of such data are planar data $\theta : \mathbb{R}^2 \to \mathbb{R}^q$, for example geodata over small areas where the curvature of the globe can be ignored, electrophoresis images, and image volumes $\theta : \mathbb{R}^3 \to \mathbb{R}^q$. A pair of 2D electrophoresis gels and an MRI slice of a human brain along with its segmented reconstruction can be found in Figure 1.4.

Imaging technologies have been used in scientific fields for several decades, but the literature on modeling of spatial data in a stochastic setting is surprisingly underdeveloped. While classical statistical models are being used for analyzing e.g. neuroimage data (Worsley & Friston 1995), the vast data sizes encountered in this field often render direct analysis infeasible.

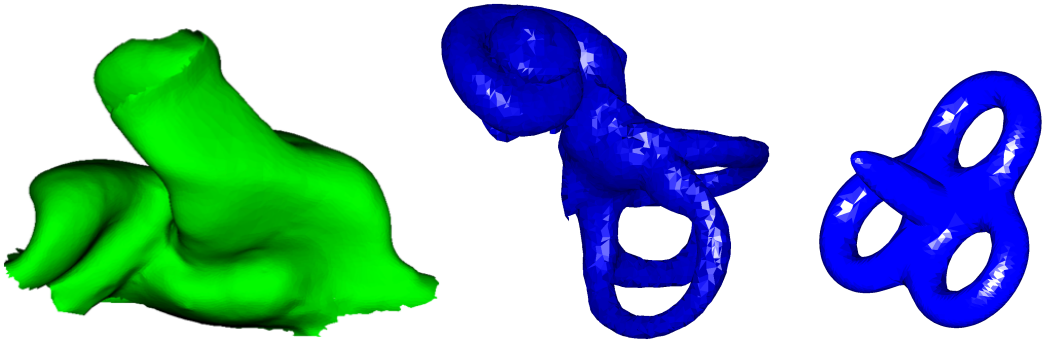
Data with topological structure As a result of the cyclic nature of the gait data in Figure 1.2, one would expect the underlying curve representing the gait cycle to be closed. Thus, it is natural to consider models where the fixed effect is defined on the circle, that is $\theta : \mathbb{S}^1 \to \mathbb{R}^q$. Other data types are naturally defined on domains with topological structure. For example, global measurements on the earth are naturally defined on a spherical domain. Such global measurements are typically of meteorological nature, for example temperature data (Lindgren et al. 2011) and ozone concentration measurements (Bolin & Lindgren 2011).

For planar shape data, such as outlines of cells or anatomical objects, the topological restriction that outlines are non-intersecting may be imposed. Furthermore, it is often natural to model shapes modulo translation, rotation, reflection, and possibly scaling (Dryden & Mardia 1998). These restrictions naturally complicate analysis of shape data, which is of course additionally hampered when the dimension of the domain is increased, as is the case when one consider shape analysis of the cortical surface extracted from MRI image data (Figure 1.4 (b)), where the surface can be considered a function



(a) $\theta : \mathbb{R}^2 \to \mathbb{R}$ (b) $\theta : \mathbb{R}^3 \to \mathbb{R}$

Figure 1.4: (a) Set of two 2D electrophoresis gels, and (b) slice of MRI volume of human brain with segmented reconstruction.



(a) (b) (c)

Figure 1.5: (a) shape of outer ear with cylindrical topology (from Darkner et al. 2007), (b) and (c) human cochlea along with its natural domain (from Charon 2013).[1]

defined on the two-sphere $\theta : \mathbb{S}^2 \to \mathbb{R}^3$. Shape data on nonspherical domains have not yet received much attention. Examples of such data are the shape of the outer ear in Figure 1.5 (a) which has open cylindrical topology, and the shape of the human cochlea in Figure 1.5 (b), which has the topology of a tri-torus.

[1]Data courtesy of Pr. Jose Braga from Université Paul Sabatier in Toulouse, Jean-Luc Kahn, curator at Institut d'anatomie normale et pathologique in Strasbourg, and Dr. Stanley Durrleman from Institut du cerveau et de la moëlle épinière in Paris

1.2.2 Modeling of functional effects

Functional effects are naturally infinite-dimensional while observation numbers are always finite. Due to this contrast between the acquired data and the object of interest, we have to put restrictions on functional effects in order estimate the full functional objects.

The simplest type of finite dimensional representation parametrizes functional effects in terms of all observation points, or in terms of a prespecified set of points, and interpolates intermediate points using for example linear interpolation. While such constructions are perhaps not the most aesthetic choice, they are often practical because of their simplicity, and generality. On the other hand, they can be problematic when data is not well-balanced, or when functional samples are not aligned—and well-aligned raw data is certainly a rarity.

For noisy curve data, the widely popular smoothing spline (Wahba 1975) is one of the earliest, and best studied models for functional effects. In generalized form, the smoothing spline is the functional effect θ in a suitable Sobolev or Beppo-Levi space that minimizes the penalized likelihood function

$$\ell(\theta) = \sum_{k=1}^{m} (y(t_k) - \theta(t_k))^2 + \lambda \int_{\mathcal{D}} \|\mathscr{K}\theta(t)\|^2 \, \mathrm{d}t, \tag{1.1}$$

for some differential operator $\mathscr{K}$. A rich theory has been developed around smoothing splines. In particular, the problem has been considered in a reproducing kernel Hilbert space setting, where the smoothing spline has been shown to have a sparse representation in terms of polynomials in the null-space of the smoothing operator and Green's functions for the smoothing operator $\mathscr{L} = \mathscr{K}^\dagger \mathscr{K}$ (Wahba 1990). Furthermore, generalizations to higher dimensional and spherical domains are readily available (Cox 1984, Wahba 1981). These ideas have also been generalized to curves on surfaces (Pottmann & Hofer 2005) and functions on topologically and geometrically complex surfaces (Duchamp & Stuetzle 2003).

Much work has used alternative finite-dimensional representations of functional effects in terms of basis functions (Ramsay & Silverman 2005). Popular choices of bases include the Fourier system, B-splines, and wavelets. On top of the regularization caused by the finite dimensional representation, one may penalize smoothness of the solution further (Ramsay & Silverman 2005, Hastie et al. 2009). One of the most popular basis function representations for curve data is the *penalized* spline, or P-spline, which is the result of finding θ as the combination of the B-splines in the given basis that minimizes the penalized likelihood function (1.1).

From the point of view of doing statistical analysis of more complex objects, such as three-dimensional shapes, discrete interpretable representation are very valuable since they may reduce both the mathematical complexity of the analysis and the computational burden. In this respect, deformable shape models based on medial skeletal representations have been a successful alternative to conventional descriptions (Pizer et al. 2005).

Recently, a new type of skeletal model was proposed by Pizer et al. (2013). These so-called s-reps are developed such that the representation is consistent across all shapes in the population at hand, which in turn makes the representation more suited for statistical analysis. This consistency also makes the representation a promising option for mixed-effect modeling of shape data. Modeling of complex functional objects is however still an underdeveloped area with many open problems.

1.2.3 Uncertainty structure

The canonical model of functional data analysis assumes that m noisy discrete observations $\boldsymbol{y}_i = y_i(t_k)_{1 \leq k \leq n_i}$ are made of a deterministic functional object θ, according to the statistical model

$$y_i(t) = \theta(t) + \varepsilon_i(t), \tag{1.2}$$

where ε_i is a zero-mean white noise process with possibly zero variance.

Let us first restrict ourselves to the case of curve data $\theta : \mathbb{R} \to \mathbb{R}$ in order to introduce the fundamental classes of uncertainty that we will consider. Curve data is of course both the simplest and most studied case of model (1.2). Classical examples such as the acceleration signals and growth velocity curves in Figure 1.1 have often been analyzed using model (1.2). A closer inspection of these two datasets, however, reveal two elements that are not accounted for by the model.

For the signature data, the deviations around the mean curve seem to be systematic for the different repetitions. A more proper statistical modeling would assume that sample i included a serially correlated effect x_i that models the systematic amplitude variation

$$y_i(t) = \theta(t) + x_i(t) + \varepsilon_i(t). \tag{1.3}$$

Model (1.3) is a simple example of a *functional mixed-effects* model. A wide variety of models in this class have been considered in the literature. Wang (1998) proposes a class of models where the fixed effect θ is modeled using a smoothing spline, and the random effects x_i are assumed to be realizations of zero-mean Gaussian process with pre-specified parametric covariance structure. Guo (2002) develops functional mixed-effects models where both fixed and random effects are modeled using smoothing splines, which, using the equivalence between smoothing spline models and Gaussian process models, can be considered a special case of the model of Wang (1998). In this class of models, Guo (2002) introduces some valuable computational tools. In a similar fashion, Chen & Wang (2011) consider modeling of fixed and random effects using P-splines. A recent development is the functional mixed-effect framework proposed by Markussen (2013). In this framework fixed effects are modeled using a pointwise representation which makes it possible to efficiently approximate likelihood calculations using operator calculus.

An alternative approach to functional mixed-effect models considers the problem in a nonparametric setting, where no distributional or parametric assumptions are made on the random effects. An example of this approach is given by Boularan et al. (1994), who consider modeling of growth curves, assuming only that population and individual effects are twice differentiable, and propose an estimation scheme based on kernel smoothing. For a recent review of linear functional mixed-effects models we refer to Liu & Guo (2012).

A closer inspection of the growth velocity data reveals another feature that is not modeled by the mixed-effect model (1.3): there seems to be *phase* variability—variability in the time domain. The overall shape of the growth velocity curves is consistent, but not aligned along the age-axis. Such alignment effects are present in almost all functional data, and can also be spotted in the signature data, although to a lesser extent.

As can be seen for the growth velocity data in Figure 1.1, ignoring phase variability will inevitably lead to overly smooth estimates of fixed effects (dashed curve), that lack the details of the individual samples. The typical solution to the problem consists in warping data as a preprocessing step, and then carry out analysis on the processed data. From a statistical point of view, this approach can however be problematic. Considering the growth data, the differences in growth velocity peaks across the different subjects can be ascribed to a large number of factors such as genetics, nutrition, and exposure to hormone-disrupting chemicals. Thus, it is natural to consider the alignment differences a random effect, fully comparable to the serially correlated effects of model (1.3). Pre-alignment of data will thus exclude this stochastic element from the analysis, which may bias the resulting estimates and conclusions. Thus, it is natural to consider the nonlinear functional mixed-effects models

$$y_i(t) = \theta(v_i(t)) + x_i(t) + \varepsilon_i(t) \tag{1.4}$$

and

$$y_i(t) = \theta(v_i(t)) + x_i(v_i(t)) + \varepsilon_i(t) \tag{1.5}$$

where v is a sample-specific (possibly stochastic) warping function.

In model (1.4) the warping function is assumed to only affect the fixed effect θ. This model is relevant when the amplitude variation x_i is assumed to follow the non-warped domain of the experiment, which generally happens when the amplitude variation can be ascribed to sources that are independent of the functional signal. The alternative model (1.5) should be used when the amplitude variation x_i is tied to the functional signal, for example when x_i models biological variation around θ.

Verzelen et al. (2012) propose an interesting alternative to the mixed-effects models considered here. They propose to model curve data by a nonlinear differential equation of the form

$$\theta_i'(t) = f(t, \theta_i(t)) + x(t)$$

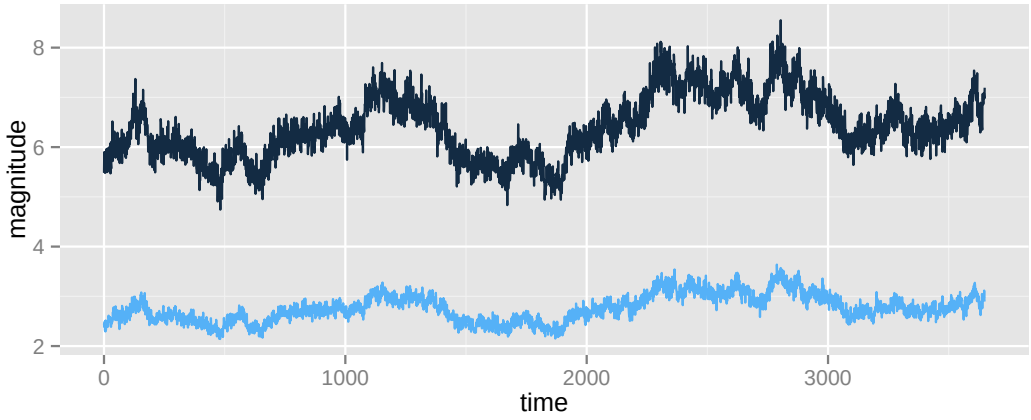


Figure 1.6: Simulated light curves from the Strong Lens Time Delay Challenge.

where x is a zero-mean stochastic process. In this model, in addition to the fixed effects θ_i describing the individual samples, one wants to estimate the nonlinear dynamics in terms of the fixed functional effect f. The generalization and adaption of this idea to the more complex types of data described in the previous sections could provide a valuable model for exploratory analysis. This type of approach does however require that a sufficiently high number of samples are available.

In the following paragraphs we will consider how the geometric and topological properties of the functional object θ plays a central role in the models one has to consider, and how the nonlinear functional mixed-effects models (1.4) and (1.5) can sometimes be related to these complex data situations.

Invariant data transformations Consider the two simulated light curves from the Strong Lens Time Delay Challenge (Dobler et al. 2013) in Figure 1.6. This type of data from gravitational lens systems can be used to infer important cosmological parameters, if one can identify the slight shifts that are present between the curves. Due to magnification, the observed curves have a difference in scale that cannot be ignored in the analysis.

The idea of analyzing data using methods build around invariance principles is getting increasingly popular. For the light curves in Figure 1.6, one could apply a scale-invariation transformation of data such as the census transform or the complete rank transform, and then use the transformed data to infer the shift (Demetz et al. 2013). Such tools are valuable for exploratory analysis because they are widely applicable, but one has to keep in mind that invariances come at a price in terms of discarded information. From the model-based point of view considered here, the proper model for the data should simply include a curve specific scaling parameter of the fixed and possibly random effects.

Many types of invariant data transformations naturally occur. Data transformations

are often encountered when one considers vector-valued functional data. Changing the dimension of the codomain is typically straight forward in terms of the statistical analysis—in models (1.4) and (1.5), one need to specify a correlation structure between the dimensions of the serially correlated effects $\boldsymbol{x}_i$, but since the curve is still indexed by a one dimensional variable, the task of registering the curves is essentially unchanged. A multi-dimensional codomains however opens up the possibility of new types of data effects. For example, the placement and orientation of an accelerometer may influence the collected data, and thus it will sometimes be natural to consider rotations of the fixed and serially correlated effects in value space to align samples.

The main idea in the formulation of the statistical models considered here, is that data transformation should be part of the model—we should include a rotation of the mean acceleration curve in value space toward the samples, rather than rotating the individual noisy samples toward the mean.

Restricted codomains Consider a model with a functional effect $\theta : \mathbb{R}^d \to \mathcal{M}$, where the codomain $\mathcal{M}$ is a smooth Riemannian submanifold of $\mathbb{R}^q$. Suppose we consider skeletal motion data in end-effector space (Figure 1.3), but instead of human motion, the observations are of a robot repeating the same motion a number of times, with independent identically distributed zero-mean observation noise. This data setup belongs in the canonical model (1.2), and the restriction of the codomain to the manifold $\mathcal{M}$ represents information that will surely improve the estimation of θ at a limited cost in complexity. On the other hand, when we are constructing models for data with serially correlated effects, which is the case in human motion data, one has to carefully consider the nature of the random effects. One could consider model (1.5), in which θ represents the mean motion in $\mathcal{M}$, x_i represents the serially correlated deviation from θ, v_i represents the difference in timing between the different repetitions, and ε_i is measurement noise. For such data, it is natural to assume that the systematic deviation x_i of the underlying motion θ takes place on the manifold $\mathcal{M}$. This means that both the underlying effect θ and the observed effect $\theta + x_i$ should be contained in $\mathcal{M}$. This intricate dependence of x_i on the fixed effect θ, makes it difficult to derive a reasonable stochastic model for such data. In this case, it is more natural to consider a completely nonlinear model of the form

$$y_i(t) = x_i^\theta(v_i(t)) + \varepsilon_i(t). \tag{1.6}$$

where x_i^θ is an $\mathcal{M}$-valued stochastic process with mean θ. Constructing general stochastic processes on manifolds, and doing inference in the resulting models can however be difficult.

Another example of restricted codomains comes from the warping functions in models (1.4) and (1.5), which are often assumed to be diffeomorphisms. This assumption is natural in many cases. Consider for example the signature data in Figure 1.1 (a).

Each acceleration peak corresponds to a specific feature of a letter, and these should come in the correct order to produce the signature. Data warping methods that ensure diffeomorphic warping functions have received considerable attention in recent years. Joshi & Miller (2000) consider landmark matching where diffeomorphisms are generated using a transport equation. The interesting idea behind this method is that not only the data domain is considered continuous, but also the path of the deformation. In the formulation of Joshi & Miller (2000), the data warping is considered a fixed functional parameter, however as noted by the authors, the spatial regularization at each time point along the path of the diffeomorphism gives the velocity fields the local structure of Gaussian random fields. If one considers formulations of random diffeomorphic warping functions, however, things get even more complex. For this reason, random diffeomorphisms have received a limited amount of attention from an applied point of view: Nielsen et al. (2008) consider so-called Brownian warps as a least-committed prior for Bayesian image registration. They derive the local distribution of the Jacobian of the Brownian warp in 2D, and describe how to use the method for aligning images. An alternative approach is given by Markussen (2007) who derives a stochastic transport equation for the same problem, and proves the equivalence of this method, and the methods of Joshi & Miller (2000) and Nielsen et al. (2008). A recent development is the second-order model of Vialard (2013) that generates random diffeomorphisms that are smooth along the path of evolution. While the mentioned works do take some steps toward defining proper statistical models for data requiring diffeomorphic warping, there is still a long way to practically applicable methods that can be used for parameter estimation and model validation.

Structured domains When building statistical models for functional data the biggest increase in mathematical complexity occurs when going from one-dimensional domains with trivial topology, to domains with higher dimension and nontrivial topology. When going from $\mathcal{D} = \mathbb{R}$ to $\mathcal{D} = \mathbb{R}^d$ $d \geq 2$, the number of well-established models for random effect x_i is significantly reduced, due to the limited practical knowledge about high-dimensional random fields compared to one-dimensional stochastic processes. The practical implication of this is that random effects are customarily assumed to be Gaussian with Matérn covariances. The complications are further increased when considering domains with topological structure. While random fields on spheres have received a fair amount of attention (see for example Fisher 1993), domains like the tri-torus used for the cochlea surfaces in Figure 1.5 represents a great challenge in terms of defining natural random fields. This complexity is even further increased if we want to model biological variation around a mean shape θ by means of a random field x_i, subject to the natural constraint that the observed shapes (including biological variation) do no self-intersect.

Likelihood-based solutions for data of this complexity does not seem likely in the near future, however, Charon (2013) analyze the cochlea data in a large deformations

by diffeomorphisms setting. Similarly to the previous described link betwen the works of Joshi & Miller (2000) and Markussen (2007), it will perhaps be possible to derive a relation to a statistical model based on stochastic differential equations for such data.

1.2.4 Practical and computational approaches to functional data analysis

The role of a statistical model and its ability to model a given phenomenon is often described by an overly negative George Box quote. When we develop statistical models it should be with the focus of making models that are good approximations of the observed phenomenon. But in this respect, we may add another set of objectives, namely simplicity and interpretability of the models. Often one can discard some of the complex data structure in the modeling, with no cost in terms of the approximative power of the model. For example, Hobolth & Jensen (2000) model cell shapes with a parametric fixed effect and a random effect modeled by a Gaussian process that changes the cell shape in the normal direction of the estimated fixed effect. This model seems to fit the considered data well, but if one were to insist that simulations from the model would never cause self-intersections of the shapes, the mathematical complexity would explode without any essential gain in accuracy.

An important aspect of statistical analysis is the ability to do significance testing. To derive reasonable testing procedures, we need a model that is a good description of the data, and a test statistic that can be used to evaluate a hypotheses. The descriptive quality of the model is important because the structure of uncertainty in the data is used to evaluate the probability of a given observation under the model.

Suppose we want to model planar shapes using a linear mixed-effects model similar to (1.3), and suppose that the natural sample variation x_i along the mean outline is modeled by a Gaussian process. If the estimated variance parameters of x_i corresponds to a process that produce self-intersecting shape samples with an overwhelming probability, the usefulness of the model is limited from a testing perspective, since the probabilities are clearly not assigned properly. On the other hand, when this is not the case, such a model may serve as a good approximation.

When it comes to test statistics, the standard choice is (restricted) likelihood-ratio tests. If one uses pointwise representations of fixed effects in the linear models described in the previous section, classical asymptotic results from multivariate analysis can be used to derive the approximate distribution of the likelihood-ratio test statistics. This approximation may however be poor because of the large number of parameters. Building on the same classical foundation, Cuevas et al. (2004) consider an anova test for functional data in the setting of model (1.2), and propose a numerical procedure to handle the asymptotic distribution. In the general linear mixed-effect setting, Antoniadis & Sapatinas (2007) develop a testing procedure that uses wavelet decomposition of both

fixed and random effects. There is an ongoing development of testing procedures in functional data analysis, but in applications, hypothesis tests based on functional models are rarely used. Furthermore, there is still a need for large scale simulation studies to assess the quality of current testing procedures.

Continuous or Discrete?

Functional data has a built in duality; data consist of discrete observations, but arise from an infinite-dimensional space. The statistical community has traditionally had a very strict focus on the discrete observations. This focus makes sense from a philosophical point of view: the observation noise of model (1.2) is generally tied to the act of doing measurements, and is thus discrete in nature. The negative log likelihood function of model (1.2) under the assumption of zero-mean Gaussian noise with variance σ^2 is

$$\ell(\sigma^2, \theta) = \frac{n}{2} \log \sigma^2 + \frac{1}{2\sigma^2} \sum_{k=1}^{n} (y(t_k) - \theta(t_k))^2, \qquad \theta \in \mathscr{H}. \tag{1.7}$$

In contrast to the discrete viewpoint of the observation model, analyses that are based on continuous formulations are often considered in other fields. The negative log likelihood function (1.7) cannot be directly used as it goes to infinity with n, and thus needs renormalization to be used in a continuous setting. An energy formulation for recovering the function θ, similar to (1.7) is of the form

$$E(\theta) = \int_{\mathcal{D}} (y(t) - \theta(t))^2 \, \mathrm{d}t, \tag{1.8}$$

which is then minimized using the discrete observations $y(t_1), \ldots, y(t_m)$, for example by interpolating unknown values. The minimizer θ of the negative log likelihood and the energy E clearly converge as the sampling rate goes to infinity, but the energy is not normalized by means of the variance, which in turn means that one cannot do direct parameter estimation from the energy. In this simple case, the solution is of course straightforward, but for more general models, linking likelihood functions to variational formulations can be difficult.

We have previously seen that model (1.2) is too simple to describe typical data effects, and both likelihood-based analyses and analyses based on continuous *energy formulations* includes one or more regularization terms of the form

$$\lambda \int_{\mathcal{D}} \|\mathscr{K} \theta(t)\|^2 \, \mathrm{d}t,$$

where $\mathscr{K}$ is a differential operator. Continuous energy formulations of this type has the advantage that one can write up the functional derivative using calculus of variations, and find the minimizer by solving the corresponding Euler-Lagrange equation.

The results of the maximum likelihood estimation and energy minimization may be quite different—in particular when data is irregularly sampled, in which case one has to pay special attention to using a renormalization schemes that properly approximate the variance normalization terms (Markussen 2013).

The likelihood method has the advantage that model parameters can be estimated directly from the likelihood function, and that hypothesis tests are possible. On the other hand, one can generally derive much faster minimization algorithms for energy formulations, and in the case of complex data, it may be significantly easier to formulate models in terms of data and regularization terms, rather than random effects.

Variational methods for analyzing functional data have been extensively used in different scientific communities, and have reached a quite mature state (Scherzer et al. 2008). The links between model-based statistical analysis of functional data and variational methods have historically only received very limited attention from the statistical community. In recent years, however, it seems that variational methods are catching on in statistics. Lindgren et al. (2011) use a finite element method to solve a partial stochastic differential equation, and these ideas have been further explored by Simpson et al. (2012*a*,*b*). On a similar node, Sangalli et al. (2013) use finite elements to efficiently estimate a spatial effect in a spatial regression model. In Chapter 2 we take these methods one step further. In addition to estimating fixed effects and predicting spatially correlated effects, we derive approximations for all the terms in the likelihood function, which enables maximum likelihood estimation of parameters for very large data sizes. This approach gives the best of both worlds: likelihood estimation of parameters combined with fast computations, and natural representation of effects.

Chapter 2

An operator-based approach to functional mixed-effect models

2.1 Introduction

In this chapter we present the first paper of the thesis (Rakêt & Markussen 2014). The paper considers spatial functional data with fixed effects $\theta : \mathbb{R}^d \to \mathbb{R}$ corrupted by spatially correlated noise and measurement noise. We show how the likelihood function can be approximated using variational formulations, and describe how these approximations can be used for doing efficient computations. We review related models, and consider the relations to the proposed method. Finally, we use the methodology to analyze a dataset of 28 pre-aligned 2D chromatograms, each of which has more than 5 million observation points. A detailed description of the chromatogram warping procedure, which is based on previous work on optical flow (Rakêt et al. 2011, Rakêt 2013), can be found in Appendix A.

Approximate inference for spatial functional data on massively parallel processors

Lars Lau Rakêt[†,*], Bo Markussen[‡]

[†]*Department of Computer Science,*
[‡]*Department of Mathematical Sciences*
University of Copenhagen, Denmark

Abstract

With continually increasing data sizes, the relevance of the big n problem of classical likelihood approaches is greater than ever. The functional mixed-effects model is a well established class of models for analyzing functional data. Spatial functional data in a mixed-effects setting is considered, and so-called operator approximations for doing inference in the resulting models are presented. These approximations embed observations in function space, transferring likelihood calculations to the functional domain. The resulting approximated problems are naturally parallel and can be solved in linear time. An extremely efficient GPU implementation is presented, and the proposed methods are illustrated by conducting a classical statistical analysis of 2D chromatography data consisting of more than 140 million spatially correlated observation points.[1]

Keywords: Functional data analysis, functional mixed-effects model, Gaussian processes, GPU, likelihood analysis, operator approximations

[1]Code for analyzing spatial functional data on graphics processing units is available as supplementary material.

[*]Corresponding author

1. Introduction

During the last half century, functional data analysis has become a well-established statistical discipline (Ramsay and Silverman, 2005; Ferraty and Vieu, 2006; Horváth and Kokoszka, 2012). The continuous sophistication of instruments gives rise to an increasing number of problems where functional aspects have to be taken in to account. Statistical analysis of functional data generally involves the ill-posed problem of inferring an infinite-dimensional function from discrete data points. This requires some sort of regularization, and the type of regularization is often chosen in terms of roughness penalties that lead to sparse representations of the inferred function in terms of simple basis functions (Wahba, 1990), thus reducing the computational complexity. The most typical specification, however, considers the inverse regularization process where a sparse basis is chosen explicitly for the given problem, which may then be further regularized through a roughness penalty (Ramsay and Silverman, 2005).

This paper takes a different path for model specification; we consider functional mixed-effects models with random effects generated by Gaussian processes, and present a framework that moves the calculations needed in such analyses from the discrete domain induced by the observations to the underlying functional domain. As a consequence it is possible to efficiently compute the functions in question, even if the regularization does not lead to sparse representations. The methods are based on the one-dimensional operator approximations of Markussen (2013), and here new results and resolution strategies are presented for high-dimensional domains.

The functional viewpoint sheds new light on some of the current challenges in statistics (Jordan, 2011), by both reducing the computational complexity of a large class of statistical problems dramatically, and at the same time revealing a natural link between partial differential equations and a large number of statistical models, including functional mixed-effects models, penalized likelihood, and Bayesian models.

In addition to reducing the computational complexity, the proposed resolution strategies are highly parallel, and naturally suited for implementation on massively parallel processors like graphics processing units (GPUs). While parallelization and GPUs have received some attention in the statistical community in recent years, the main focus has been on parallelizing matrix operations and sampling techniques (Suchard et al., 2010; da Silva, 2010). To our knowledge, this work marks the first attempt of actively formulating solutions for classical statistical problems in a way that is particularly beneficial for implementation on massively parallel hardware.

The proposed methods are illustrated by conducting a classical statistical analysis of a dataset of 2D chromatograms with more than 140 million spatially correlated observations on a GPU.

2. Model and estimation

We consider spatial functional data on a domain $\mathscr{T} \subseteq \mathbb{R}^d$. Suppose we are given k noisy vectorized functional samples $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k$ each consisting of n observation points. We assume that the observations are generated from the following functional mixed-effect model

$$y_i(\boldsymbol{t}) = \theta_{e(i)}(\boldsymbol{t}) + x_i(\boldsymbol{t}) + \varepsilon_i(\boldsymbol{t}) \tag{1}$$

where $e : \{1, \ldots, k\} \to \{1, \ldots, p\}$ is a factor, $\theta_{e(i)}$ is the fixed functional mean for group $e(i)$, x_i is a zero-mean Gaussian process with covariance function $\tau^2\mathcal{G}$, and ε_i is a Gaussian white noise process with variance σ^2.

A wide variety of functional mixed-effects models have previously been considered. One of the dominant approaches is to model functional effects using smoothing splines (Wahba, 1990). Such constructions are considered by Wang (1998) and Guo (2002). Modeling of mixed effects in terms of penalized splines is considered by Chen and Wang (2011), and Lee et al. (2013) propose a related method based on nested basis functions for spatial mixed-effects models. An alternative approach to functional mixed-effect models

considers the problem in a nonparametric setting, where no distributional or parametric assumptions are made on the random effects. Boularan et al. (1994) considered modeling of growth curves, assuming only that population and individual effects were twice differentiable, and proposed kernel smoothing estimates for the effects. On a similar note, Núñez-Antón et al. (1999) considered a nonparametric three-level model and applied it to speech recognition data. For the use of nonparametric statistical modeling techniques for functional data we refer to the monograph by Ferraty and Vieu (2006), and for a review on functional mixed-effects models we refer to Liu and Guo (2012).

Now, let $\boldsymbol{y}$ be the concatenation of all the vectorized observations of length $N = kn$. The discrete observation $\boldsymbol{y}$ generated by function evaluation at the points $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_n$ in the model (1) may be modeled by a conventional linear mixed-effects model

$$\boldsymbol{y} = \boldsymbol{\Gamma}\boldsymbol{\theta} + \boldsymbol{x} + \boldsymbol{\varepsilon}, \tag{2}$$

where $\boldsymbol{\Gamma} = \mathbb{I}_n \otimes \boldsymbol{\Gamma}_0$ is the design matrix corresponding to the factor e and $\boldsymbol{\theta} \in \mathbb{R}^{np}$ is a vector of parameters describing the group mean functions pointwise, $\boldsymbol{x}$ consists of the spatially correlated effects, $\boldsymbol{x} \sim \mathcal{N}(0, \mathbb{I}_k \otimes \tau^2\boldsymbol{\Sigma})$ with covariance matrix $\boldsymbol{\Sigma} = \left\{\mathcal{G}(\boldsymbol{t}_i, \boldsymbol{t}_j)\right\}_{i,j}$, and $\boldsymbol{\varepsilon}$ is independent, identically distributed Gaussian noise $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2\mathbb{I}_N)$. Since the design is constant across all observations, i.e. given by $\boldsymbol{\Gamma}_0$, the fixed effect $\boldsymbol{\theta}$ can be estimated pointwise. The solution strategy presented below may also be adapted to the situation with a low rank design matrix following Markussen (2013).

Functional mixed-effect models are typically modeled with fixed effects of a functional nature. For simplicity, we parametrize the fixed effect with one parameter per observation point, mimicking classical mixed-effects models. The adaption to functional fixed effects given by a limited number of basis functions can be done following the previously mentioned references. In particular, the computations needed for fixed effects parametrized in terms of

smoothing splines closely follow the computations related to the spatially correlated effect $\boldsymbol{x}$, and the presented methods naturally extend to such parametrizations.

The best linear unbiased prediction for the spatially correlated effects in the model (2) is done by means of the conditional expectation (Robinson, 1991)

$$\mathrm{E}\,[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,] = \big(\mathbb{I}_k \otimes \tau^2\boldsymbol{\Sigma}\big)\,\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\Gamma}\hat{\boldsymbol{\theta}}), \tag{3}$$

where $\boldsymbol{V} = \sigma^2\mathbb{I}_N + \mathbb{I}_k \otimes \tau^2\boldsymbol{\Sigma}$. The variance parameters are typically estimated by minimizing the negative log restricted likelihood (Harville, 1977; Lee et al., 2006)

$$\ell_{\boldsymbol{y}}(\sigma, \tau) = \log\det \boldsymbol{V} + \log\det[\boldsymbol{\Gamma}^\top \boldsymbol{V}^{-1}\boldsymbol{\Gamma}] + (\boldsymbol{y} - \boldsymbol{\Gamma}\hat{\boldsymbol{\theta}})^\top \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\Gamma}\hat{\boldsymbol{\theta}}). \tag{4}$$

For later use it is noted that the last term in the likelihood function can be written as

$$\begin{aligned}(\boldsymbol{y} - \boldsymbol{\Gamma}\hat{\boldsymbol{\theta}})^\top \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\Gamma}\hat{\boldsymbol{\theta}}) &= \frac{1}{\sigma^2}(\boldsymbol{y} - \boldsymbol{\Gamma}\hat{\boldsymbol{\theta}})^\top (\boldsymbol{y} - \boldsymbol{\Gamma}\hat{\boldsymbol{\theta}} - \mathrm{E}\,[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,]) \\ &= \frac{1}{\sigma^2}(\boldsymbol{y} - \boldsymbol{\Gamma}\hat{\boldsymbol{\theta}} - \mathrm{E}\,[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,])^\top (\boldsymbol{y} - \boldsymbol{\Gamma}\hat{\boldsymbol{\theta}} - \mathrm{E}\,[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,]) \\ &\quad + \frac{1}{\sigma^2}\mathrm{E}\,[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,]^\top (\boldsymbol{y} - \boldsymbol{\Gamma}\hat{\boldsymbol{\theta}} - \mathrm{E}\,[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,]).\end{aligned} \tag{5}$$

3. Operator approximations

For many common covariances $\mathcal{G}$ the underlying functional structure of the covariance matrix $\boldsymbol{\Sigma}$ can be exploited, so that one may approximate calculations involving $\boldsymbol{\Sigma}$. The functional counterpart to $\boldsymbol{\Sigma}$ is the integral operator $\mathscr{G}$ given by

$$\mathscr{G}f = \int_{\mathscr{T}} \mathcal{G}(\,\cdot\,, \boldsymbol{t}) f(\boldsymbol{t})\,\mathrm{d}\boldsymbol{t}.$$

To ease notation it is assumed that $k = 1$. The general case follows easily. Furthermore, assume for simplicity that the observations are equidistantly spaced within $[0, 1]^d$. For non-equidistant observations, one can introduce a normalization operator following Markussen (2013). Let $\mathscr{E}_\cdot : \mathbb{R}^n \to \mathcal{C}(\mathscr{T}, \mathbb{R})$ be a linear embedding of the observation space into the space of piecewise linear functions on $\mathscr{T}$. For n large, one has the Riemannian sum approximation of the integral

$$\boldsymbol{\Sigma z} \approx \{n \mathscr{G} \mathscr{E}_{\boldsymbol{z}}(\boldsymbol{t}_i)\}_i. \tag{6}$$

Assuming that $\mathcal{G}$ is two times continuously differentiable within the d-cubes spanned by the observation points, the approximation error can be specified explicitly by applying the trapezoidal rule on the right-hand side integrals, mimicking Proposition 1 in Markussen (2013). The error is of order $\sum_{i=1}^{d} O(n_i^{-1})$ where n_i denotes the number of sample points across data dimension i, i.e. $n = n_1 \cdots n_d$.

Denote by $\mathscr{L} = \mathscr{G}^{-1}$ the precision operator corresponding to $\mathscr{G}$, i.e.

$$\mathscr{L}\mathcal{G}(\,\cdot\,, \boldsymbol{t}) = \delta_{\boldsymbol{t}} \tag{7}$$

where $\delta_{\boldsymbol{t}}$ is the Dirac delta function at $\boldsymbol{t}$. In many cases $\mathscr{L}$ is a differential operator with $\mathcal{G}$ as its corresponding Green's function. For a general introduction to Green's functions we refer to the monograph by Duffy (2001). The relation between covariance functions and differential operators can be used to approximate calculations involving the covariance matrix $\boldsymbol{\Sigma}$.

First we consider the conditional expectation (3). One may rewrite the matrix product, to get

$$\mathrm{E}\,[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,] = \left(\mathbb{I}_n + \frac{\sigma^2}{\tau^2}\boldsymbol{\Sigma}^{-1}\right)^{-1} (\boldsymbol{y} - \boldsymbol{\Gamma}\hat{\boldsymbol{\theta}}).$$

By using the approximation (6) and the fact that inversion is a continuous operation, one can derive (component-wise) operator approximations of the

conditional expectation (3)

$$\widehat{\mathrm{E}}\left[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,\right] = \left(\mathbb{I} + \frac{\sigma^2}{n\tau^2}\mathscr{L}\right)^{-1} \mathscr{E}_{\boldsymbol{y}-\boldsymbol{\Gamma}\hat{\boldsymbol{\theta}}}, \tag{8}$$

which means that the conditional expectation can be approximated by applying an integral operator with smoothing kernel corresponding to the Green's function of $\mathbb{I} + \frac{\sigma^2}{n\tau^2}\mathscr{L}$ on the continuously embedded residual $\boldsymbol{y} - \boldsymbol{\Gamma}\hat{\boldsymbol{\theta}}$. As opposed to the original conditional expectation (3) that requires inversion of a possibly dense covariance matrix, the operator approximation (8) require the inversion of an operator. This may be done explicitly, and the approximation (8) can typically be evaluated in linear time, and may in fact often be evaluated at all observation points in linear time (Markussen, 2013). Furthermore, convolving high-dimensional data with possibly non-isotropic smoothing kernels can be done very efficiently on massively parallel processors (Hartung et al., 2012).

By applying the differential operator $\mathbb{I} + \frac{\sigma^2}{n\tau^2}\mathscr{L}$ on both sides of equation (8) one gets that $f = \widehat{\mathrm{E}}\left[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,\right]$ is the solution to the partial differential equation

$$\mathscr{L} f = \frac{n\tau^2}{\sigma^2}(\mathscr{E}_{\boldsymbol{y}-\boldsymbol{\Gamma}\hat{\boldsymbol{\theta}}} - f). \tag{9}$$

In general, numerical solution of the differential equation (9) is the most efficient choice for obtaining the approximated conditional expectation (8). In particular, GPUs are very suited for efficiently solving differential equations based on finite difference approximations (Micikevicius, 2009).

In the following, point evaluation of $\widehat{\mathrm{E}}\left[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,\right]$ will be assumed to be done at all data points, giving a vector object directly comparable to $\mathrm{E}\left[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,\right]$. Point evaluation is always done after applications of operators, for example differentiation.

Considering the differential equation (9), one can derive a numerically stable expression for the last part of the expanded quadratic term (5). By

inserting the functional approximations of the conditional expectation in the term and using (9), one gets that

$$\widehat{\mathrm{E}}\left[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,\right]^{\top}\left(\mathscr{E}_{\boldsymbol{y}-\boldsymbol{\Gamma}\hat{\boldsymbol{\theta}}}-\widehat{\mathrm{E}}\left[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,\right]\right)=\frac{\sigma^2}{n\tau^2}\widehat{\mathrm{E}}\left[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,\right]^{\top}\mathscr{L}\widehat{\mathrm{E}}\left[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,\right].$$

Assuming that the covariance function $\mathcal{G}$ is positive definite, a square root $\mathscr{K}$ of $\mathscr{L}$ exists, such that $\mathscr{L}=\mathscr{K}^{\dagger}\mathscr{K}$, which means that the last term may also be written as a sum of squares

$$\frac{\sigma^2}{n\tau^2}(\mathscr{K}\widehat{\mathrm{E}}\left[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,\right])^{\top}(\mathscr{K}\widehat{\mathrm{E}}\left[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,\right]).$$

Finally, to approximate the determinant terms in the restricted likelihood function (4), one notes that

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\log\det\left[\mathbb{I}_n+\alpha\boldsymbol{\Sigma}\right]=\mathrm{tr}\big((\mathbb{I}_n+\alpha\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}\big),$$

which means that

$$\log\det[\mathbb{I}_n+\boldsymbol{\Sigma}]=\int_0^1\sum_{\ell=1}^{n}\boldsymbol{e}_\ell^{\top}(\mathbb{I}+\alpha\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}\boldsymbol{e}_\ell\,\mathrm{d}\alpha,$$

where the vectors $\boldsymbol{e}_\ell$ constitute an orthonormal basis for $\mathbb{R}^n$. By approximating the matrix computations with their operator counterparts, one gets that

$$\log\det[\sigma^2\mathbb{I}_n+\tau^2\boldsymbol{\Sigma}]\approx\int_0^1\int_{\mathscr{T}}\left(\alpha\mathbb{I}+\tfrac{\sigma^2}{n\tau^2}\mathscr{L}\right)^{-1}\delta_{\boldsymbol{t}}(\boldsymbol{t})\,\mathrm{d}\boldsymbol{t}\,\mathrm{d}\alpha+n\log\sigma^2.$$

The integral term integrates over a family of Green's functions, and for many common covariance functions $\mathcal{G}$, the integral may be explicitly computed, resulting in constant time computation of the approximated log-determinant.

The explicit link between the covariance $\mathcal{G}$ and the differential operator $\mathscr{L}$ can be convenient in model specification. For some models, it may be natural to start out assuming that the random effect has a specific covariance

function, and for other it may be straightforward to specify the differential operator.

Many well-known covariance functions $\mathcal{G}$ correspond to simple differential operators $\mathscr{L}$ with suitable boundary conditions. This will be illustrated in the following examples.

Example 3.1. Let $\mathscr{T} = [0,1]^d$ and $\mathscr{L} = \partial_{t_1}^2 \cdots \partial_{t_d}^2$. For homogeneous Dirichlet boundary conditions the corresponding Green's function is

$$\mathcal{G}(\boldsymbol{t}, \boldsymbol{t}') = (t_1 \wedge t_1' - t_1 t_1') \cdots (t_d \wedge t_d' - t_d t_d'),$$

which is the covariance of the tied down Brownian bridge on $\mathscr{T}$. Alternatively, assuming homogeneous Dirichlet boundaries along the 0-boundaries, and corresponding Neumann boundaries along the 1-boundaries results in the Green's function

$$\mathcal{G}(\boldsymbol{t}, \boldsymbol{t}') = (t_1 \wedge t_1') \cdots (t_d \wedge t_d'),$$

which is the covariance of the Brownian sheet.

Other boundary conditions leads to e.g. the Brownian bridge on $\mathscr{T}$. Finally, assuming homogeneous Neumann boundary conditions may often be a good choice from a modeling point of view, as this corresponds to a Brownian process with a free level. Even though this will only make $\mathscr{L}$ and the corresponding covariance $\mathcal{G}$ positive semi-definite, all calculations can be done completely analogous to the cases where $\mathscr{L}$ is positive definite. ∘

Example 3.2. Let $\mathscr{T} = [0,1]^d$ and $\mathscr{L} = (-\Delta)^\ell + \varepsilon$ where Δ denotes the Laplace operator, $\epsilon > 0$, and $\ell \geq 2$. Under suitable boundary conditions and with $\varepsilon = 0$, this class of precision operators corresponds to penalizing the squares of derivatives (Wahba and Wendelberger, 1980), which is commonly used for regularization.

For Homogeneous Dirichlet boundary conditions one gets the covariance

$$\mathcal{G}(\boldsymbol{t}, \boldsymbol{t}') = \sum_{i_1,\ldots,i_d=1}^{\infty} \frac{2^d}{\pi^{2\ell}(i_1^2 + \cdots + i_d^2)^\ell + \varepsilon} \prod_{j=1}^{d} \sin(i_j \pi t_j) \sin(i_j \pi t_j'). \tag{10}$$

For Neumann boundaries the covariance function is similar, only with the sine functions substituted by cosines. When $\varepsilon = 0$, the covariance function is no longer positive, but the above expression is well defined, and so in practice one may choose $\varepsilon = 0$.

For some choices of d and ℓ more compact descriptions are available (Duffy, 2001, chap. 5). Finally it is worth noting that these Green's functions may take the value $+\infty$ on the diagonal, corresponding to infinite variance. This happens for example when $d = 2$ and $\ell = 1$. ∘

Example 3.3. Let $\mathscr{T} = \mathbb{R}^d$ and $\mathscr{L} = (\kappa^2 - \Delta)^{\alpha/2}$ with free boundary condtions. Assume that $\alpha = \nu + d/2$, $\kappa > 0$, and $\nu > 0$. This choice of precision $\mathscr{L}$ has the Matérn covariance function (Lindgren et al., 2011) as its Green's function

$$\mathcal{G}(\boldsymbol{t}, \boldsymbol{t}') = \frac{\|\boldsymbol{t} - \boldsymbol{t}'\|^v}{2^{\nu-1}\Gamma(\nu + d/2)(4\pi)^{d/2}\kappa^v} K_v(\kappa\|\boldsymbol{t} - \boldsymbol{t}'\|).$$

∘

Example 3.4. Suppose that $\boldsymbol{x}$ from (2) is a tied down Brownian sheet on $[0, 1]^2$, i.e.

$$\mathcal{G}((t_1, t_2), (t_1', t_2')) = (t_1 \wedge t_1' - t_1 t_1')(t_2 \wedge t_2' - t_2 t_2').$$

The Green's function $\mathcal{G}^\alpha((t_1, t_2), (t_1', t_2'))$ for the differential operator $\mathscr{L} + \alpha\mathbb{I}$ is given by

$$\sum_{i=1}^{\infty} \frac{2\sinh(\frac{\sqrt{\alpha}}{i\pi}(1 - t_2 \vee t_2'))\sinh(\frac{\sqrt{\alpha}}{i\pi}(t_2 \wedge t_2'))}{i\pi\sqrt{\alpha}\sinh(\frac{\sqrt{\alpha}}{i\pi})} \sin(i\pi t_1)\sin(i\pi t_1').$$

With this expression one can explicitly compute (8). Furthermore, one can

derive the following log-determinant approximation

$$\log\det[\sigma^2\mathbb{I}_n + \tau^2\boldsymbol{\Sigma}] \approx n\log\sigma^2 + \sum_{i=1}^{\infty}\log\left(\frac{i\pi\sigma}{\sqrt{n}\tau}\sinh\left(\frac{\tau\sqrt{n}}{i\pi\sigma}\right)\right) \tag{11}$$

which can be evaluated by cutting the sum off at some sufficiently high value of i. This provides an interesting generalization to the known log-determinant approximation for the Brownian bridge under Gaussian noise (Markussen, 2013) which is

$$n\log\sigma^2 + \log\left(\frac{\sigma}{\sqrt{n}\tau}\sinh\left(\frac{\tau\sqrt{n}}{\sigma}\right)\right).$$

Finally, due to the symmetry of the eigenfunctions of $\mathscr{L}$ under Dirichlet and Neumann boundary conditions, the approximation (11) is identical to the expression one would get with Neumann boundary conditions. ∘

Example 3.5. Assume that $\mathscr{L} = (-\Delta)^\ell + \varepsilon$ and $\mathscr{T} = [0,1]^2$ with homogeneous Dirichlet or Neumann boundary conditions. Using (10) the following log-determinant approximation is easily derived

$$\log\det[\sigma^2\mathbb{I}_n + \tau^2\boldsymbol{\Sigma}] \approx n\log\sigma^2 + \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\log\left(1 + \frac{\tau^2 n}{\sigma^2}\frac{1}{\pi^{2\ell}(i^2+j^2)^\ell + \varepsilon}\right). \tag{12}$$

∘

Example 3.6. To compare the computation time of the conditional expectation (3) with the approximation given by the solution of the differential equation (9), the two solutions were calculated for $m \times m$ images. The matrix solution (3) was calculated by efficiently inverting the matrix $\boldsymbol{V}$ in BLAS using the Cholesky decomposition and a single thread on a 3.4 GHz Intel Core i7. The differential equation (9) was solved using the explicit diffusion scheme described in detail in Appendix A. The scheme was implemented in CUDA C and executed on an NVIDIA GeForce GTX 680MX GPU with

1536 CUDA cores. The runtime results, excluding the construction time for the matrix $\boldsymbol{V}$ for the matrix approach, can be seen in Figure 1. We note that for $m = 50$, the runtime of the matrix computation is a factor 1200 slower than the solution of the differential equation. For the given observation sizes, we note that the GPU runtimes only differ slightly, with an average runtime increase of approximately 10% from $m = 10$ to $m = 50$ despite of the factor 25 increase in observation size. This is caused by the GPU not being fully utilized for data sizes in the given range, and the runtime is dominated by memory bandwidth. The runtime increase from $m = 10$ to $m = 1000$ of the GPU implementation was found to be merely a factor 33.

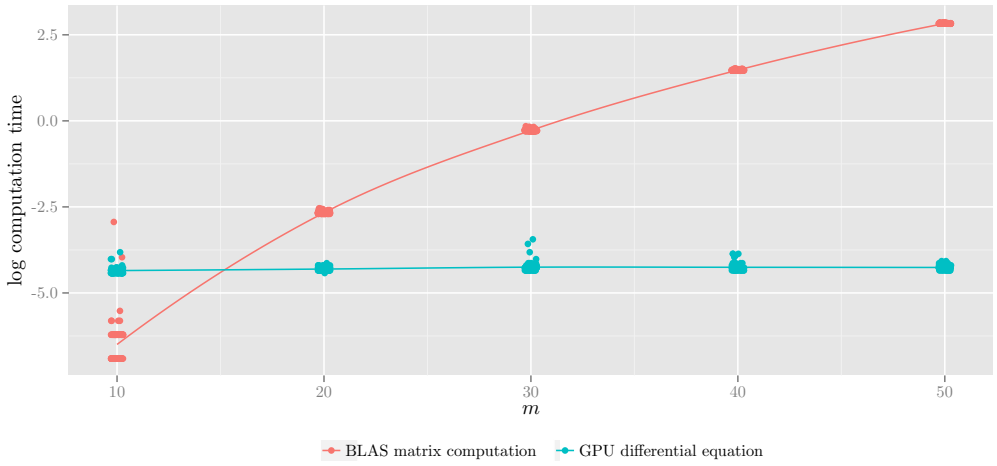


Figure 1: Runtime for the prediction of the conditional expectation using respectively the matrix formulation (3) and the differential equation (9) based on 100 replications for $m = 10, 20, 30, 40, 50$.

○

Example 3.7. To assess the quality of the approximations, observations of tied down Brownian sheets on $[0, 1]^2$ with added Gaussian noise have been generated. The observation points are on an equidistant $m \times m$ grid, for varying values of m. The parameters in terms of the model (2) were $\boldsymbol{\Gamma} = \mathbf{0}$,

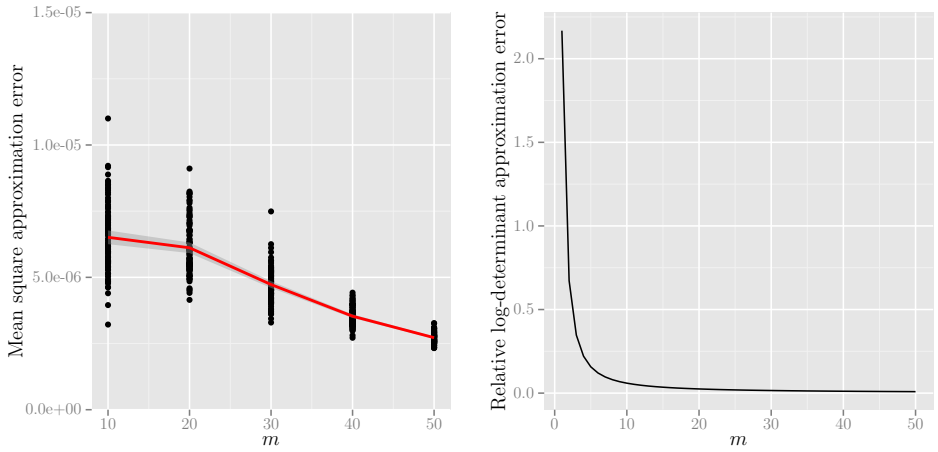


Figure 2: Mean square error of the approximated conditional expectation (8) computed by solving (9), based on 100 replications per m (left) and relative approximation error of the log-determinant (right).

$\sigma^2 = 0.1$, and $\tau^2 = 1$.

Figure 2 shows the mean square error of the approximated conditional expectation (8) with respect to the original conditional expectation (3), and the relative error of the log-determinant approximation (11). The approximated conditional expectation was computed by solving the differential equation (9) using the same setup as described in the previous example. The log-determinant approximation was computed using the formula (11) where n was replaced by $n+1$ in the second term to correct for the Dirichlet boundary conditions, and the sum was cut off after 10,000 terms.

Both approximations clearly improve as m increases. In particular, it is worth noting that the relative error of the log-determinant approximation seems to converge faster than $O(m^{-1})$. ○

3.1. Related models

The model (2) is closely related to other types of models. In particular, assuming that $k = 1$ and $\mathbf{\Gamma} = \mathbf{0}$, one arrives at the classical functional data

model (Ramsay and Silverman, 2005)

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{\varepsilon}. \tag{13}$$

which is typically written in functional form as

$$y(\boldsymbol{t}) = x(\boldsymbol{t}) + \varepsilon(\boldsymbol{t}).$$

One can think of the model (13) as a Bayesian model with $\boldsymbol{x}$ as the prior of the observed function. In this model, the conditional expectation corresponds to the Bayes estimator of the function. Alternatively one can see the conditional expectation as the minimizer of the penalized likelihood function

$$\ell_{\boldsymbol{y}}(x) = (\boldsymbol{y} - x(\boldsymbol{t}_i)_i)^\top (\boldsymbol{y} - x(\boldsymbol{t}_i)_i) + \lambda \int_{\mathscr{T}} x(\boldsymbol{t}) \mathscr{L} x(\boldsymbol{t}) \, \mathrm{d}\boldsymbol{t}, \tag{14}$$

where the λ parameter corresponds to σ^2/τ^2 in the mixed-effects and Bayesian model, and $x(\boldsymbol{t}_i)_i$ is the column vector consisting of the function x evaluated at the points $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_n$. In these cases one would typically estimate parameters by means of marginalized likelihood methods or the generalized cross validation criterion (Craven and Wahba, 1978)

$$\mathrm{GCV}(\lambda) = \frac{n}{(n - \mathrm{df}(\lambda))^2} (\boldsymbol{y} - \hat{\boldsymbol{x}}_\lambda)^\top (\boldsymbol{y} - \hat{\boldsymbol{x}}_\lambda),$$

where $\hat{\boldsymbol{x}}_\lambda$ is the conditional expectation (3), with $\lambda = \sigma^2/\tau^2$, and $\mathrm{df}(\lambda)$ is the trace of the matrix $\frac{1}{2\lambda}\boldsymbol{\Sigma}(\mathbb{I} + \frac{1}{2\lambda}\boldsymbol{\Sigma})^{-1}$. Similarly to the calculations for the log-determinant, one can approximate

$$\mathrm{df}(\lambda) \approx \int_{\mathscr{T}} \mathcal{G}^*_\lambda(\boldsymbol{t}, \boldsymbol{t}) \, \mathrm{d}\boldsymbol{t},$$

where $\mathcal{G}^*_\lambda$ is the Green's function corresponding to the differential operator $\frac{2\lambda}{n}\mathscr{L} + \mathbb{I}$, and thus carry out the generalized cross validation using operator approximations. If a marginalized likelihood approach is preferred, the

likelihood can be approximated using the already presented approximations.

In addition to the connection between the mentioned statistical models, the differential equation (9) naturally links mathematical models governed by this type of equation to the models described here. This in turn allows the use of the mentioned criteria to estimate parameters in such mathematical models.

3.2. Related work

It was noticed by Dolph and Woodbury (1952) that covariance functions of stochastic processes and Green's functions were related through stochastic differential equations. The solution $\boldsymbol{x}$ to the stochastic partial differential equation

$$\mathscr{L}\boldsymbol{x}(\boldsymbol{t}) = \boldsymbol{w}(\boldsymbol{t}), \tag{15}$$

where $\boldsymbol{w}$ is Gaussian white noise and $\mathscr{L}$ is positive definite, is a Gaussian random field with covariance $\mathcal{G}$—the Green's function of $\mathscr{L}$. In a somewhat similar fashion to what has been described in the present paper, Dolph and Woodbury (1952) used this representation to pose prediction problems for continuously observed curves as solutions to differential equations.

More recently, Lindgren et al. (2011) used the connection (15) with $\mathscr{L} = (\kappa^2 - \Delta)^{\alpha/2}$ as the definition of the class of Matérn fields, and derived a computationally efficient Markov representation of the solution. In contrast this paper poses the prediction of the corresponding stochastic differential equation as a partial differential equation in the functional domain, and does not use any explicit representation of the data. Because of this relation to the stochastic differential equation formulation, the presented method can also be generalized to domains that are smooth manifolds, by simply changing the domain of (9), completely analogous to the mainfold generalization by Lindgren et al. (2011). In addition, the presented method can handle a large class of covariance functions since the presented methods only need to

identify the corresponding differential operator and solve a partial differential equation.

4. Example: Glyphosate data

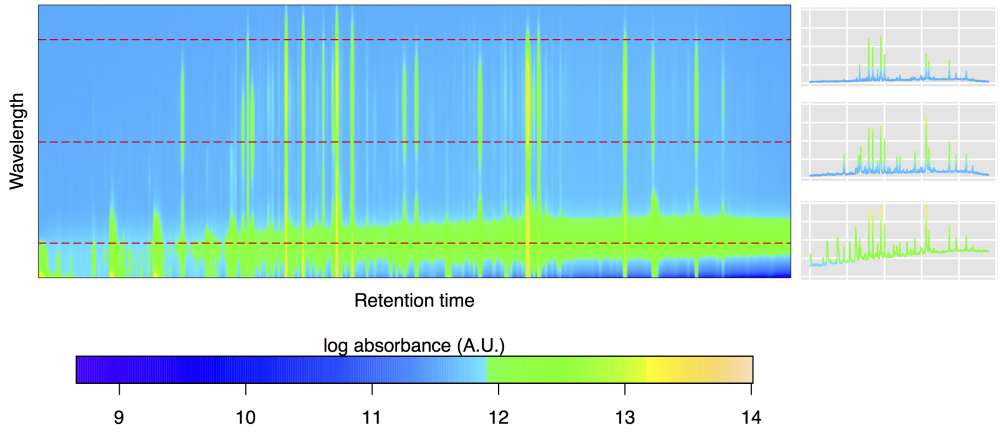


Figure 3: Example of a chromatogram along with absorbance curves for three fixed wavelengths (corresponding to the dashed red lines) on log-scale.

Consider a dataset consisting of $k = 28$ chromatograms $(\boldsymbol{y}_i)_{1 \leq i \leq 28}$, each of which consists of $n = 209 \times 24,000$ (wavelength $\times$ retention time) observations of absorbance (A.U.). The chromatograms have been generated using ultra-high-performing liquid chromatography with diode-array-detection (Petersen et al., 2011). The subjects of the analysis are rapeseed seedlings having been exposed to different levels of glyphosate, commonly known as Roundup®.

The original data have been preprocessed prior to the analysis. The chromatograms have been registered in retention time using a method similar to the so-called TV-L^1 optical flow algorithm (Zach et al., 2007; Rakêt et al., 2011). First, the observations of each glyphosate-level group have been iteratively registered toward the group mean. Next, warping functions of all group means toward the maximum-glyphosate-level group mean are computed. Finally, these warps are applied to the intra-group registered observations,

such that all samples follow a similar coordinate system. For the algorithmic details we refer to Rakêt (2013). Furthermore, the data does not have homogeneous variance; in flat regions, little or no noise is present while noise around peaks is stronger. To alleviate this problem Gaussian noise with variance $2 \cdot 10^{-4}$ has been added to the logarithm of the registered absorbances. Figure 3 displays one of the preprocessed chromatograms, and from the scale of the log absorbance it is clear that the added noise is minuscule compared to the signal.

The logarithm of the absorbance is modeled according to (2)

$$\log(\boldsymbol{y}_i + 1) = \boldsymbol{\theta}_{e(i)} + \boldsymbol{x}_i + \boldsymbol{\varepsilon}_i \tag{16}$$

where the factor $e : \{1, \ldots, 28\} \to \{0, 1, 5, 10, 20, 30, 50\}$ with $p = 7$ levels gives the glyphosate exposure (in μM), and each $\boldsymbol{\theta}$ is $209 \times 24,000$ dimensional. The $\boldsymbol{x}_i$s are independent $209 \times 24,000$ dimensional free Brownian sheets (i.e. $n = 5,016,000$, $N = 140,448,000$ and $\mathscr{L} = \partial_s^2 \partial_t^2$ with Neumann boundary conditions) with variance parameter $\tau^2 = \sigma^2 \xi^2$, and the $\boldsymbol{\varepsilon}_i$s are independent, identically distributed Gaussian noise $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_n)$. Brownian sheets have folds parallel to the axes, which also carry over to the associated posteriors (see e.g. Figure 10). This behaviour makes the Brownian sheet a natural model for the present data, where responses at individual retention times are expected to extend along wavelengths. Furthermore, a multiplicative difference between chromatograms can be expected for this data. This gives a constant level shift after the log transformation. The Neumann boundary conditions (corresponding to a free level, cf. Example 3.1) are then natural for the problem, since the level shift may be captured in the prediction of the spatially correlated effects.

To approximate the restricted likelihood function (4) it is first noted that the determinant terms can be simplified

$$\det \boldsymbol{V} = \sigma^{2N} \det[\mathbb{I} + \xi^2 \boldsymbol{\Sigma}]^k, \qquad \det[\boldsymbol{\Gamma}^\top V^{-1} \boldsymbol{\Gamma}] = \sigma^{-2np} \left(\frac{k}{p}\right)^{np} \det[\mathbb{I} + \xi^2 \boldsymbol{\Sigma}]^{-p},$$

both of which are approximated using the operator approximation (11). In the given parametrization, a closed form restricted maximum likelihood estimate for σ^2 can be derived

$$\hat{\sigma}^2 = \frac{1}{N-np}\Big((\boldsymbol{y} - \boldsymbol{\Gamma}\hat{\boldsymbol{\theta}} - \widehat{\mathrm{E}}\,[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,])^\top(\boldsymbol{y} - \boldsymbol{\Gamma}\hat{\boldsymbol{\theta}} - \widehat{\mathrm{E}}\,[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,]) + \frac{1}{c\xi^2}(\mathscr{K}\widehat{\mathrm{E}}\,[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,])^\top(\mathscr{K}\widehat{\mathrm{E}}\,[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,])\Big).$$

The conditional expectation is computed as the solution to the differential equation (9), which is solved numerically using a finite difference approximation with a stabilized explicit diffusion scheme on a GPU. We refer to Appendix A for the details.

The fixed effects $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \boldsymbol{\theta}_5, \boldsymbol{\theta}_{10}, \boldsymbol{\theta}_{20}, \boldsymbol{\theta}_{30}, \boldsymbol{\theta}_{50}$ are estimated pointwise, and the contrasts to baseline $\boldsymbol{\theta}_0$ can be found in Figure 4. Examples of the predicted spatially correlated effect can be found in Figure 5. We note that the range of the log absorbance values in the predicted spatially correlated effect is around one fifth of the range for the estimated fixed effect contrasts. The estimates of the variance parameters are 91.96 and $1.363 \cdot 10^{-2}$ respectively for ξ and σ.

Figure 6 displays a QQ plot of the conditional residual quantiles against normal quantiles and a scatter plot of conditional residuals against the estimated fixed effects. While the QQ plot shows non-normal tail behavior, this is caused by approximately 0.2% of the observations, and their effect on the estimate of σ is small. The residual plot shows an unnaturally large variation of the residuals corresponding to low absorbance, and for log absorbance levels of around 12.2. Nevertheless, these effects are again caused by very few observations, and the vast majority of the observations, that lie between log absorbance levels of 11.5 and 12, behave as one would expect.

Figure 7 shows the difference in log-likelihood evaluated at the maximum likelihood estimates between the original model (16) and the six models corresponding to collapsing the zero-exposure group with each of the other ex-

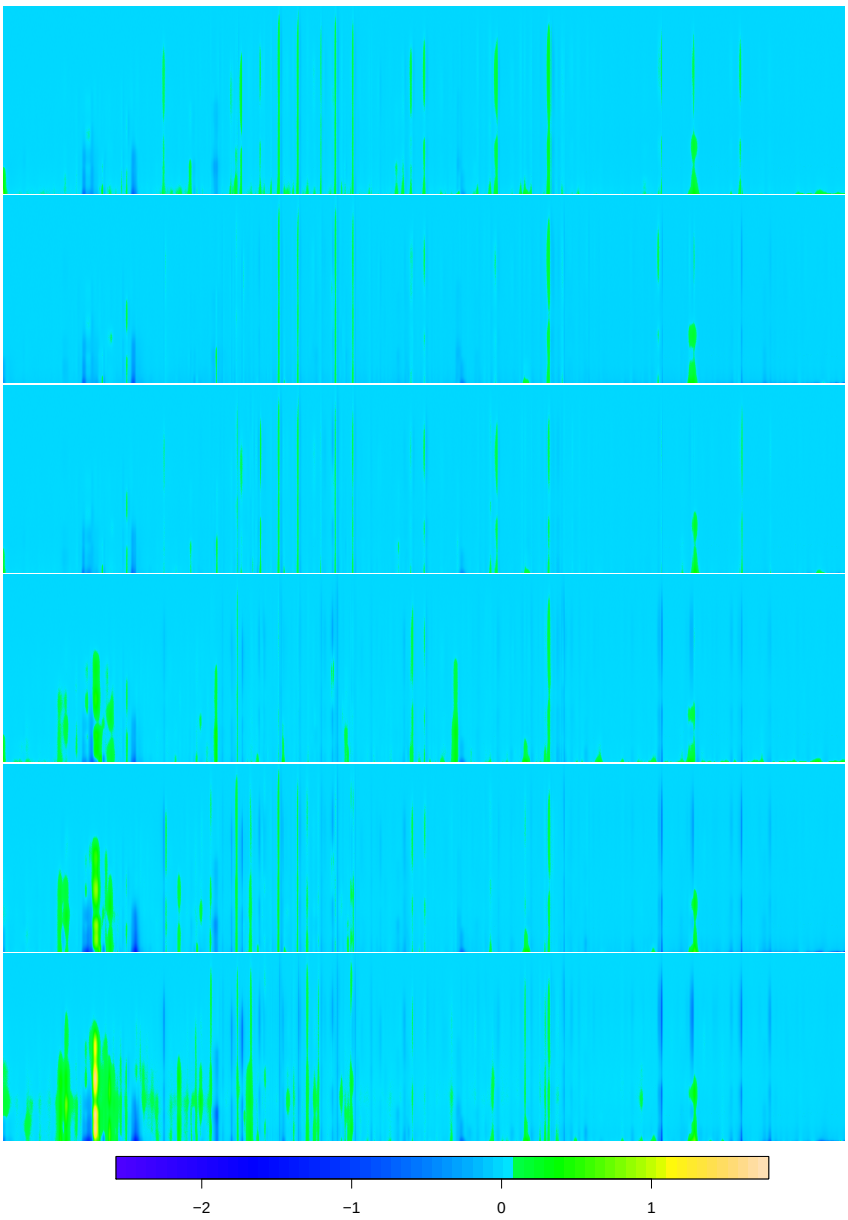


Figure 4: Differences between the estimated fixed effects $\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_5, \hat{\boldsymbol{\theta}}_{10}, \hat{\boldsymbol{\theta}}_{20}, \hat{\boldsymbol{\theta}}_{30}, \hat{\boldsymbol{\theta}}_{50}$ and baseline $\hat{\boldsymbol{\theta}}_0$ (from top to bottom).

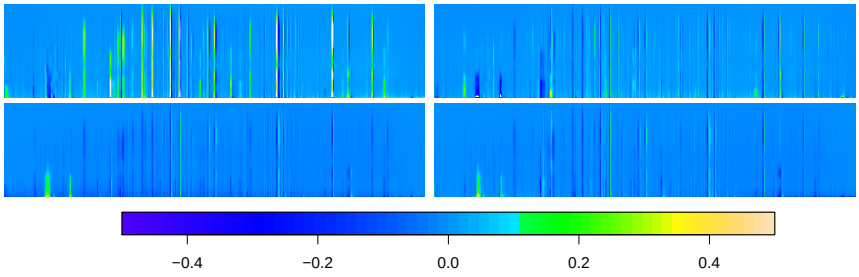


Figure 5: Predictions of the spatially correlated effects $\boldsymbol{x}_i$ for the four observations with glyphosate exposure level $1\,\mu M$.

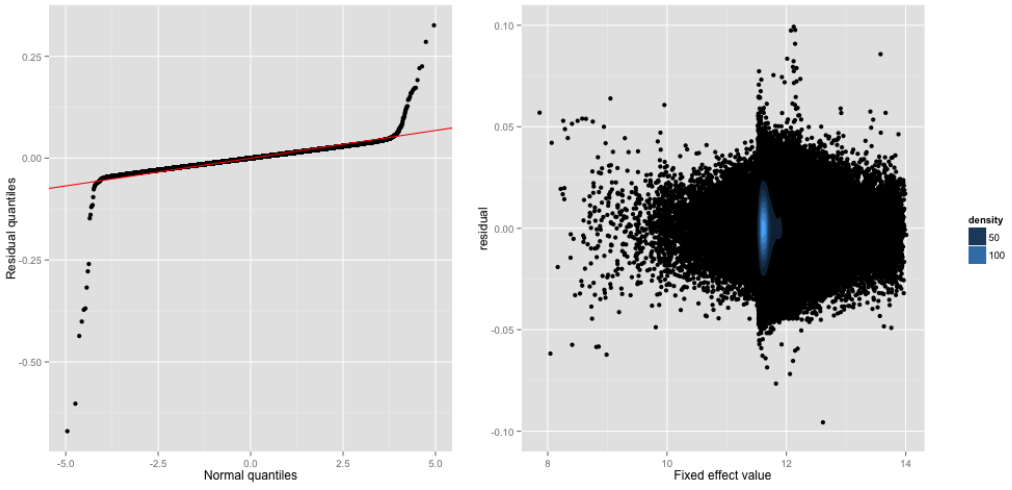


Figure 6: QQ plot and residual plot of a random sample consisting of 0.1% of the conditional glyphosate data residuals (1,404,480 data points), with the 38 most severe outliers removed from the residual plot. The line in the QQ plot shows the estimated standard deviation. For the residual plot the conditional residuals are plotted against the fitted values $\hat{\boldsymbol{\theta}}_{e(i)}$, and the point density is indicated in blue.

posure level groups. The likelihood has been used instead of the restricted likelihood in order to invoke Wilk's likelihood ratio statistic (Pawitan, 2001). Classical asymptotical behavior would prescribe twice the difference in log-likelihood to be approximately χ^2-distributed with degrees of freedom equal to n. In this example the test statistics of order $17 \cdot 10^6$ thus could be evaluated at approximately $5 \cdot 10^6$ degrees of freedom. However, since the validity of a χ^2-test with this many degrees of freedom is questionable, we have not computed p-values. However, there seems to be no doubt concerning the significant difference between the exposure groups. Apart from the $1\mu M$ exposure group, that has a somewhat irregular fixed effect (Figure 4), the log-likelihood differences behave as one would expect; differences increase with glyphosate level. The irregularity of the $1\mu M$ group is mainly caused by one observation with very strong peaks. The prediction of the corresponding spatially correlated effect can be seen in Figure 5 (top left).

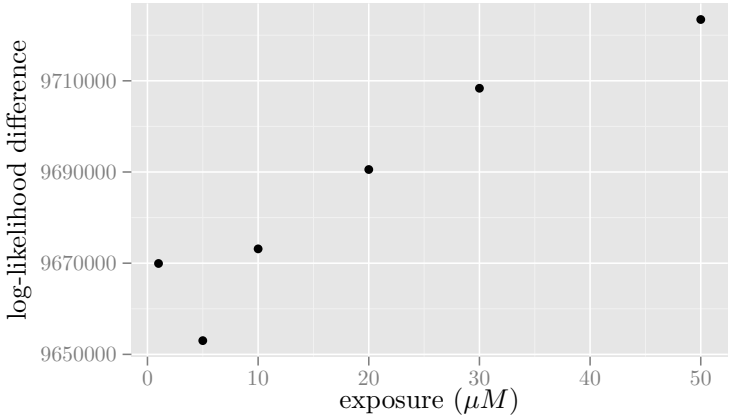


Figure 7: Log-likelihood differences between the model with the marked exposure level and zero-exposure level combined and the full model (16). The likelihood functions have been evaluated at the maximum likelihood estimates.

5. Example: Simulated data

In this simulation example, $k = 25$ images on $\mathscr{T} = [0,1]^2$ sampled at 200×200 equidistant points have been generated from the model

$$\boldsymbol{y}_i = f_\alpha(\boldsymbol{t}_j)_j + g_{\beta_i,\gamma_i}(\boldsymbol{t}_j)_j + \boldsymbol{\varepsilon}_i, \tag{17}$$

where the functions f and g at a point $\boldsymbol{t} = (t_1, t_2)$ are given as

$$f_\alpha(t_1, t_2) = \sin(2\alpha t_1) - \sin(\alpha t_1 t_2)\cos(5t_2) + t_2,$$

$$g_{\beta,\gamma}(t_1, t_2) = g^\star_{\beta,\gamma}(t_1, t_2) - E[g^\star_{\beta,\gamma}(t_1, t_2)],$$

with

$$g^\star_{\beta,\gamma}(t_1, t_2) = \frac{1}{2}(\sin(\beta t_1 t_2)\cos(\gamma t_1)t_2^2 - \cos(\beta\gamma t_2)).$$

Here $\alpha \in \{1, \dots, 10\}$ is a fixed integer, $\beta_i \sim \mathcal{N}(1, 4)$, $\gamma_i \sim \mathcal{N}(1, 9)$, $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_n)$ with $n = 40,000$ and variance $\sigma^2 = 0.1$, and all random variables are independent across the different samples. Images of the functions f_α and $g_{\beta,\gamma}$ with different parameters can be found in figures 8 and 9.

The spatially correlated part of the model is simulated from a parametric random effect model with two degrees of freedom, and it is investigated how the developed model performs under misspecification. This is relevant since one would expect the functional model to be misspecified in most real data applications.

The parametrization and calculations from the previous example trivially carries over to this example. Figures 10 and 11 show examples of the conditional expectation under the assumption of a free Brownian sheet effect and of an effect with biharmonic precision $\mathscr{L} = \Delta\Delta$. For the presented figures $\alpha = 6$ was used and the spatially correlated effects shown correspond to those of Figure 9. In this setting the smoother predictions from the biharmonic precision consistently lead to better predictions of the spatially correlated effects. QQ plots of the conditional residuals can be found in Figure 12. While both plots look very reasonable, it can be seen that the biharmonic

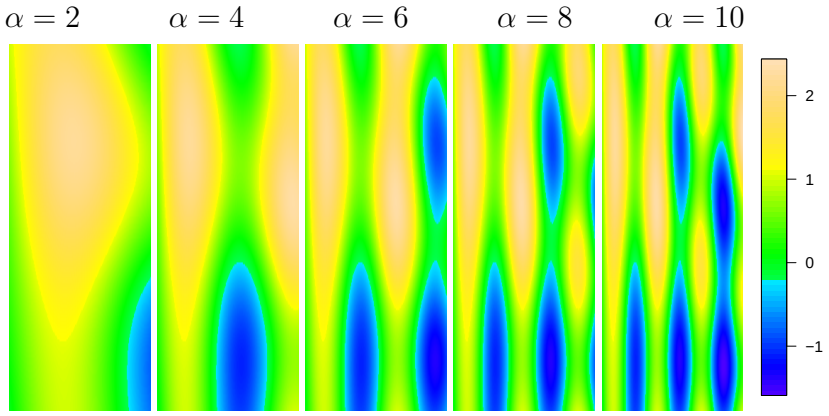


Figure 8: The function f_α for different values of α.

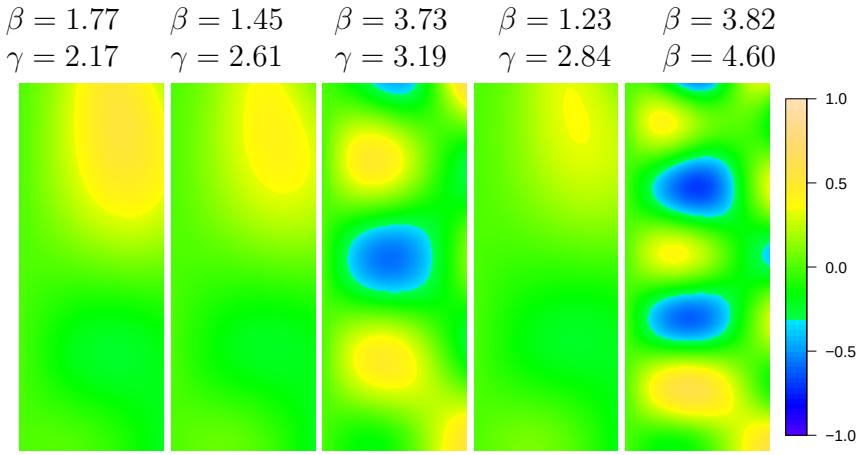


Figure 9: The function $g_{\beta,\gamma}$ with β and γ values simulated following $\beta \sim \mathcal{N}(1, 4)$, $\gamma \sim \mathcal{N}(1, 9)$.

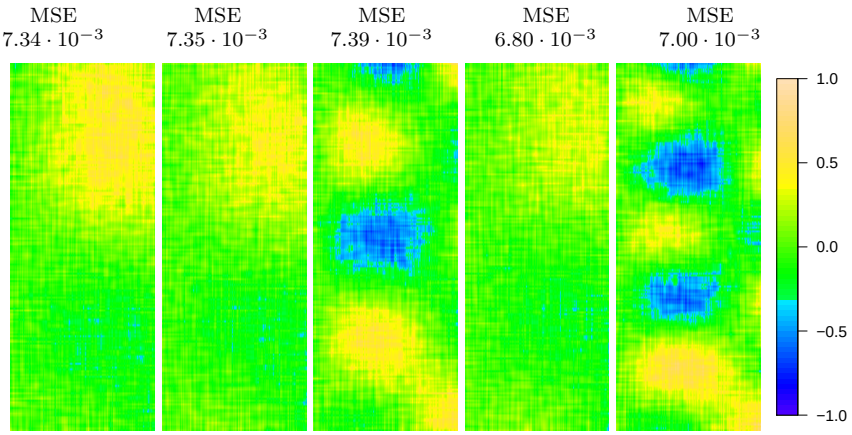


Figure 10: Predictions of the spatially correlated effects from Figure 9 in the model (17) with $\alpha = 6$ under the assumption of a free Brownian sheet effect, with $\hat{\xi} = 0.115$, along with mean squared errors (MSEs).

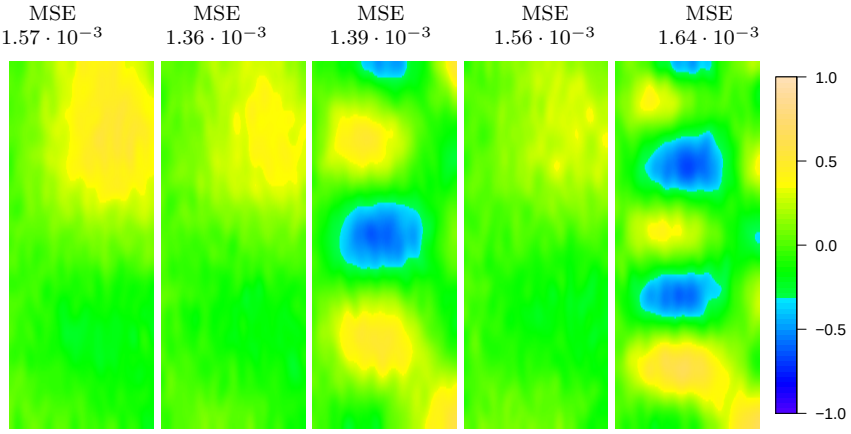


Figure 11: Predictions of the spatially correlated effects from Figure 9 in the model (17) with $\alpha = 6$ under the assumption of a precision operator $\mathscr{L} = \Delta\Delta$, with $\hat{\xi} = 0.0535$, along with mean squared errors (MSEs).

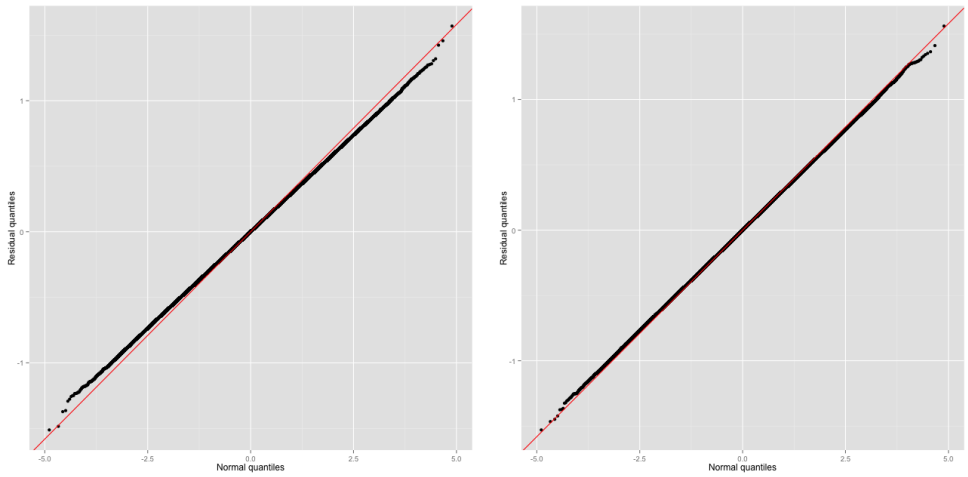


Figure 12: QQ plot of the conditional residuals from the model with a Brownian (left) and biharmonic (right) spatially correlated effect (1,000,000 data points). The lines show the true standard deviation.

model gives a better variance estimate. This is caused by the inherent roughness of the Brownian sheet prior, that will capture some of the noise in the prediction of the spatially correlated effects.

To quantify the behaviour of the variance parameter estimators 100 independent replications (10 for each value of α) of data from the model (17) have been generated. Figure 13 shows a histogram of $\hat{\sigma}^2$ under the assumption of a Brownian and biharmonic correlated effect, respectively. The previously mentioned property that the Brownian sheet effect results in underestimation of the true standard deviation (0.3162) is clearly visible. It is also seen that the biharmonic effect underestimates the standard deviation, although to a much smaller extent.

6. Discussion

This work presents a new method for conducting classical statistical analyses of functional data. By avoiding a direct representation of the data, and doing calculations in the functional domain, the computational complexities

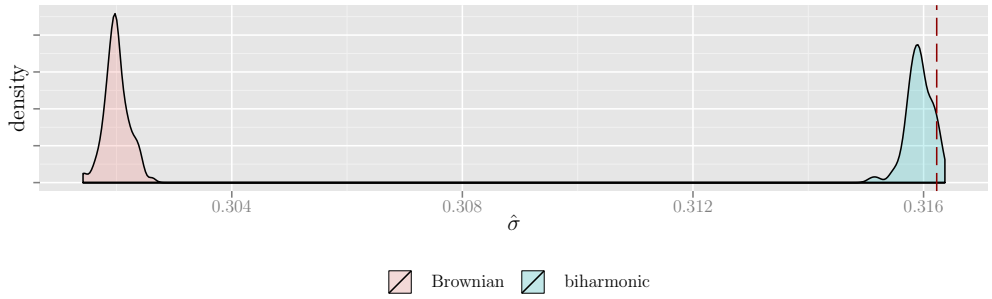


Figure 13: Histograms of parameter estimates in the model (17) under assumption of Brownian and biharmonic correlated effects. The dashed red line in the right histogram shows the true standard deviation.

of the likelihood function and the predictions of spatially correlated effects are significantly reduced. In addition to reducing the computational complexity, the problem of predicting spatially correlated effects may be posed as a partial differential equation. Solvers for such partial differential equations are easily implemented on massively parallel processors, which drastically decrease computation times. CUDA C and R (R Core Team, 2012) code for conducting the presented analyses on NVIDIA graphics hardware is available as supplementary material.

The presented methods allow for analyzing data that are orders of magnitude larger than what has previously been feasible. Using a massively parallel implementation, it was demonstrated that statistical analysis of a dataset of 2D chromatograms, consisting of more than 140 million spatially correlated observations can be done in a matter of minutes.

The considered model was kept simple to illustrate the computational methods, but a number of generalizations can be done. Extensions to vector valued data and more complex designs, including functional fixed effects, are straightforward, and the approximations may be useful in e.g. hierarchical functional models (Staicu et al., 2010). Furthermore, the results are also easily adapted to the case of the domain $\mathscr{T}$ being more complex than what

was considered here, e.g. a smooth manifold. Further generalizations that are relevant from the perspective of achieving valid statistical models, but also require new methodological work, is to allow for variance heterogeneity (Pintore et al., 2006; Yue et al., 2012b,a) and to incorporate data registration directly in the mixed-effects model.

References

Boularan, J., Ferré, L., Vieu, P., 1994. Growth curves: a two-stage nonparametric approach. Journal of Statistical Planning and Inference 38 (3), 327–350.

Chen, H., Wang, Y., 2011. A penalized spline approach to functional mixed effects model analysis. Biometrics 67 (3), 861–870.

Craven, P., Wahba, G., 1978. Smoothing noisy data with spline functions. Numerische Mathematik 31, 377–403.

da Silva, A. F., December 2010. cudaBayesreg: Bayesian Computation in CUDA. The R Journal 2 (2), 48–55.

Dolph, C. L., Woodbury, M. A., 1952. On the relation between Green's functions and covariances of certain stochastic processes and its application to unbiased linear prediction. Transactions of the American Mathematical Society, 519–550.

Duffy, D. G., 2001. Green's Functions with Applications. Chapman & Hall/CRC.

Ferraty, F., Vieu, P., 2006. Nonparametric Functional Data Analysis. Springer.

Grewenig, S., Weickert, J., Bruhn, A., 2010. From box filtering to fast explicit diffusion. In: Goesele, M., Roth, S., Kuijper, A., Schiele, B., Schindler,

K. (Eds.), Pattern Recognition. Vol. 6376 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 533–542.

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

Hartung, S., Shukla, H., Miller, J. P., Pennypacker, C., 2012. GPU acceleration of image convolution using spatially-varying kernel. In: Image Processing (ICIP), 2012 19th IEEE International Conference on. IEEE, pp. 1685–1688.

Harville, D. A., 1977. Maximum likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association 72, 320–340.

Horváth, L., Kokoszka, P., 2012. Inference for functional data with applications. Vol. 200. Springer.

Jordan, M. I., 2011. Message from the President: What are the open problems in Bayesian statistics? ISBA Bulletin 18, 1–4.

Lee, D.-J., Durbán, M., Eilers, P., 2013. Efficient two-dimensional smoothing with P-spline ANOVA mixed models and nested bases. Computational Statistics & Data Analysis 61, 22 – 37.

Lee, Y., Nelder, J. A., Pawitan, Y., 2006. Generalized Linear Models With Random Effects: Unified Analysis Via H-Likelihood. Chapman & Hall.

Lindgren, F., Rue, H., Lindström, J., 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73 (4), 423–498.

Liu, Z., Guo, W., 2012. Functional mixed effects models. Wiley Interdisciplinary Reviews: Computational Statistics 4 (6), 527–534.

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

Micikevicius, P., 2009. 3D finite difference computation on GPUs using CUDA. In: Proceedings of 2nd Workshop on General Purpose Processing on Graphics Processing Units. ACM, pp. 79–84.

Núñez-Antón, V., Rodríguez-Póo, J. M., Vieu, P., 1999. Longitudinal data with nonstationary errors: a nonparametric three-stage approach. Test 8 (1), 201–231.

Pawitan, Y., 2001. In All Likelihood. Oxford University Press.

Petersen, I. L., Tomasi, G., Sørensen, H., Boll, E. S., Hansen, H. C. B., Christensen, J. H., 2011. The use of environmental metabolomics to determine glyphosate level of exposure in rapeseed (Brassica napus L.) seedlings. Environmental Pollution 159 (10), 3071 – 3077.

Pintore, A., Speckman, P., Holmes, C. C., 2006. Spatially adaptive smoothing splines. Biometrika 93 (1), 113–125.

R Core Team, 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
URL `http://www.R-project.org/`

Rakêt, L. L., 2013. Duality based optical flow algorithms with applications. University of Copenhagen prize thesis in Computer Science, Copenhagen University Library.

Rakêt, L. L., Roholm, L., Nielsen, M., Lauze, F., 2011. TV-L^1 optical flow for vector valued images. In: Boykov, Y., Kahl, F., Lempitsky, V., Schmidt, F. (Eds.), Energy Minimization Methods in Computer Vision and Pattern Recognition. Vol. 6819 of Lecture Notes in Computer Science. Springer, pp. 329–343.

Ramsay, J. O., Silverman, B. W., 2005. Functional Data Analysis, 2nd Edition. Springer.

Robinson, G. K., 1991. That BLUP is a good thing: The estimation of random effects. Statistical Science 6 (1), pp. 15–32.

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

Suchard, M. A., Wang, Q., Chan, C., Frelinger, J., Cron, A., West, M., 2010. Understanding GPU programming for statistical computation: Studies in massively parallel massive mixtures. Journal of Computational and Graphical Statistics 19 (2), 419–438.

Wahba, G., 1990. Spline Models for Observational Data. Society for Industrial and Applied Mathematics.

Wahba, G., Wendelberger, J., 1980. Some new mathematical methods for variational objective analysis using splines and cross validation. Monthly Weather Review 108, 1122.

Wang, Y., 1998. Mixed effects smoothing spline analysis of variance. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 60 (1), 159–174.

Weickert, J., Schnörr, C., 2001. Variational optic flow computation with a spatio-temporal smoothness constraint. Journal of Mathematical Imaging and Vision 14, 245–255.

Yue, Y. R., Simpson, D., Lindgren, F., Rue, H., 2012a. Bayesian adaptive smoothing spline using stochastic differential equations. arXiv preprint arXiv:1209.2013.

Yue, Y. R., Speckman, P. L., Sun, D., 2012b. Priors for Bayesian adaptive spline smoothing. Annals of the Institute of Statistical Mathematics 64 (3), 577–613.

Zach, C., Pock, T., Bischof, H., 2007. A duality based approach for realtime TV-L^1 optical flow. In: Hamprecht, F., Schnörr, C., Jähne, B. (Eds.), Pattern Recognition. Vol. 4713 of Lecture Notes in Computer Science. Springer, pp. 214–223.

A. Solving the fourth order PDEs

Consider the differential equation (9) with $\mathscr{L} = \partial_{t_1}^2 \partial_{t_2}^2$ (the case $\mathscr{L} = \Delta\Delta$ is treated similarly), i.e.

$$(\partial_{t_1}^2 \partial_{t_2}^2 + c)f = g. \tag{A.1}$$

This equation is solved using an explicit diffusion scheme, with an added artificial time variable. The solution to (A.1) is found as the steady state of the corresponding diffusion equation.

In order to get numerically stable solutions, $\partial_{t_1}^2 \partial_{t_2}^2$ is approximated using a 5×5 stencil, and furthermore the diffusion is stabilized by evaluating the center point of the stencil at the future time point (Weickert and Schnörr, 2001). This scheme is stable for time steps of 0.125, but the convergence rate is greatly accelerated by using the so-called fast explicit diffusion (FED) method of Grewenig et al. (2010), which cleverly mix stable and unstable time steps. In the following example the procedure is demonstrated for a one-dimensional example.

Example A.1. To illustrate the solution procedure, consider the one-dimensional version of the differential equation (A.1), i.e. the differential equation (9) with $\mathscr{L} = -\partial_t^2$. We approximate $\mathscr{L}$ by a standard five-point stencil, so assuming equidistant observations we get

$$\mathscr{L} f(t)|_{t=t_i} \approx \frac{f(t_{i-2}) - 16f(t_{i-1}) + 30f(t_i) - 16f(t_{i+1}) + f(t_{i+2})}{12(t_i - t_{i-1})^2}$$

The one-dimensional version of the differential equation (A.1) is given by

$$(-\partial_t^2 + c)f = g. \tag{A.2}$$

Instead of considering this equation directly, we introduce an artificial time variable τ and consider the diffusion equation

$$\partial_\tau f(t,\tau) = g(t) - (-\partial_t^2 + c)f(t,\tau), \tag{A.3}$$

where $f(t,0)$ is initialized using the observed data values. The steady state of the differential equation (A.3) in τ, e.g. when $\partial_\tau f(t,\tau) = 0$ will solve the original differential equation (A.2). We discretize

$$\partial_\tau f(t,\tau)|_{\tau=\tau_j} \approx \frac{f(t,\tau_{j+1}) - f(t,\tau_j)}{\tau_{j+1} - \tau_j}$$

and $\mathscr{L} f(t_i,\tau)|_{t=t_i,\tau=\tau_j}$ is approximated by

$$\frac{f(t_{i-2},\tau_j) - 16f(t_{i-1},\tau_j) + 30f(t_i,\tau_{j+1}) - 16f(t_{i+1},\tau_j) + f(t_{i+2},\tau_j)}{12(t_i - t_{i-1})^2},$$

where the future time point τ_{j+1} is used in the term $f(t_i,\tau_{j+1})$ for stability. The equation (A.3) is now solved iteratively by considering its finite difference representation, and taking time steps of size $\tau_{j+1} - \tau_j$, where at each step we solve for $f(t_i,\tau_{j+1})$. ∘

For the glyphosate data from Section 4, the diffusion is assumed to have reached its steady state once the artificial time reaches 5,000, corresponding to 40,000 iterations using a step size of 0.125, or a mere 346 FED steps.

The presented solver has been implemented in CUDA C, in order to utilize the thousands of cores on modern GPUs. The runtime (including writing to GPU memory) for computing the solution to (A.1) for a single $209 \times 24,000$ chromatogram is on average 2.0 seconds on an NVIDIA GeForce GTX 680MX. The resulting average computation time for the restricted likelihood function (4) is 69 seconds on the full glyphosate dataset presented in Section 4.

Chapter 3

Data alignment as a nonlinear effect

3.1 Introduction

In this chapter we present the second and third papers of the thesis.

The first paper, Paper P.2 (Rakêt, Sommer & Markussen 2014), considers the nonlinear mixed-effects model (1.4), where in addition to an additive random effect, a nonlinear warping effect is considered. We present a linearization scheme for doing likelihood inference in the resulting model. This in turn gives a model where the phase and amplitude effects in data are modeled simultaneously and all variance parameters are estimated by means of maximum likelihood.

Estimation of warping functions is often considered in a Bayesian context through maximum a posteriori estimation of warps (for classical examples, see Glasbey & Mardia 1998). By considering the warping function a random effect, it becomes natural to predict warps by means of the posterior in the likelihood framework. Thus many existing methods using maximum a posteriori estimation of warps to preprocess data before further analysis can be considered a single iteration of the presented method.

The model is presented curve data, but can effortlessly be generalized to higher dimensional domains. These generalizations come at a computational cost, but for warps parametrized by a limited number of variables, this cost will be tied to the spatially correlated effects, in which case one can use the framework described in Paper P.1.

The second paper, Paper P.3 (Raket, Grimme, Markussen, Schöner & Igel 2014), considers human movement analysis. The proposed model is an extension of the model considered in Paper P.2. In the movement model, we consider a hierarchical structure of both the fixed amplitude effect and of the warping that in the given setup represents timing of the motion. Furthermore, the model is presented in a slightly more elegant setup where linear fixed effects are modeled using B-splines, which eliminates the need to back-warp noisy data.

A nonlinear mixed-effects model for simultaneous smoothing and registration of functional data

Lars Lau Rakêt[a,*], Stefan Sommer[a], Bo Markussen[b]

[a]*Department of Computer Science, University of Copenhagen, Denmark*
[b]*Department of Mathematical Sciences, University of Copenhagen, Denmark*

Abstract

We consider misaligned functional data, where data registration is necessary for proper statistical analysis. This paper proposes to treat misalignment as a nonlinear random effect, which makes simultaneous likelihood inference for horizontal and vertical effects possible. By simultaneously fitting the model and registering data, the proposed method estimates parameters and predicts random effects more precisely than conventional methods that register data in preprocessing. The ability of the model to estimate both hyperparameters and predict horizontal and vertical effects are illustrated on both simulated and real data.

Keywords: data alignment, functional mixed-effects model, nonlinear mixed-effects model, phase variation, amplitude variation, smoothing

1. Introduction

The current standard practice of analyzing functional data in a number of sequential steps is problematic. Analyses are often carried out by performing one or more independent preprocessing steps prior to the final statistical

*Corresponding author
Email addresses: larslau@diku.dk (Lars Lau Rakêt), sommer@diku.dk (Stefan Sommer), bomar@life.ku.dk (Bo Markussen)

analysis (Ramsay and Silverman, 2005). Typical examples are data registration, pre-smoothing, and dimensionality reduction. Such preprocessing steps can cause problems since the final analysis does not take the resulting data modifications (and their related uncertainty) into account. In the worst case this may invalidate the conclusions of the final analysis.

This paper considers misaligned functional data, where proper registration is key to analyzing the data. Treating data registration as a preprocessing step can cause problems. In particular, noisy observations can skew registration results such that noise rather than signal is aligned. Since this type of overfitting happens prior to the statistical analysis, it will lead to both wrongly predicted warps and underestimation of the noise variance. To deal with these issues we propose to simultaneously do likelihood-based smoothing and data registration in a general class of nonlinear functional mixed-effects models. By computing both registration and smoothing at the same time, we will get the optimal registration given the prediction of the functional mixed-effects and vice versa.

The mixed effects are assumed to be observations of Gaussian processes, and the resulting calculations are carried out by iteratively linearizing the model and estimating parameters from the resulting likelihood function. In addition to allowing estimation of the optimal combination of smoothing and registration, all parameters can be estimated by maximum-likelihood estimation. This contrasts most previous works on simultaneous smoothing and registration (see e.g. Lord et al. (2007) and Kneip and Ramsay (2008)) where parameters have to be adjusted (semi-)manually. Some notable exceptions are Rønn (2001), Gervini and Gasser (2005), and Rønn and Skovgaard (2009) who presents methods for doing full likelihood inference for time-transformed curves, and Allassonnière et al. (2007) who derive a rigorous Bayesian framework for estimating data deformation and related parameters. In contrast to the mentioned works, the model we present seeks to align fixed effects, but allows for serially correlated effects that cannot

be matched across functional samples. Since much functional data contains serially correlated noise, e.g. from the measuring device or individual sample differences, a model that allows the separation of such amplitude variations from the phase variation is a considerable step forward.

It is worth noting the differences with pair-wise data registration as is often employed in for example medical imaging. Instead of the common approach of choosing parameters of the registration model either by heuristic arguments or by cross-validation, incorporating the entire dataset or population in the analysis allows parameters to be estimated by maximum-likelihood inference. In addition, instead of searching for a similarity measure that is invariant to certain types of serially correlated effects, e.g. mutual information (Viola and Wells, 1995), the explicit modeling of the serially correlated effects removes the need for invariance in the similarity measure.

The proposed methods are illustrated and compared to conventional preprocessing alignment on simulated dataset, and a general model for alignment is proposed and evaluated on four real datasets.

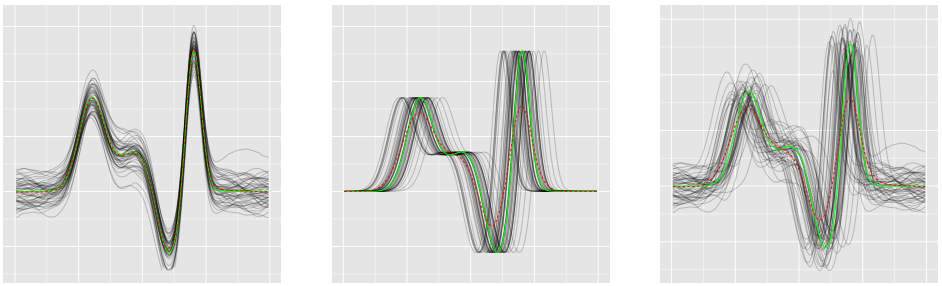
2. Motivation and preliminaries

Two of the major challenges when analyzing functional data are modeling of individual sample effects and aligning of functional samples. Figure 1 illustrates these effects on their own, and in combination, on a one-dimensional functional dataset.

In order to handle individual variation (corresponding to the situation in Figure 1 (a)), one can consider a linear functional mixed-effects model where the kth observation point of functional sample i from the dataset $\boldsymbol{y}$ is assumed to be generated as follows

$$y_i(t_k) = \theta(t_k) + x_i(t_k) + \varepsilon_{ik}, \tag{1}$$

where θ is a fixed effect, x_i is a zero-mean Gaussian process with covariance function $\sigma^2\mathcal{S}$, and ε_{ik} is independent identically distributed Gaussian noise



(a) Individual variations (b) Alignment variations (c) Alignment variations plus individual variations

Figure 1: Different types of variation in a one-dimensional functional dataset. The true underlying curve is shown in green, the average curve is shown in dashed red.

with variance σ^2. Inference in this class of models has been considered in numerous works (Guo, 2002).

In contrast to the vertical variation due to individual sample differences one may encounter horizontal variation due to non-aligned samples (Figure 1 (b)). To align samples, one wishes to estimate so-called *warping functions* v that model the horizontal variation. Similarly to the vertical variation, one may consider the following functional mixed-effects model for this setup

$$y_i(t_k) = \theta(v(t_k, \boldsymbol{w}_i)) + \varepsilon_{ik}, \tag{2}$$

where θ and ε_{ik} are as in (1), and v is a warping function depending on $\boldsymbol{w}_i$ that is a vector of Gaussian parameters with covariance matrix C_0. This model can be considered a nonlinear mixed-effects model, and many known registration algorithms can be thought of as methods for predicting the warping parameters in the model (2), with a known fixed effect θ.

The model (2) has been considered in a statistical setting by Rønn (2001), Gervini and Gasser (2005), and Rønn and Skovgaard (2009), who all consider the problem in a nonparametric maximum likelihood setting. An alternative view is taken in shape analysis, where the interest is on the common shape

θ, while the warping functions are considered nuisance parameters, and data is generally considered free of observation noise. From this viewpoint Kurtek et al. (2011) and Srivastava et al. (2011) have recently proposed an estimation procedure for θ based on the Fisher-Rao metric, that is invariant to diffeomorphic data warping. The mean shape is subsequently used for estimating the warping functions and aligning data. This approach produces state-of-the-art results on numerous examples, but is not generally applicable to all types of data, since the invariance to diffeomorphic warping may lead to overfitting when significant noise is present.

In practice, data often exhibit both vertical and horizontal variation. Figure 1 (c) shows alignment variations of the fixed effect with added serially correlated effects, i.e. a combination of the models (1) and (2)

$$y_i(t_k) = \theta(v(t_k, \boldsymbol{w}_i)) + x_i(t_k) + \varepsilon_{ik}. \tag{3}$$

This type of model describe the fixed effect as a deformation of θ and allows a serially correlated effect x_i that follows the coordinate system of the observation. For some examples, it may be natural to consider the correlated effects x_i in the coordinate system of the fixed effect θ. That model will not be considered here, but inference may be done completely analogous to the procedure described for model (3).

Data modeling following the lines of model (3) have received little attention. One notable exception is the paper by Bigot and Charlier (2011) who consider the sample Fréchet mean as an estimator for θ in the model (3) where the effect x_i also undergo warping by v, and give conditions under which the estimator is consistent. They do however not consider parameter estimation and prediction of random effects. In another related work, Elmi et al. (2011) derive a B-spline based nonlinear mixed-effects model in a maximum likelihood setting. The model allows incorporation of data registration, and is applied to labour curve data, where amplitude variation is modeled parametrically, with random additive and multiplicative effects. Another ap-

plication of this type of model is considered by Chambolle and Pock (2011) in the setting of motion estimation in image sequences. They propose to include a spatially correlated effect that plays the role of lighting differences between the images in question. Their approach, however, does not take the uncertainty related to the prediction of the spatially correlated effect into account in the estimation of the warp, and do not consider the question of parameter estimation.

In the following we will derive inference methodology for the model (3). In contrast to conventional preprocessing approaches that register raw data, the proposed methods can separate horizontal and vertical variation, and allows for maximum-likelihood estimation of all hyperparameters.

3. Estimation

Consider model (3), where the functional data is defined on a domain $\mathscr{T} \subseteq \mathbb{R}$, with m vectorized samples $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$, each of which consists of n points.

The estimation procedures consists of interleaved steps of estimating (a) the fixed effect and the warps; and (b) the parameters of the model and the serially correlated effects. In order to do likelihood estimation of the parameters, we iteratively linearize the model (3) around the given prediction of the warping parameters $\boldsymbol{w}$. This approach is similar to Lindstrom and Bates's (1990) strategy for obtaining maximum likelihood estimates in nonlinear mixed-effects models. It is however more general from the point of view that we predict both linear and nonlinear random effects and estimate the function θ causing the nonlinearity simultaneously.

In pursuance of generality, we will assume that θ is parametrized by its n values at the positions t_k, and that in-between values can be found by interpolation (e.g. cubic spline interpolation). This parametrization mimics the parametrization one would use in a conventional mixed effects model, and follows the well-established convention of interpolation used for motion

estimation in image sequences (Sun et al., 2010). We will assume differentiability of the estimated effect, so the type of interpolation chosen should reflect this. More explicit control of the smoothness of θ can be achieved by specifying a parametric subspace for θ, given by a set of smooth basis functions, or by means of a roughness penalty (Liu and Guo, 2012). Such constructions will not be pursued here.

Using the smoothness of θ, the model (3) can be linearized in the warping parameters $\boldsymbol{w}_i$ around a given prediction $\boldsymbol{w}_i^0$ by means of the first order Taylor approximation,

$$\theta(v(t_k, \boldsymbol{w}_i)) \approx \theta(v(t_k, \boldsymbol{w}_i^0)) + \partial_t \theta(v(t_k, \boldsymbol{w}_i^0)) \nabla_{\boldsymbol{w}} v(t_k, \boldsymbol{w}_i^0)(\boldsymbol{w}_i - \boldsymbol{w}_i^0).$$

The derivative of θ may be computed explicitly from the interpolation function, or it may be estimated by a finite difference approximation.

Let $N = mn$ be the total number of observation points, and let n_w be the dimension of the warping parameters $\boldsymbol{w}_i$. We can write the linearization of model (3) as a vectorized linear mixed-effects model

$$\boldsymbol{y} = \boldsymbol{\theta}^{\boldsymbol{w}^0} + Z(\boldsymbol{w} - \boldsymbol{w}^0) + \boldsymbol{x} + \boldsymbol{\varepsilon} \tag{4}$$

where

$$\begin{gathered}
\boldsymbol{\theta}^{\boldsymbol{w}^0} \approx \{\theta(v(t_k, \boldsymbol{w}_i^0))\}_{i,k} \in \mathbb{R}^N, \\
Z = \operatorname{diag}(Z_i)_{1 \le i \le m}, \qquad Z_i = \{\partial_t \theta(v(t_k, \boldsymbol{w}_i^0)) \nabla_{\boldsymbol{w}} v(t_k, \boldsymbol{w}_i^0)\}_k \in \mathbb{R}^{n \times n_w}, \\
\boldsymbol{w} = (\boldsymbol{w}_i)_{1 \le i \le m} \sim \mathcal{N}_{mn_{\boldsymbol{w}}}(0, \sigma^2 C), \qquad C = \mathbb{I}_m \otimes C_0, \\
\boldsymbol{x} = \{x_i(t_k)\}_{k,i} \sim \mathcal{N}_N(0, \sigma^2 S), \qquad S = \mathbb{I}_m \otimes \{\mathcal{S}(t_k, t_\ell)\}_{k,\ell}, \\
\boldsymbol{\varepsilon} \sim \mathcal{N}_N(0, \sigma^2 \mathbb{I}_N),
\end{gathered}$$

and $\otimes$ denotes the Kronecker product.

The first step of the analysis consists in estimating the fixed effect θ at the positions t_k. Assuming that $\boldsymbol{w}_i^0$ is a correct prediction, back-warping the

observations y_i with $v(t_k, \boldsymbol{w}_i^0)$, and using the non-linearized model we get that

$$y_i(v^{\leftarrow}(t_k, \boldsymbol{w}_i^0)) = \theta(t_k) + x_i(v^{\leftarrow}(t_k, \boldsymbol{w}_i^0)) + \tilde{\varepsilon}_{ik},$$

where $^{\leftarrow}$ indicates inversion of the warp. Ignoring the slight change in variance caused by the back-warping, and hence assuming equal covariances across the different functional samples, the best linear unbiased estimate (Henderson, 1975) of θ given the warp is defined pointwise by

$$\hat{\theta}(t_k) = \frac{1}{m} \sum_{i=1}^{m} y_i(v^{\leftarrow}(t_k, \boldsymbol{w}_i^0)). \tag{5}$$

This estimate should in principle be computed such that the interpolation of the data performed in relation to the back-warping is taken into account. While such computations are feasible, we will not consider that here, since the practical difference is minimal.

With this estimate of θ we estimate the variance parameter σ^2 and possible variance parameters in the covariance matrices C and S from twice the negative log likelihood of the linearized model, which has the form

$$\ell(\sigma^2, C, S) = N \log \sigma^2 + \log \det V + \sigma^{-2}(\boldsymbol{y} - \hat{\boldsymbol{\theta}}^{\boldsymbol{w}^0} + Z\boldsymbol{w}^0)^\top V^{-1}(\boldsymbol{y} - \hat{\boldsymbol{\theta}}^{\boldsymbol{w}^0} + Z\boldsymbol{w}^0),$$

where $V = S + ZCZ^\top + \mathbb{I}_N$. Following Markussen (2013), the double negative log likelihood is rewritten as

$$\begin{aligned} \ell(\sigma^2, C, S) = {} & nm \log \sigma^2 + \log \det V + \sigma^{-2} \boldsymbol{r}^\top \boldsymbol{r} \\ & + \sigma^{-2} \mathrm{E}[\,\boldsymbol{w} \,|\, \boldsymbol{y}\,]^\top C^{-1} \mathrm{E}[\,\boldsymbol{w} \,|\, \boldsymbol{y}\,] \\ & + \sigma^{-2} \mathrm{E}[\,\boldsymbol{x} \,|\, \boldsymbol{y}\,]^\top S^{-1} \mathrm{E}[\,\boldsymbol{x} \,|\, \boldsymbol{y}\,], \end{aligned} \tag{6}$$

where $\boldsymbol{r} = \boldsymbol{y} - \hat{\boldsymbol{\theta}}^{\boldsymbol{w}^0} - Z(\mathrm{E}[\,\boldsymbol{w} \,|\, \boldsymbol{y}\,] - \boldsymbol{w}^0) - \mathrm{E}[\,\boldsymbol{x} \,|\, \boldsymbol{y}\,]$. The best linear unbiased predictor of $\boldsymbol{w}$ and the spatially correlated effects $\boldsymbol{x}$ in the linearized model

are given by their conditional expectations given data (Robinson, 1991)

$$\mathrm{E}[\,\boldsymbol{w}\,|\,\boldsymbol{y}\,] = (C^{-1} + Z^\top(\mathbb{I}_N + S)^{-1}Z)^{-1}Z^\top(\mathbb{I}_N + S)^{-1}(\boldsymbol{y} - \hat{\boldsymbol{\theta}}^{\boldsymbol{w}^0} + Z\boldsymbol{w}^0) \quad (7)$$

and

$$\mathrm{E}[\,\boldsymbol{x}\,|\,\boldsymbol{y}\,] = S(\mathbb{I}_N + S)^{-1}(\boldsymbol{y} - \hat{\boldsymbol{\theta}}^{\boldsymbol{w}^0} - Z(\mathrm{E}[\,\boldsymbol{w}\,|\,\boldsymbol{y}\,] - \boldsymbol{w}^0)). \quad (8)$$

The estimation process is now iterated: Given the estimates of θ and the variance parameters, the new warping parameters $\boldsymbol{w}^0$ are predicted by minimizing the nonlinear negative log posterior (Lindstrom and Bates, 1990)

$$\begin{aligned}\wp(\boldsymbol{w}) &= (\boldsymbol{y} - \hat{\boldsymbol{\theta}}^{\boldsymbol{w}})^\top(S + \mathbb{I}_N)^{-1}(\boldsymbol{y} - \hat{\boldsymbol{\theta}}^{\boldsymbol{w}}) + \boldsymbol{w}^\top C^{-1}\boldsymbol{w} \\ &= (\boldsymbol{y} - \hat{\boldsymbol{\theta}}^{\boldsymbol{w}} - \mathrm{E}[\,\boldsymbol{x}\,|\,\boldsymbol{w},\boldsymbol{y}\,])^\top(\boldsymbol{y} - \hat{\boldsymbol{\theta}}^{\boldsymbol{w}} - \mathrm{E}[\,\boldsymbol{x}\,|\,\boldsymbol{w},\boldsymbol{y}\,]) \\ &\quad + \mathrm{E}[\,\boldsymbol{x}\,|\,\boldsymbol{w},\boldsymbol{y}\,]^\top S^{-1}\mathrm{E}[\,\boldsymbol{x}\,|\,\boldsymbol{w},\boldsymbol{y}\,] + \boldsymbol{w}^\top C^{-1}\boldsymbol{w}\end{aligned} \quad (9)$$

where

$$\mathrm{E}[\,\boldsymbol{x}\,|\,\boldsymbol{w},\boldsymbol{y}\,] = S(S + \mathbb{I}_N)^{-1}(\boldsymbol{y} - \hat{\boldsymbol{\theta}}^{\boldsymbol{w}}).$$

We note how $\wp$ differs from conventional methods of estimating data warps by the explicit modeling of the residual $\boldsymbol{y} - \hat{\boldsymbol{\theta}}^{\boldsymbol{w}}$ in terms of $\mathrm{E}[\,\boldsymbol{x}\,|\,\boldsymbol{w},\boldsymbol{y}\,]$ and the corresponding complexity cost. This way we allow for probable data differences that are captured well by the predicted amplitude effect $\boldsymbol{x}$.

The entire estimation procedure is outlined in Algorithm 1. The inner loop produces the estimates for the fixed effect and the warps. The outer loop produces the estimates for the parameters and the predictions of the serially correlated effects.

4. Experimental results

In this section we study the performance of the estimation procedure. We first consider a simulation study, where we show that the estimation pro-

Algorithm 1: Inference in the model (3).

Data: $\boldsymbol{y}$
Result: Estimates of the fixed effect and variance parameters of the model (3), and the resulting predictions of the serially correlated effects $\boldsymbol{x}$ and the warping parameters $\boldsymbol{w}$

```
// Initialize parameters
```
Initialize $\boldsymbol{w}^0$
Compute $\hat{\boldsymbol{\theta}}^{\boldsymbol{w}^0}$ following (5)
for $i = 1$ *to* $i_{\max}$ **do**
    `// Outer loop: parameters, serially correlated effects`
    Estimate variance parameters and predict serially correlated effects by minimizing the double negative log linearized likelihood (6)
    for $j = 1$ *to* $j_{\max}$ **do**
        `// Inner loop: fixed effect, warping parameters`
        Predict warping parameters by minimizing (9)
        Update linearization points $\boldsymbol{w}^0$ to current prediction
        Recompute $\hat{\boldsymbol{\theta}}^{\boldsymbol{w}^0}$ from (5)
    end
end

cedure is able to correctly predict the parameters of the underlying model used for generating the data, and illustrate how the simultaneous estimation of warps and serially correlated effects increases the precision of the predictions. This is followed by an example of a general class of models that can be used for modeling non-aligned data. We illustrate the models on four real datasets.

4.1. Simulation study

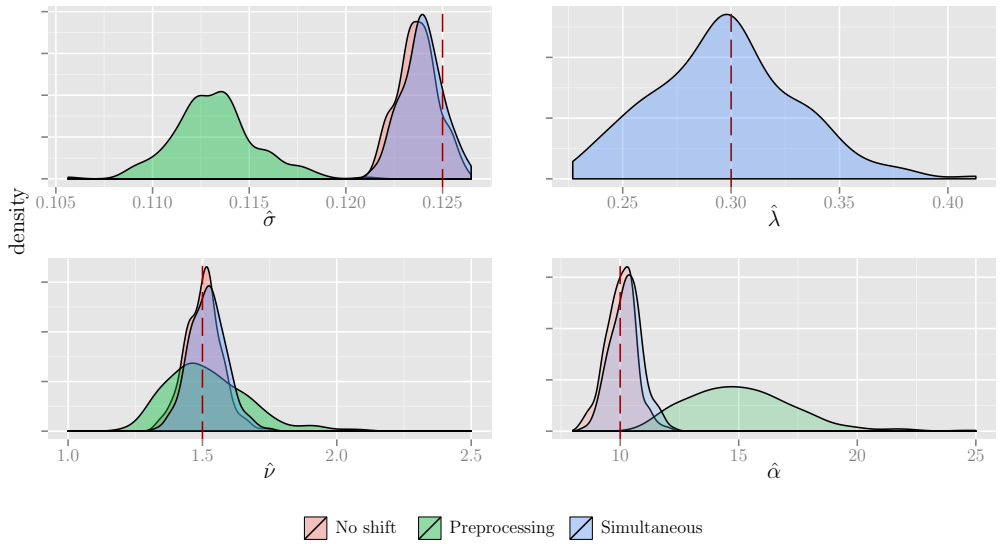


Figure 2: Density plots of variance parameter estimates from 200 independent realizations of the model (10). Seven outliers have been removed in the buttom left plot (4 *Simultaneous*, 3 *Preprocessing*)

Consider synthetic data generated from the model

$$y_i(t_k) = \theta(t_k + w_i) + x_i(t_k) + \varepsilon_{ik} \tag{10}$$

where the w_is and ε_{ik}s are respectively independent identically distributed $\mathcal{N}(0, \sigma^2\lambda^2)$ and $\mathcal{N}(0, \sigma^2)$ variables, the x_is are independent zero-mean Gaussian processes with Matérn covariances $\sigma^2\mathcal{S}$

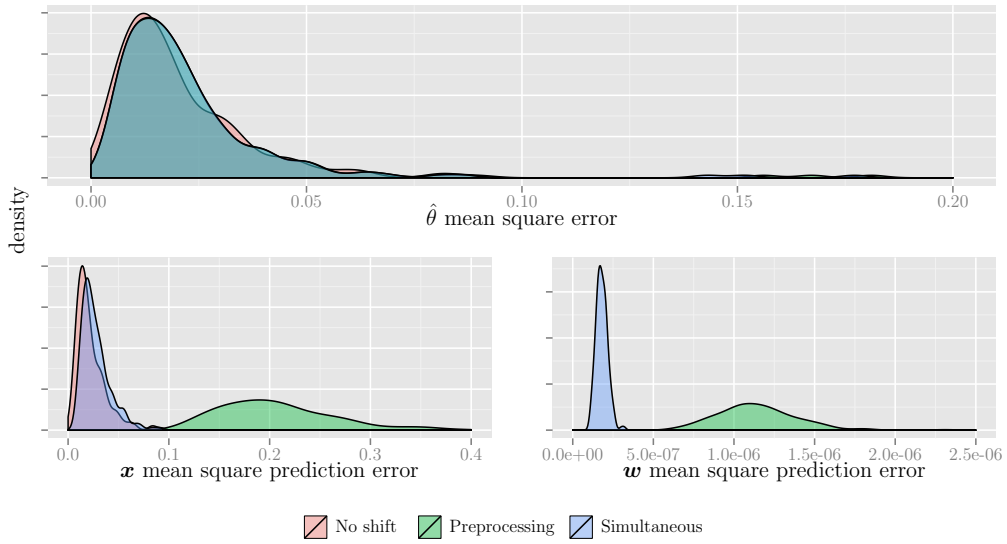


Figure 3: Density plots of mean square errors of $\hat{\theta}$ (top) and predictions of the serially correlated effects x (buttom left) and the warping parameters w (buttom right) from 200 independent realizations of the model (10). Ten outliers have been removed in the buttom left plot (4 *Simultaneous*, 6 *Preprocessing*).

$$\mathcal{S}(s,t) = \frac{1}{\sigma^2\Gamma(\nu)2^{\nu-1}} \left(\sqrt{2\nu}\alpha\|s-t\|\right)^{\nu} K_\nu \left(\sqrt{2\nu}\alpha\|s-t\|\right), \qquad (11)$$

where K_ν is the modified Bessel function of the second kind, and θ is given by

$$\theta(t) = \varphi(t, 0.3, 0.05^2) + \varphi(t, 0.5, 0.1^2) - \varphi(t, 0.6, 0.05^2) + \varphi(t, 0.7, 0.03^2)$$

where $\varphi(t, \mu, \varsigma^2)$ is the normal density with mean μ and variance ς^2. The variance parameters of the model were chosen as follows

$$\sigma = 0.125, \qquad \lambda = 0.3, \qquad \nu = 1.5, \qquad \alpha = 10.$$

Figure 1 (c) displays noiseless samples from this model, i.e. with $\boldsymbol{\varepsilon} = \mathbf{0}$.

We generated 200 independent functional dataset with $m = 50$ functional samples, each consisting of $n = 200$ observation points.

The presented method, denoted by *Simultaneous*, was applied to the simulated datasets. The fixed effect θ was interpolated using a natural cubic spline and the shifts w_i were initialized as the minimizers of the least squares criterion

$$(\boldsymbol{y} - \hat{\boldsymbol{\theta}}^{\boldsymbol{w}})^\top (\boldsymbol{y} - \hat{\boldsymbol{\theta}}^{\boldsymbol{w}}).$$

The algorithm used $i_{\max} = 5$ outer iterations and $j_{\max} = 10$ inner iterations, after which convergence was assumed.

The method was compared to a *Preprocessing* approach where the warping parameters $\boldsymbol{w}$ were predicted by minimizing

$$(\boldsymbol{y} - \hat{\boldsymbol{\theta}}^{\boldsymbol{w}})^\top (\boldsymbol{y} - \hat{\boldsymbol{\theta}}^{\boldsymbol{w}}) + \lambda^{-2} \boldsymbol{w}^\top \boldsymbol{w}$$

using the ground truth λ value. This procedure corresponds to performing the inner iterations of Algorithm 1, which is equivalent to iteratively minimizing the negative log posterior of model (2), i.e. (9) with $S = \mathbf{0}$, updating the estimate θ after each iteration. The resulting predictions were then used to back-warp data (i.e. each y_i was shifted by $-\hat{w}_i$), which was subsequently analyzed using model (1). Finally the simulated datasets without shifts were analyzed using model (1), producing a reference points for the optimal performance of the other methods. We denote this method by *No shift*.

Figure 2 shows density plots of the estimated variance parameters, and Figure 3 displays density plots of the mean square errors of the estimated fixed effects $\hat{\theta}$ evaluated at all observation points t_k, and the predictions of the serially correlated effects $\boldsymbol{x}$ and warping parameters $\boldsymbol{w}$. We see that the proposed method produces good parameter estimates and generally mimics the results of *No shift*. *Preprocessing* on the other hand, generally underestimates the variance of the noise and overestimate the variance of the corre-

lated effects, which is symptomatic of bad alignment. Figure 3 shows that all methods estimate $\hat{\boldsymbol{\theta}}$ reasonably well, but that the ability of *Preprocessing* to predict the serially correlated effects $\boldsymbol{x}$ and the warping parameters $\boldsymbol{w}$ is significantly worse than *Simultaneous*. The simultaneous parameter estimation and prediction of $\boldsymbol{x}$ and $\boldsymbol{w}$ clearly increases the precision of the predictions, and generally mimics the optimal behavior of *No shift*.

4.2. Real data

In this section we consider a general application of model (3) for simultaneously aligning data and modeling individual amplitude effects. We consider four real datasets: Handwriting signature acceleration data (Kneip and Ramsay, 2008); gene expression data (Leng and Müller, 2006); growth velocity data for male subjects in the Berkeley growth study[1]; and spike train data (Wu and Srivastava, 2011). These four datasets has previously been analyzed in the context of data registration by Srivastava et al. (2011), who also give detailed descriptions of the datasets.

For the spatial covariance $\sigma^2\mathcal{S}$ we use the exponential covariance function

$$\mathcal{S}(s,t) = \beta \exp(-\alpha\|s-t\|), \qquad \alpha,\beta \in (0,\infty)$$

which is a special case of the Matérn covariance (11).

We consider two different models for the distribution of the warps of the time axis $[0,1]$. The first one is given by linear interpolation of a discretized Brownian bridge evaluated at the points $t'_1,\ldots,t'_{n_w}$, i.e. the covariance matrix C_0 of $\boldsymbol{w}_i = (w_{i1},\ldots,w_{in_w})$ is given by evaluation of the covariance function

$$\mathcal{C}(t,t') = \lambda^2(t \wedge t' - tt'),$$

where $\wedge$ denotes the minimum operator. The second model instead assumes a Brownian motion, i.e.

[1] `http://www.psych.mcgill.ca/faculty/ramsay/datasets.html`

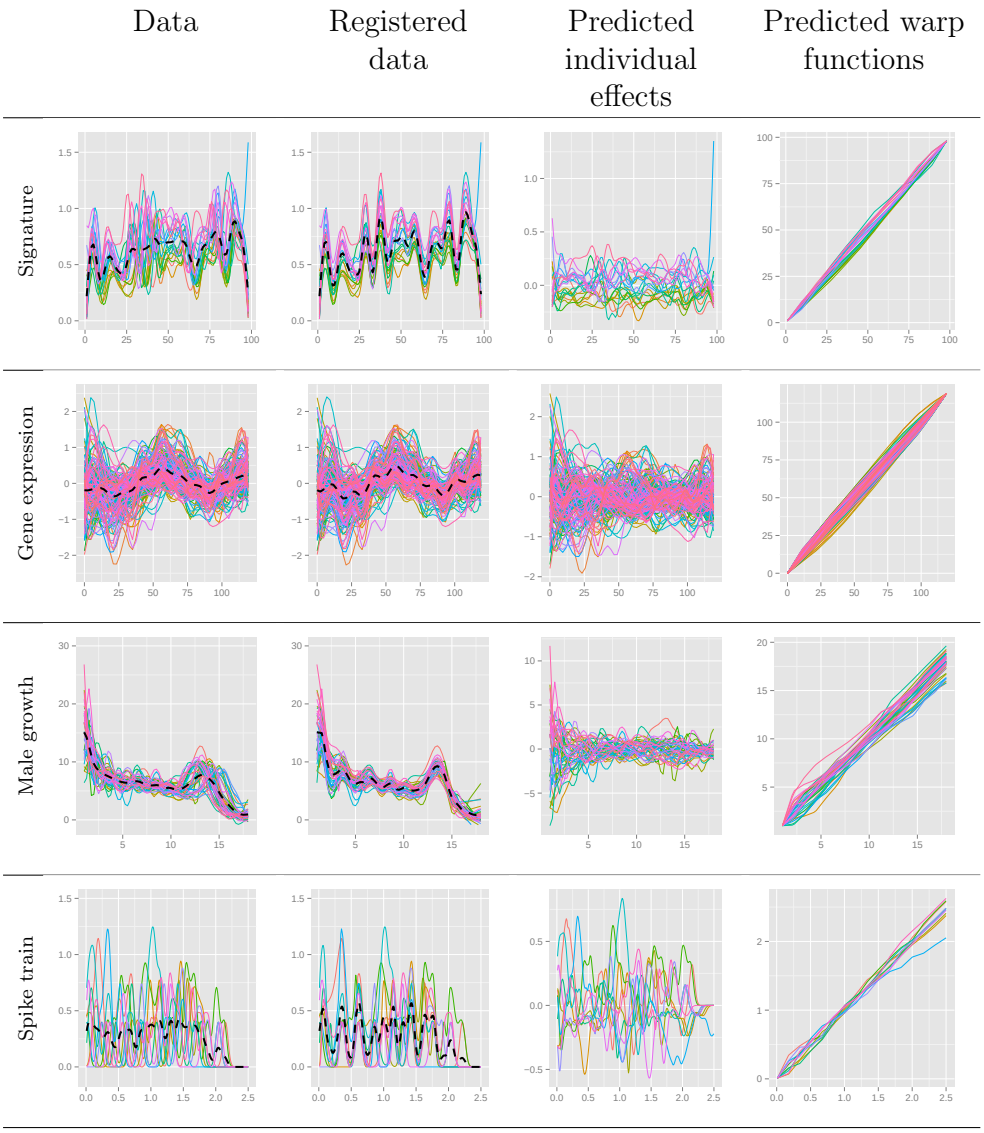


Figure 4: Results of Algorithm 1 on four datasets. Black dashed curves show the mean curve.

$$\mathcal{C}(t, t') = \lambda^2(t \wedge t').$$

The corresponding warping function is

$$v(t_k, \boldsymbol{w}_i) = t_k + \mathcal{E}_{\boldsymbol{w}_i}(t_k),$$

where $\mathcal{E}_{\boldsymbol{w}_i}$ is the linear interpolation function of $\boldsymbol{w}_i$. The Brownian bridge model is useful for data where the observed endpoints of the functional samples correspond to the endpoints of the fixed effect. The Brownian motion model is suitable when the variance of the warp increase with t, and the right endpoints of the functions are different, thus allowing warping of the fixed effect outside of the right endpoint.

While these models assign positive probability to non-diffeomorphic warps, a sufficiently small λ-value will make the predicted warps diffeomorphisms with high probability. As we will see, the maximum likelihood estimates for the given datasets do not lead to any non-diffeomorphic warping functions.

The Brownian bridge model was used for the signature and gene expression data, while the Brownian motion model was used for the male growth data and the spike train data, where warping effects seem to accumulate over time. We used $n_w = 15$ equidistant warping points in $[0, 1]$ and the number of inner iterations $j_{\max}$ was fixed to 10. In order to have comparable results all datasets were normalized to $[0, 1]$ prior to the analysis. We note that since the linearization is a local approximation, we may get stuck in a local minimum depending on the initialization of the warps—in particular if the warps severely overfit the data in a non-diffeomorphic fashion. For this reason we initialize the warps by running 10 inner iterations of minimizing the nonlinear posterior (9) using the parameters $\lambda = 1$, $\beta = 10$ (Brownian bridge) and $\beta = 100$ (Brownian motion), and $\alpha = 1$, which produce initial warps that only deviate slightly from the identity. Table 1 contains information about data sizes, runtime, and number of outer iterations $i_{\max}$ needed for convergence. Table 2 contain the parameter estimates for the four datasets, a relative warp variance (rwv) measure that is computed as the average relative variance contribution of the warp in the linearized model (4), i.e.

$$\frac{1}{N}\sum_{i=1}^{m}\sum_{k=1}^{n}\frac{\mathrm{Var}(\partial_t\theta(v(t_k,\boldsymbol{w}_i^0))\nabla_{\boldsymbol{w}}v(t_k,\boldsymbol{w}_i^0)\boldsymbol{w}_i)}{\mathrm{Var}(y_i(t_k))}.$$

Furthermore, Table 2 hold three different measures of data synchronization (Srivastava et al., 2011).

Table 1: Data sizes, number of iterations needed for convergence, and total runtime (3.4 GHz Intel Core i7, single core) of Algorithm 1 for the four datasets. Convergence was assumed when the variance parameters did not change in two consecutive outer iterations.

	m	n	$i_{\max}$	runtime
Signature	20	98	77	2509 sec
Gene expression	159	52	31	2388 sec
Male growth	39	156	36	1181 sec
Spike train	10	250	51	5883 sec

Table 2: Estimated variance parameters for the four real datasets, along with measures of model fit. *rwv* denotes the average relative data variation ascribed to the warp (see text), and *ls*, *pc*, and *sls* denotes respectively cross-validated least squares, pairwise correlation, and Sobolev least squares (see Srivastava et al. (2011) for details).

	$\hat{\sigma}$	$\hat{\lambda}$	$\hat{\beta}$	$\hat{\alpha}$	*rwv*	*ls*	*pc*	*sls*
Signature	$1.96\cdot10^{-4}$	230	$4.33\cdot10^{5}$	1.65	0.19	0.59	1.07	0.26
Gene expression	$2.03\cdot10^{-4}$	282	$2.12\cdot10^{5}$	2.98	0.05	0.94	1.19	0.81
Male growth	$1.41\cdot10^{-4}$	751	$2.47\cdot10^{5}$	2.86	0.35	0.77	1.11	0.42
Spike train	$1.67\cdot10^{-4}$	536	$1.04\cdot10^{5}$	2.53	0.51	0.77	0.98	0.58

The results of the registration procedure on the four datasets can be seen in Figure 4. Visually, the improved alignment of the curves is immediate. For the signature and male growth data, the data synchronization measures in Table 2 are comparable to the results of Srivastava et al. (2011), while the synchronization for the gene expression and spike train datasets is lower. These less obviously aligned samples however fit well with the goal of the model—we want to decompose data variation into horizontal and vertical components. In particular we see that the average relative warp variance is

only 0.05 for the gene expression data, which indicates that the model found that the amplitude variation in the data was so large, that only large scale structures could be matched.

Finally, we notice that for the gene expression and male growth data, the predicted individual effects seem to imply a bigger variability at the beginning of the samples. Modeling the covariance of the x_is to follow the underlying physical heterogeneity of the data, could possibly improve the model fit.

5. Conclusion and outlook

We have introduced a statistical model that includes data warping for misaligned functional data. Compared to previous works, the model incorporates serially correlated effects explicitly and simultaneously provided estimates of the model parameters. The corresponding estimation algorithm was compared to conventional data analysis where registration is done as preprocessing in the simplest case of misaligned data; the fixed-effect curve being shifted across samples. The comparison demonstrated that parameters were estimated significantly better using the simultaneous approach, and that serially correlated effects were predicted more precisely. Furthermore, we demonstrated that the model can be applied to real data with good registration results.

The proposed model can be extended in several directions. In its presented form, the model allows for parametric warping of data. Replacing the warping parameters $\boldsymbol{w}$ in model (3) by a continuous Gaussian processes would allow for fully non-parametric warping. Furthermore the model is easily generalized to more complex experimental designs or data on high-dimensional domains, such as images.

The presented algorithm is computationally demanding for large data sizes, because of the need to invert the dense covariance matrices of the individual effects. For models with low-dimensional parametric warps, the computationally attractive approximations for predicting individual effects of

Markussen (2013) and Rakêt and Markussen (2014) are directly applicable. New methodological work is however still required in order to use the presented model on very large datasets requiring non-parametric registration, e.g. neuroimage data.

Acknowledgements

The authors wish to thank Wei Wu and Anuj Srivastava for providing the four datasets analyzed in Section 4.2.

References

Allassonnière, S., Amit, Y., Trouvé, A., 2007. Towards a coherent statistical framework for dense deformable template estimation. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69 (1), 3–29.

Bigot, J., Charlier, B., 2011. On the consistency of Fréchet means in deformable models for curve and image analysis. Electronic Journal of Statistics 5, 1054–1089.

Chambolle, A., Pock, T., 2011. A first-order primal-dual algorithm for convex problems with applications to imaging. Journal of Mathematical Imaging and Vision 40, 120–145.

Elmi, A., Ratcliffe, S. J., Parry, S., Guo, W., 2011. A B-spline based semiparametric nonlinear mixed effects model. Journal of Computational and Graphical Statistics 20 (2), 492–509.

Gervini, D., Gasser, T., 2005. Nonparametric maximum likelihood estimation of the structural mean of a sample of curves. Biometrika 92 (4), 801–820.

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

Henderson, C. R., 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics, 423–447.

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

Kurtek, S. A., Srivastava, A., Wu, W., 2011. Signal estimation under random time-warpings and nonlinear signal alignment. In: Advances in Neural Information Processing Systems. pp. 675–683.

Leng, X., Müller, H.-G., 2006. Time ordering of gene coexpression. Biostatistics 7 (4), 569–584.

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

Liu, Z., Guo, W., 2012. Functional mixed effects models. Wiley Interdisciplinary Reviews: Computational Statistics 4 (6), 527–534.

Lord, N., Ho, J., Vemuri, B., oct. 2007. USSR: A unified framework for simultaneous smoothing, segmentation, and registration of multiple images. In: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on. pp. 1 –6.

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

Rakêt, L. L., Markussen, B., 2014. Approximate inference for spatial functional data on massively parallel processors. Computational Statistics & Data Analysis 72, 227 – 240.

Ramsay, J. O., Silverman, B. W., 2005. Functional Data Analysis, 2nd Edition. Springer.

Robinson, G. K., 1991. That BLUP is a good thing: The estimation of random effects. Statistical Science 6 (1), 15–32.

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

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

Srivastava, A., Wu, W., Kurtek, S., Klassen, E., Marron, J., 2011. Registration of functional data using Fisher-Rao metric. arXiv preprint arXiv:1103.3817.

Sun, D., Roth, S., Black, M. J., 2010. Secrets of optical flow estimation and their principles. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, pp. 2432–2439.

Viola, P., Wells, W., 1995. Alignment by maximization of mutual information. In: Computer Vision, 1995. Proceedings., Fifth International Conference on. pp. 16–23.

Wu, W., Srivastava, A., 2011. Towards statistical summaries of spike train data. Journal of Neuroscience Methods 195 (1), 107–110.

Statistical analysis of human arm movements using timing and motion separation

Lars Lau Raket
Department of Computer Sciences
University of Copenhagen
larslau@di.ku.dk

Britta Grimme
Institut für Neuroinformatik
Ruhr-Universität Bochum
britta.grimme@rub.de

Bo Markussen
Department of Mathematical Sciences
University of Copenhagen
bomar@life.ku.dk

Gregor Schöner
Institut für Neuroinformatik
Ruhr-Universität Bochum
gregor.schoener@ini.rub.de

Christian Igel
Department of Computer Science
University of Copenhagen
igel@di.ku.dk

Abstract

A major challenge in the study of human motion is to determine systematic patterns and differences in movements between subjects. In this paper, we consider human arm movements in various obstacle avoidance tasks. We identify the natural types of variation in such data and use these to derive a hierarchical nonlinear functional mixed-effects model, which decomposes movements into task specific aspects, individual effects, and noise. We demonstrate how to do maximum likelihood estimation of the parameters. The learnt individual movement templates are evaluated in a classification scenario. The results support the use of nonlinear mixed-effects models as a well-grounded alternative to conventional movement analysis frameworks.

1 Introduction

The human movement apparatus has more degrees of freedom than needed to realize any particular motor task (e.g., on the level of joint angles and the muscle system) [1]. A major question in the neuroscientific study of movements is how the central nervous systems copes with this freedom; *how does it pick a particular solution out of the endless possibilities?* Research on decoding the formation and arm controlling mechanisms of the central nervous system typically consider the question in terms of optimization principles or invariances. The former assumes that evolutionary and learning processes have led to the maximization of movement benefits. Such optimization principles of arm movement include: maximizing smoothness [2, 3], minimization of movement effort [4], and minimization of end-effector error [5]. The latter research path seeks to answer the question by searching for invariant movement characteristics that can serve as building blocks for a theoretical description of movements. Successful instances of this approach are the discovery of the piecewise planarity of end-effector paths [6], the 2/3 power law [7], and the isochrony principle [8]. The opposing perspective on movement generation focuses on differences and the discrimination of individual characteristics. Instead of looking for building blocks of movements that are common for all subjects, the focus of intersubject variability of movements can enable the identification

of subjects from single movement trajectories. Such motion classification finds many commercial and non-commercial uses. From a product usability point of view, detecting the person behind a given motion can be valuable, as the system can, for example, automatically load custom profiles. Alternative uses of such motion classification are access control and threat detection [9, 10].

When observing movement data, the natural atoms of the analysis are curves that describe position, velocity and/or some other derived quantity of a given repetition. To analyse such functional data, one wants to decompose the observed signal into a common task specific trajectory, subject-specific motion traits, and noise. In the arm movement setting, timing of the movement has a major influence on the data. The conventional approach to modeling functional data with time-warping effects is to pre-align samples under an oversimplified noise model [11], in the hope of eliminating the effects of movement timing. In contrast, we propose an analytic framework where the decomposition of the signal is done simultaneously with the estimation of movement timing effects, so that samples are continually aligned under the estimated noise model. This allows maximal extraction of details of the trajectories.

The desired decomposition into common effect (the task specific aspects of the movement trajectory), individual effect, and noise naturally leads to a mixed-effects formulation [12]. The addition of nonlinear timing effects gives the model the structure of a hierarchical nonlinear mixed-effects model [13]. We present a framework for maximum-likelihood estimation and demonstrate that the method leads to high-quality templates that foster subsequent analysis (e.g., classification). Furthermore, the model can be used for testing of invariance hypotheses across participants and experiments.

In this study, we consider simple grasping tasks with obstacle avoidance. We model the acceleration curves, which are composed of a common pattern and the individual deviations from it, to capture amplitude effects. The timing of the acceleration profiles is determined by individual time warping functions. The identified time warping functions will be of higher quality than conventional estimates, since both timing and movement noise are modeled simultaneously. The high quality of the estimates are demonstrated in a motion classification setup, where the model produces superior classification results.

2 Experimental setup

Ten participants performed a series of simple obstacle avoidance tasks on a table by relocating a cylindrical object from a starting position to a target position. Between the starting position and target, obstacles of varying heights and positions were placed. The participants were instructed to avoid the obstacles by lifting the cylindric object over them, see Figure 1.

The movements were recorded with the Visualeyez (Phoenix Technologies Inc.) motion capture system VZ 4000. Two trackers, each equipped with three cameras, were mounted on the wall 1.5 m above the working surface, so that both systems had an excellent view of the table. A wireless infrared light-emitting diode (IRED) was attached to the object. The trajectories of markers were recorded in three Cartesian dimensions at a sampling rate of 110 Hz based on a reference frame anchored on the table. The starting position projected to the table was taken as the origin of each trajectory in three-dimensional Cartesian space. The acceleration curves considered here were obtained by using finite difference approximations of the raw velocity magnitude data.

Fifteen obstacle avoidance tasks were performed (one for every combination of obstacle height *S*, *M*, or *T* and obstacle distance from starting position $d \in \{15, 22.5, 30, 37.5, 45\}$) as well as a control experiment with no obstacle. The participants repeated each task ten times, giving $m = 100$ functional samples per experiment, and a total of $N_f = 1600$ functional samples in the dataset, with a total data size of $N = 133{,}133$ observation points.

3 Model and estimation

In the following, we describe inference for a single experimental setup. The extension to simultaneous inference for all experiments is straight-forward. For a given experimental setup—an object that needs to be moved to a target and an obstacle that needs to be avoided—we assume there will be a common underlying pattern in all acceleration curves; all m participants will lift the object and move it toward the target, at some point lifting it over the obstacle. We will denote the underlying

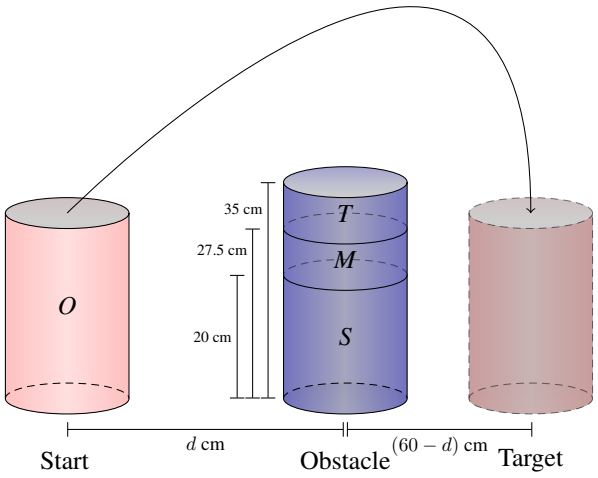


Figure 1: Obstacle avoidance setup. Participants have to move the cylindrical object *O* from the starting position to the target position by lifting it over an obstacle. Obstacles of three different heights, small (*S*), medium (*M*), and tall (*T*), were used in the experiment, and the distance from starting position to obstacle d were varied across the experiments.

common acceleration curve for the experiment by θ. In addition to this fixed acceleration profile, we assume that subject i has a typical deviation φ_i from θ, in other words, the solution strategy of subject i produces the acceleration profile $\theta + \varphi_i$. On top of this ideal trajectory, we assume that there will be a layer of additive correlated noise, that is, for repetition j of the experiment we have an additive random effect x_{ij} that causes deviation from the ideal path. Finally we assume that the data contains observation noise ε_{ij} tied to the tracking system.

In addition to these linear amplitude effects, we assume that each person has an individual, consistent timing of the movement, which corresponds to a time deformation of the acceleration curve. We call this time warping function ν_i. Finally, we assume that person i's jth repetition of the experiment will contain random timing variation around ν_i in the form of a random warping function v_{ij}.

Altogether, this gives the following model for the observed acceleration trajectories across subjects

$$y_{ij}(t) = (\theta + \varphi_i) \circ (\nu_i + v_{ij})(t) + x_{ij}(t) + \varepsilon_{ij}(t) \tag{1}$$

where t denotes time, θ, φ_i, $\nu_i : \mathbb{R} \to \mathbb{R}$ are fixed effects and v_{ij}, x_{ij} and ε_{ij} are random effects. The serially correlated effect x_{ij} is assumed to be a zero-mean Gaussian process with known parametric covariance function $\mathcal{S} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$; the randomness of the warping function v_{ij} is assumed to be completely characterized by a vector of n_w zero-mean Gaussian random variables $\boldsymbol{w}_{ij}$ with covariance matrix $\sigma^2 C$; and ε_{ij} is Gaussian white noise with variance σ^2.

3.1 Estimation

Model (1) is parametrized by a considerable number of parameters, and contains both linear and nonlinear parameters and effects that interact. This renders direct simultaneous likelihood estimation intractable. Instead we propose a scheme where fixed effects and parameters are estimated and random effects are predicted iteratively on three different model levels in an EM-type fashion:

Fixed warp model At the fixed warp level, we fix the person-specific warping effect ν_i at the conditional maximum likelihood estimate, and the random warping function v_{ij} at the predicted values. The resulting model is an approximate linear mixed-effects model with Gaussian random effects x_{ij} and ε_{ij}, that allows direct maximum-likelihood estimation of the remaining fixed effects θ and φ_i.

Nonlinear model At the nonlinear level, we consider the original model (1), and simultaneous perform conditional likelihood estimation of the person specific warping functions and prediction of the random warping functions from the negative log posterior. All other parameters remain fixed.

Linearized model At the linearized level, we consider the first order Taylor approximation of model (1) in the random warp v_{ij}. This linearization is done around the estimate of ν_i plus the

given prediction of v_{ij} from the nonlinear model. The result is again a linear mixed-effects model, for which one can compute the likelihood explicitly, while taking the uncertainty of all random effects—including the nonlinear effect v_{ij}—into account. At this level all variance parameters are estimated using maximum-likelihood estimation.

The estimation/prediction procedure is similar to the algorithmic framework proposed in [14], which in turn builds on the conventional linearization scheme suggested in [13]. In the given setting, we have adapted the estimation procedure to the hierarchical structure of data. Furthermore, we refine the algorithm in [14] in several respects, in particular we avoid back-warping of the noisy data when estimating linear fixed effects. In the following we will go through the steps of the estimation.

Let $\boldsymbol{y}_{ij}$ be the vector of the n_{ij} observations for person i's jth replication of the given experiment, and let $\boldsymbol{y}_i$ and $\boldsymbol{y}$ denote the concatenation of all functional observations of person i in the experiment and all functional observations in the given experiment, respectively. We denote the lengths of these vectors by n_i and n. Furthermore, let $\sigma^2 S_{ij}$, $\sigma^2 S_i$ and $\sigma^2 S$ denote the covariance matrices of respectively $\boldsymbol{x}_{ij} = (x_{ij}(t_k))_k$, $\boldsymbol{x}_i = (\boldsymbol{x}_{ij})_j$, and $\boldsymbol{x} = (\boldsymbol{x}_i)_i$. We note that the index set for k depends on i and j since the covariance matrices S_{ij} vary in size due to the different durations of the movements and because of possible missing values due to sensor occlusions.

We note that all random effects are scaled by the noise standard deviation σ. This parametrization is chosen because it simplifies the likelihood, as we shall see. Finally, we denote the norm induced by the covariance matrix A by $\|\boldsymbol{z}\|_A^2 = \boldsymbol{z}^\top A^{-1}\boldsymbol{z}$.

Fixed warp level We model the underlying curve θ and the person-specific variation around this curve φ_i in the common (warped) functional basis $\boldsymbol{\Phi}$, with weights $\boldsymbol{c} = (c_1, \ldots, c_K)$ for θ and $\boldsymbol{d}_i = (d_{i1}, \ldots, d_{iK})$ for φ_i. We assume that the person-specific variations φ_i are around θ and thus $\sum_i \boldsymbol{d}_i = \boldsymbol{0}$. Furthermore, we penalize the square magnitude of the weights $\boldsymbol{d}_i$ with a weighting factor η. This penalization will direct the alignment process in the direction of the highest possible level of detail in the common profile θ.

For fixed warping functions ν_i and $v_{\boldsymbol{w}ij}$, the negative log likelihood function in $\theta = \boldsymbol{\Phi}\boldsymbol{c}$ becomes

$$\ell(\boldsymbol{c}) = \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{c}\|^2_{\mathbb{I}_n + S}$$

which yields the estimate

$$\hat{\boldsymbol{c}} = (\boldsymbol{\Phi}^\top(\mathbb{I}_n + S)^{-1}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top(\mathbb{I}_n + S)^{-1}\boldsymbol{y}.$$

The penalized profile likelihood for the weights $\boldsymbol{d}_i$ for φ_i is

$$\ell(\boldsymbol{d}_i) = \|\boldsymbol{y}_i - \boldsymbol{\Phi}_i(\hat{\boldsymbol{c}} + \boldsymbol{d}_i)\|^2_{I_{n_i} + S_i} + \eta \boldsymbol{d}_i^\top \boldsymbol{d}_i,$$

which gives the maximum likelihood estimator

$$\hat{\boldsymbol{d}}_i = (\boldsymbol{\Phi}_i^\top(\mathbb{I}_{n_i} + S_i)^{-1}\boldsymbol{\Phi}_i + \eta\mathbb{I}_K)^{-1}\boldsymbol{\Phi}_i^\top(\mathbb{I}_{n_i} + S_i)^{-1}(\boldsymbol{y}_i - \boldsymbol{\Phi}_i\hat{\boldsymbol{c}}).$$

Nonlinear level Similarly to the linear mixed-effects setting [15], it is natural to predict nonlinear random effects from the posterior [13]. Since the conditional negative (profile) log likelihood function in ν_i given the random warping function v_{ij} and the negative (profile) log posterior for $\boldsymbol{w}_{ij}$ coincide, we propose to simultaneously estimate the fixed warping effects ν_i and predict the random warping effects v_{ij} from the joint conditional negative log likelihood/negative log posterior

$$p(\nu_i, \boldsymbol{w}_{ij}) = \sum_j \|\boldsymbol{y}_{ij} - (\hat{\theta} + \hat{\varphi}_i) \circ (\nu_i + v_{ij})(t_k)_k\|^2_{\mathbb{I}_{n_{ij}} + S_{ij}} + \sum_j \|\boldsymbol{w}_{ij}\|^2_C, \tag{2}$$

where the Gaussian variables $\boldsymbol{w}_{ij}$ parametrize the randomness of ν_{ij}. Since these variables can be arbitrarily transformed through the choice of warping function v_{ij}, the assumption that variables are Gaussian is merely one of computational convenience.

Linearized level We can write the linearization of model (1) in the random warping parameters $\boldsymbol{w}_{ij}$ around a given prediction $\boldsymbol{w}_{ij}^0$ as a vectorized linear mixed-effects model

$$\boldsymbol{y} \approx \boldsymbol{\vartheta} + Z(\boldsymbol{w} - \boldsymbol{w}^0) + \boldsymbol{x} + \boldsymbol{\varepsilon} \tag{3}$$

with effects given by

$$\boldsymbol{\vartheta} = \{(\theta + \varphi_i) \circ (\nu_i + v_{ij}^0)(t_k)\}_{ijk} \in \mathbb{R}^n,$$
$$Z = \operatorname{diag}(Z_{ij})_{ij}, \qquad Z_{ij} = \{\partial_{\boldsymbol{t}}(\theta + \varphi_i) \circ (\nu_i + v_{ij}^0)(t_k)(\nabla_{\boldsymbol{w}} v_{ij}^0(t_k))^\top\}_k \in \mathbb{R}^{n_i \times n_w},$$
$$\boldsymbol{w} = (\boldsymbol{w}_{ij})_{ij} \sim \mathcal{N}_{mn_{\boldsymbol{w}}}(0, \sigma^2 \mathbb{I}_m \otimes C), \qquad \boldsymbol{x} \sim \mathcal{N}_n(0, \sigma^2 S), \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2 \mathbb{I}_n),$$

where v_{ij}^0 indicates that the warping function is evaluated at the prediction $\boldsymbol{w}_{ij}^0$ and $\otimes$ denotes the Kronecker product.

Altogether, twice the negative profile log likelihood function for the linearized model (3) is

$$\ell(\sigma^2, C, S) = n \log \sigma^2 + \log \det V + \sigma^{-2} \|\boldsymbol{y} - \hat{\boldsymbol{\vartheta}} + Z\boldsymbol{w}^0\|_V^2 \tag{4}$$

where $V = S + Z(\mathbb{I}_m \otimes C)Z^\top + \mathbb{I}_N$.

3.2 Modeling choices and algorithmic setup

The model (1) has so far only been presented in a general setting. In this section we consider the specific modeling choices. The data has been rescaled using a common scaling for all experiments, such that the span of data values has length 1 and the global timespan is considered as $[0, 1]$.

For the modeling of the amplitude effects, we use a functional basis $\boldsymbol{\Phi}$ consisting of K B-spline functions [11].

We require that the fixed warping function ν_i is a piecewise linear homeomorphism parametrized by n_w equidistant anchor points in $(0, 1)$, and assume that v_{ij} is of the form

$$v_{ij}(t) = t + \mathcal{E}_{ij}(t),$$

where $\mathcal{E}_{ij}(t)$ is the linear interpolation at t of the values $\boldsymbol{w}_{ij}$ placed at the n_w anchor points in $(0, 1)$. We model $\boldsymbol{w}_{ij}$ as a discretely observed zero-drift Brownian Bridge with scale $\sigma^2\gamma^2$ [16, Chapters 8-9], which means that the covariance matrix $\sigma^2 C$ is given by point evaluation of the covariance function

$$\mathcal{C}(t, t') = \sigma^2\gamma^2\, t(1 - t')$$

for $t \leq t'$. When predicting the warps from the negative log posterior we restrict the search space to warps ν_i and $\nu_i + v_{ij}$ that are homeomorphic maps of the domain $[0, 1]$ onto itself. The conditional distribution of ν_{ij} given this restriction is slightly changed. We will however use the original Brownian model as an approximation of the true distribution.

We assume that the sample paths of the serially correlated effects x_{ij} are smooth and that the process is stationary [2]. A natural choice of covariance is then the Matérn covariance with smoothness parameter 2 (producing continuously differentiable sample paths), scale $\sigma^2\tau^2$ and range $1/\alpha$ [17].

Finally, in order to consistently penalize the person specific spline across experiments with varying variance parameters, we will use penalization weights that are normalized with the variance of the amplitude effects, i.e. $\eta = \lambda/(1 + \tau^2)$.

The algorithm for doing inference in model (1) is outlined in Algorithm 1. We have found that $i_{\max} = j_{\max} = 5$ outer and inner loops are sufficient for convergence.

The number of basis functions K, the number of warping anchor points n_w, and the regularization parameter λ were determined by the average 5-fold cross-validation score on each of three experimental setups ($d = 30$ cm and obstacle heights *S*, *M*, and *T*). The models were fitted using the method described in the previous section, and the quality of the models were evaluated by means of classification accuracy of person for a given movement in the test set, using the L^2 distance between the sample and the combined estimated fixed effects $(\theta + \varphi_i) \circ \nu_i$. The cross-validation was done over a grid of the following values $K \in \{20, 25, 30, 35\}$, $n_w \in \{40, 50, 60\}$, and $\lambda \in \{1, 2, 3, 4\}$. The best values were found to be $K = 30$, $n_w = 50$ and $\lambda = 3$.

4 Results and discussion

Variance parameters We fitted the model to data from each of the 15 obstacle avoidance tasks, using experiment-specific variance parameters. The estimated fixed effects can be found in Figure 2,

Algorithm 1: Inference in the model (1).

Data: $\boldsymbol{y}$
`// Initialize parameters`
Compute $\hat{\theta}$ and $\hat{\varphi}_1, \ldots, \hat{\varphi}_m$ assuming an identity warp
Initialize ν_i and $\boldsymbol{w}_{ij}^0$ by minimizing the posterior (2) with $\gamma^2 = 0$
for $i = 1, \ldots, i_{\max}$ **do**
  `// Outer loop`
  Estimate variance parameters by minimizing the linearized likelihood (4)
  for $j = 1, \ldots, j_{\max}$ **do**
    `// Inner loop`
    Estimate and predict warping functions by minimizing the posterior (2)
    Update linearization points $\boldsymbol{w}_{ij}^0$ to current prediction
    Recompute $\hat{\theta}$ and $\hat{\varphi}_1, \ldots, \hat{\varphi}_m$
  end
end

and the estimated variance parameters can be found in Table 1. Higher resolution figures are available in the supplementary material. From Figure 2, the most visible effect is that the intersubject variability seem to decrease with obstacle distance. Furthermore, duration of the movement also seems to increase with obstacle height. When considering the estimated variance parameters in Table 1, the variability of the estimated noise standard deviation σ is perhaps mildly disturbing, but is likely caused by acceleration spikes that are not accounted for in the model. The most interesting parameters are the standard deviation $\sigma\tau$ of the serially correlated effects x_{ij}, and the scale $\sigma\gamma$ of the random warping functions v_{ij}. We note that both the estimated scales of the serially correlated effects $\hat{\sigma}\hat{\tau}$ and of the random warping functions $\hat{\sigma}\hat{\gamma}$ seem quite stable across the experiments, with no clear pattern in the variability. One could have expected that the variability was tied to the experimental setup, and that a higher variability in timing was present for the tall obstacle size, where the movement durations appear longer than for the small and medium obstacles, for example. We note that since starting position and velocity are known, we can directly map the estimated point correspondences back into three-dimensional space to conduct the analysis on the spatial paths.

$(10^6 \cdot \hat{\sigma}, 10^4 \cdot \hat{\sigma}\hat{\tau}, \hat{\alpha}, 10^3 \cdot \hat{\sigma}\hat{\gamma})$	*S*	*M*	*T*
15.0 cm	$(151, 264, 234, 140)$	$(125, 234, 229, 104)$	$(121, 253, 224, 106)$
22.5 cm	$(193, 263, 251, 104)$	$(164, 278, 222, 128)$	$(115, 239, 217, 126)$
30.0 cm	*(224, 260, 228, 123)*	*(157, 251, 244, 115)*	*(227, 233, 237, 113)*
37.5 cm	$(225, 253, 255, 111)$	$(283, 268, 244, 134)$	$(117, 241, 207, 126)$
45.0 cm	$(155, 238, 239, 114)$	$(214, 256, 262, 98)$	$(115, 253, 206, 117)$

Table 1: Estimates of the variance parameters in the 15 obstacle avoidance setups. *Italics* indicates data used in the cross-validation to determine the model parameters.

Classification An interesting use of movement data is person classification based on motion traits. Such applications are becoming increasingly relevant with the recent technological advances in motion tracking systems, and the growing array of digital sensors in handheld consumer electronics.

We consider template based classification, where a characteristic acceleration profile is calculated for each subject, and used for classification. In the following, we describe the different methods considered. All stated parameters have been chosen by 5-fold cross-validation on the experiments with obstacle distance $d = 30.0$ cm. The used grids are given in the supplementary material. *Nearest Centroid* (NC) classification estimates the pointwise, unaligned mean functions and uses these for classification using the L^2 distance. *Nearest Centroid Percentual time* (NCP) classification aligns the samples linearly according to percentual motion time, where all motion endpoints correspond to 100% percentual time. The classification is then done analogously to NC. *Modified Band Median* (MBM) classification estimates templates using the modified band median proposed in [18], which under certain conditions is a consistent estimator of the underlying fixed amplitude effects warped

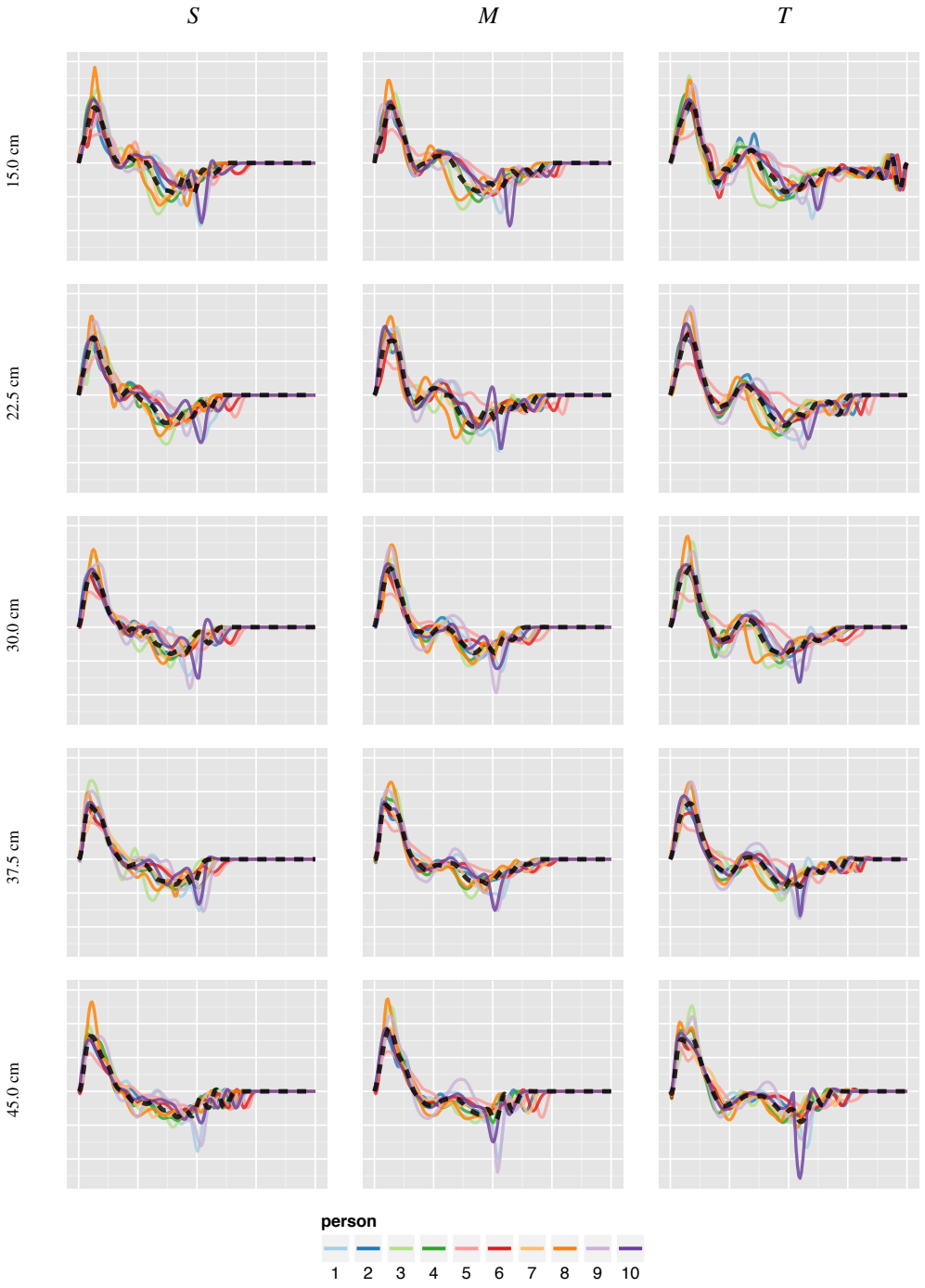


Figure 2: Estimated fixed effects $(\theta + \varphi_i) \circ \nu_i$ in the 15 obstacle avoidance experiments. The dashed curve shows the estimate for θ.

according to the modified band medians of the warping functions. Classification is done using L^2 distance to the estimated templates. In the computations we count the number of bands defined by $J = 4$ curves [18, Section 2.2]. *Robust Manifold Embedding* (RME) classification estimates templates using the robust manifold embedding algorithm proposed in [19], which, assuming that data lies on a low-dimensional smooth manifold, approximates the geodesic distance and computes the

empirical Fréchet median function. Classification is done using L^2 distance to the estimated templates. *Fisher-Rao* (FR) classification estimates templates as Karcher means under the Fisher-Rao Riemannian metric [20] of the data represented using a single principal component [21]. Classification is done using L^2 distance to the estimated templates. *Elastic Fisher-Rao* (FR_E) classification estimates templates analogously to FR, but classifies using the weighted sum of elastic amplitude and phase distances [22, Definition 1 and Section 3.1]. The phase distance was weighted by a factor 1.5. The proposed *Timing and Motion Separation* (TMS) classification estimates templates of the fixed effects $(\theta + \varphi_i) \circ \nu_i$ using Algorithm 1. Classification is done using L^2 distance to the estimated templates. *Posterior Timing and Motion Separation* (TMS_P) classification estimates templates analogously to TMS, but classifies using distance measured in the negative log posterior (2) as a function of the test samples.

We evaluate classification accuracy using 5-fold cross-validation, which means that eight samples are available in the training set for every person. The folds of the cross-validation are chosen chronologically, such that the first fold contains replications 1 and 2, the second contains 3 and 4 and so on. The results are available in Table 2. We see that TMS and TMS_P achieves the highest classification rates, followed by FR_E and RME.

The classification results across all methods suggests two interesting phenomena, namely that the difficulty of the classification task increase with obstacle distance, and is also markedly increased for the medium obstacle height *M*, compared to *S* and *T*. These observations suggest that the movements are either more tightly clustered together across subjects or contain a higher level of random variability for greater obstacle distances, and for the medium height obstacle.

d	obstacle	NC	NCP	MBM	RME	FR	FR_E	TMS	TMS_P
15.0 cm	*S*	0.36	0.48	0.53	0.55	0.47	**0.62**	0.56	0.53
	M	0.36	0.46	0.38	0.41	0.36	**0.47**	0.44	0.39
	T	0.41	0.47	0.41	0.49	0.32	0.49	**0.52**	0.50
22.5 cm	*S*	0.36	0.49	0.34	0.37	0.44	0.50	**0.52**	0.47
	M	0.38	0.44	0.42	0.46	0.32	0.42	**0.47**	0.45
	T	0.36	0.49	0.45	0.46	0.48	**0.53**	0.48	0.42
30.0 cm	*S*	0.27	0.29	*0.37*	0.41	*0.40*	*0.46*	***0.47***	*0.43*
	M	0.30	0.42	*0.38*	0.40	*0.36*	*0.46*	***0.49***	*0.44*
	T	0.37	0.44	*0.42*	0.50	*0.37*	*0.39*	*0.48*	***0.50***
37.5 cm	*S*	0.28	0.45	0.41	0.42	0.36	0.39	0.52	**0.58**
	M	0.26	0.33	0.33	0.35	0.35	0.32	**0.39**	**0.39**
	T	0.31	0.43	0.38	0.40	**0.50**	0.49	0.39	0.42
45.0 cm	*S*	0.25	0.38	0.33	0.32	0.32	0.37	**0.45**	0.44
	M	0.29	0.31	0.29	0.38	0.36	0.38	0.38	**0.50**
	T	0.29	0.39	0.45	0.48	0.39	0.44	**0.50**	0.45
average		0.322	0.418	0.393	0.426	0.387	0.449	**0.470**	0.461

Table 2: Classification accuracies of various methods. **Bold** indicates best result(s), *italics* indicates that the given experiments were used for training.

Hypotheses testing The model allows for testing of hypotheses about human motion. In the current form, the described setup allows for testing of motion hypotheses through the parameters of the model. Classical asymptotic behavior would suggest that the difference in twice the log likelihood functions (4) under the model and hypothesis can be approximated by a χ^2-distribution with degrees of freedom given by the difference in parameter count under the model and under the hypothesis. In practice, the validity of this approximation is often questionable for functional data [23]. For the given dataset, fitting a combined model to all 16 experiments with individual parameters for each setup gives a likelihood value (4) of -1417861. Assuming a common noise variance parameter σ^2 across experiments increases the value to-1414632. Finally, assuming common noise variance σ^2 and warp variance γ^2 gives a likelihood value of -1402531. The test statistics of going from the first model to the second is thus 3229 and from the second to the third is 12101, which when evaluated against a χ^2-distribution with 15 degrees of freedom gives p-values that are essentially zero. Note however that the *assumed true* model that measurement noise is equal across experiments seems much more likely than the hypothesis that timing variation is comparable across experiments.

In order to state credible p-values we need additional model validation, and to use either specialized tests for the functional mixed-effects setting [24], or to approximate the distribution of the test statistic better, for example by simulation studies.

5 Conclusions

We have proposed a statistical framework for modeling of human arm movements. The hierarchical nonlinear mixed-effects model systematically decomposes movements into common effects, individual effects, and noise and considers nonlinear timing effects. We have outlined a method for doing likelihood estimation in the model, and analyzed a dataset consisting of acceleration trajectories in an obstacle avoidance task. The quality of the estimates were evaluated in a classification task, where our model produced superior results compared to state-of-the-art template based curve classification methods. These results indicate that our templates are both more consistent and richer in detail. Because of its conceptual properties and our experiments, we suggest nonlinear mixed-effects modeling as a preferable alternative to conventional movement analysis frameworks.

References

[1] N. A. Bernshtein. *Co-ordination and regulation of movements*. Oxford, New York, Pergamon Press, 1967.

[2] T. Flash and N. Hogan. The coordination of arm movements: an experimentally confirmed mathematical model. *The Journal of Neuroscience*, 5(7):1688–1703, 1985.

[3] Y. Uno, M. Kawato, and R. Suzuki. Formation and control of optimal trajectory in human multijoint arm movement. *Biological cybernetics*, 61(2):89–101, 1989.

[4] Z. Hasan. Optimized movement trajectories and joint stiffness in unperturbed, inertially loaded movements. *Biological cybernetics*, 53(6):373–382, 1986.

[5] C. M. Harris and D. M. Wolpert. Signal-dependent noise determines motor planning. *Nature*, 394(6695):780–784, 1998.

[6] J. F. Soechting and C. A. Terzuolo. Organization of arm movements in three-dimensional space. Wrist motion is piecewise planar. *Neuroscience*, 23(1):53–61, 1987.

[7] F. Lacquaniti, C. Terzuolo, and P. Viviani. The law relating the kinematic and figural aspects of drawing movements. *Acta psychologica*, 54(1):115–130, 1983.

[8] P. Viviani and G. McCollum. The relation between linear extent and velocity in drawing movements. *Neuroscience*, 10(1):211–218, 1983.

[9] W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 34(3):334–352, 2004.

[10] T. Ko. A survey on behavior analysis in video surveillance for homeland security applications. In *37th IEEE Applied Imagery Pattern Recognition Workshop, 2008*, pages 1–8. IEEE, 2008.

[11] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, second edition, 2005.

[12] J. C. Pinheiro and D. M. Bates. *Mixed effects models in S and S-PLUS*. Springer, 2000.

[13] M. J. Lindstrom and D. M. Bates. Nonlinear mixed effects models for repeated measures data. *Biometrics*, 46(3):673–687, 1990.

[14] L. L. Rakêt, S. Sommer, and B. Markussen. A nonlinear mixed-effects model for simultaneous smoothing and registration of functional data. *Pattern Recognition Letters*, 38:1–7, 2014.

[15] G. K. Robinson. That BLUP is a good thing: The estimation of random effects. *Statistical Science*, 6(1):15–32, 1991.

[16] P. Billingsley. *Convergence of probability measures*, volume 493 of *Wiley Series in Probability and Statistics*. John Wiley & Sons, 1999.

[17] M. L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, 1999.

[18] A. Arribas-Gil and J. Romo. Robust depth-based estimation in the time warping model. *Biostatistics*, 13(3):398–414, 2012.

[19] C. Dimeglio, S. Gallón, J.-M. Loubes, and E. Maza. A robust algorithm for template curve estimation based on manifold embedding. *Computational Statistics & Data Analysis*, 70:373–386, 2014.

[20] S. A. Kurtek, A. Srivastava, and W. Wu. Signal estimation under random time-warpings and nonlinear signal alignment. In *Advances in Neural Information Processing Systems*, pages 675–683, 2011.

[21] J. D. Tucker. *fdasrvf: Elastic Functional Data Analysis*, 2014. R package version 1.4.2.

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

[23] A. Cuevas, M. Febrero, and R. Fraiman. An anova test for functional data. *Computational Statistics & Data Analysis*, 47(1):111–122, 2004.

[24] A. Antoniadis and T. Sapatinas. Estimation and inference in functional mixed-effects models. *Computational Statistics & Data Analysis*, 51(10):4793–4813, 2007.

Chapter 4

Conclusion

4.1 Contributions

The work presented in this thesis marks the two first steps toward a framework for analyzing functional objects based on stochastic processes. In Chapter 2 we presented an approximation method that allowed efficient likelihood analyses of functional data on high-dimensional Euclidean spaces, and in Chapter 3 we presented a model that included both phase and amplitude variation in functional data, thus accounting for the two most natural types of variation.

The principles behind the presented work has been that one should avoid oversimplification and preprocessing of the data at hand, but instead try to build statistical models that account for relevant data effects.

4.2 Future work

Statistical analysis of functional objects is a young field with many open problems. While a general likelihood-based framework that seamlessly handle the more exotic functional objects from Chapter 1 still seem out of reach, there are many simpler problems that can be addressed.

In addition to development of new methodology for data of different complexities, there is also a need to study and compare existing methodology in the functional setting, and to develop methodology to answer new types of questions.

The following four subsections describe extensions to the work described here, that we are currently considering.

4.2.1 Local significance

For many types of experiments, the question of interest is not *whether* there is an effect, but *where* the effect of e.g. treatment or disease can be found in the data (this question is for example relevant for the Glyphosate data in Paper P.1). While functional regression, or local per-observation regression on the distinguishing variable may provide ad hoc answers to this question, a more reliable answer would consider the question in connection to the data model. A possible solution is to answer the question using p-values obtained from classical statistical testing. This approach may however be problematic from the point of view of current modeling paradigms in functional data analysis. Firstly, the standard asymptotic basis of statistical testing fails for immense numbers of parameters, and gives rise to a multiple testing problem. Secondly, for models formulated in terms of discrete bases, the individual basis functions—which may have non-obvious interactions—typically do no represent stand-alone effects in the functional signal, and thus significance testing of local effects in terms of basis functions will generally give skewed results. To address the multiple testing problem, alternative extremity-based measures of significance have been developed in bioinformatics (Altschul et al. 1997). In the functional setting, such constructions does not seem to have been considered. Compared to the methods of Altschul et al. (1997), the functional alternative require a number of adaptions due to the underlying geometric structure of the data.

In a future work we will consider extremity based measures for assessing local signficance. This will be done by considering linear functional mixed-effects models in a scale space representation (Lindeberg 1993), where the estimates are analyzed after different amounts of spatial smoothing. By choosing a pointwise representation of the fixed effect θ, one can calculate the distribution of the estimated effects on any scale. For a given scale, one can use results from stochastic geometry on extrema of random fields (Leadbetter et al. 1982), to locally measure the extremity (e.g. in terms of the expected number of observations taking more extreme values) of the given estimate under the null hypothesis of no local effect. By computing such measures across all scales, one can do multi-scale segmentation (Olsen & Nielsen 1997) of extremity, to produce a robust measure of the locally most extreme, and thus significant effects.

4.2.2 Non-Gaussian processes

In the literature on functional mixed-effects models, Gaussian processes seem to be the universal choice for modelling functional random effects. In practice, however, data may have structure that cannot be properly modeled in a Gaussian setting.

The approximations in Paper P.1 relied heavily on the fact that noise and random effects were Gaussian. However, one can consider distributions that exhibit similar structure. The class of elliptically contoured distributions (Cambanis et al. 1981) have

probability densities of the form

$$p(\boldsymbol{y}) = \frac{1}{\sqrt{\det \Sigma}}\, g((\boldsymbol{y}-\boldsymbol{\theta})^\top \Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\theta})).$$

In this class of models, the multivariate generalized Laplace distribution (Kotz et al. 2001) seems particularly interesting for functional data analysis because of its robustness and its sparsity-inducing properties. In order to derive operator approximations similar to the ones in Paper P.1, we need covariances of the form $\Sigma = \Sigma_0 + \mathbb{I}$ where Σ_0 is a functional covariance matrix. This covariance structure in connection with general elliptically contoured distributions implies that we typically cannot construct models with independent observation noise, but that there will be a layer of uncorrelated (but not independent) noise in the random effect.

4.2.3 Models for shape data

As already mentioned, one major tasks ahead is to define mixed-effect models based on random fields, and calculate the resulting likelihood functions in the setting of more complex functional objects. We are currently considering model (1.5) for data on the spherical domains $\mathbb{S}^1$ and $\mathbb{S}^2$. This extension will give a natural model for shape analysis, which may provide likelihood based answers to some of the fundamental questions in shape analysis. In particular, point correspondences to the observed shapes can be automatically estimated, and the estimated variance of the warping function will give an indication of the extent to which point correspondences are present in the data at hand.

4.2.4 Constrained stochastic processes

The warping functions in Paper P.2 were modeled as coarsely observed Brownian motions or bridges. Sample paths of these processes are nowhere monotone, nor are the predicted processes necessarily monotone. As a result, we had to restrict the solution space to homeomorphisms in Paper P.3. As previously mentioned, one approach of guaranteeing diffeomorphic processes is to define them as solutions to certain stochastic transport equations (Markussen 2007). This is however mathematically complex, and from a modeling point of view, a class of monotone processes with known distribution would be very convenient. We are currently considering this issue by defining classes of diffeomorphic warping processes using a conditional hierarchical structure.

Appendix A

Data registration with L^1 data terms

1.1 Introduction

This appendix describes a framework for registering functional data using L^1-norm data terms. The content of this appendix is adapted from Rakêt et al. (2011) and Rakêt (2013), and focusses on registration of the Glyphosate data described in Paper P.1. For an extensive review of applications and extensions we refer to Rakêt (2013).

Let two functional samples $I_0, I_1 : \mathbb{R}^d \to \mathbb{R}^q$ be given. In the following we will think of I_0 and I_1 as planar images, i.e. $d = 2$, but the presented methods are generally applicable for any d. The problem we are considering is to compute a map v, such that the difference between I_1 warped with v and I_0 is close to zero

$$I_1(\boldsymbol{x} + v(\boldsymbol{x})) - I_0(\boldsymbol{x}) \approx 0. \tag{A.1}$$

Solving (A.1) equal to zero is problematic in a number of ways. It is generally ill-posed; images often have one-dimensional values, and for each point $\boldsymbol{x}$ we need to estimate a two-dimensional displacement vector $v(\boldsymbol{x})$. In addition the problem is nonlinear. To address these issues, a common approach is to linearize equation (A.1) by means of its first order Taylor approximation in v. So-called *local* methods assume that the displacement $v(\boldsymbol{x})$ is similar in a neighborhood of $\boldsymbol{x}$, which typically gives enough linear independent equations in the channels of v for proper estimation (Lucas & Kanade 1981). In contrast *global* methods typically use a pointwise data term based on the linearization of (A.1), but adds a regularization term, that penalizes erratic behavior of v, giving an energy that must be minimized in order to estimate v (Horn & Schunck 1981). Here we will focus on global methods in a setup inspired by the so-called duality-based optical flow method (Zach et al. 2007).

1.2 Energy formulation and estimation

Given a subdomain $\mathscr{T} \subseteq \mathbb{R}^d$ containing all data values and two images $I_0, I_1 : \mathscr{T} \to \mathbb{R}^q$, we want to estimate the warping function $v : \mathscr{T} \to \mathscr{T}$ that aligns I_1 to I_0. We will consider a variational approach where the flow v is estimated as a minimizer of an energy on the form

$$E(v) = \lambda \int_{\mathscr{T}} \|I_1(\boldsymbol{x} + v(\boldsymbol{x})) - I_0(\boldsymbol{x})\| \, \mathrm{d}\boldsymbol{x} + G(v) \tag{A.2}$$

where G acts as a regularization term on the warp.

Here we will focus on a half-quadratic relaxation of the problem, and consider the minimization methods in this framework. The relaxed energy is obtained by introducing an auxiliary variable, effectively splitting the minimization problem in two quadratically coupled problems

$$E(u, v) = \lambda \int_{\mathscr{T}} \|I_1(\boldsymbol{x} + v(\boldsymbol{x})) - I_0(\boldsymbol{x})\| \, \mathrm{d}\boldsymbol{x} + \frac{1}{2\theta} \int_{\mathscr{T}} \|v(\boldsymbol{x}) - u(\boldsymbol{x})\|^2 \, \mathrm{d}\boldsymbol{x} + G(u). \tag{A.3}$$

As $\theta \to 0$ a minimizer of (A.3) will also minimize (A.2). The hope of this relaxation is that for θ small, a minimizer of the relaxed energy (A.3) will be close to a minimizer of the original energy (A.2).

It may seem troublesome to introduce an auxiliary variable, since one has to iteratively minimize the two energies

$$E_1(v) = \lambda \int_{\mathscr{T}} \|I_1(\boldsymbol{x} + v(\boldsymbol{x})) - I_0(\boldsymbol{x})\| \, \mathrm{d}\boldsymbol{x} + \frac{1}{2\theta} \int_{\mathscr{T}} \|v(\boldsymbol{x}) - u(\boldsymbol{x})\|^2 \, \mathrm{d}\boldsymbol{x}, \tag{A.4}$$

$$E_2(u) = \frac{1}{2\theta} \int_{\mathscr{T}} \|v(\boldsymbol{x}) - u(\boldsymbol{x})\|^2 \, \mathrm{d}\boldsymbol{x} + G(u), \tag{A.5}$$

instead of just a single one. The splitting method, however, has the advantage that the two subproblems (A.4) and (A.5) are often much easier to solve than the original energy (A.2). This is because E_1 can be solved pointwise, and for a wide variety of regularization terms G, E_2 takes the form of a standard denoising problem for which efficient solvers are available (Scherzer et al. 2008). Another positive feature is that data-matching and regularization are done independently, so one can easily replace one without changing the minimization of the other—a fact that makes comparison of different types of energies straightforward.

In the following section we will consider the minimization of the energy (A.4). In order to minimize this energy, we iteratively linearize the data term by its first order Taylor approximation around the current estimates $\boldsymbol{v}^0_{\boldsymbol{x}}$

$$I_1(\boldsymbol{x} + v(\boldsymbol{x})) - I_0(\boldsymbol{x}) \approx I_1(\boldsymbol{x} + \boldsymbol{v}^0_{\boldsymbol{x}}) + J_{I_1}(\boldsymbol{x} + \boldsymbol{v}^0_{\boldsymbol{x}})(v(\boldsymbol{x}) - \boldsymbol{v}^0_{\boldsymbol{x}}) - I_0(\boldsymbol{x})$$

where $J_{I_1}(\boldsymbol{x})$ is the Jacobian at $I_1(\boldsymbol{x})$. This procedure is known as *warping*, and has been theoretically justified by Brox et al. (2004), who show that the warping method corresponds to minimizing an energy with non-linearized data terms using two nested fixed-point iterations. To recover large deformations and accelerate the solution scheme we do the minimization in a coarse-to-fine framework. While it is possible to directly target the original energy (A.4), the current methods to do so are typically slower and produce inferior results (Steinbrücker et al. 2009).

1.2.1 Minimizing affine L^1-L^2 energies

Consider an L^1-L^2 energy on the form

$$E_1(v) = \lambda \int_{\mathscr{I}} \|A_{\boldsymbol{x}} v(\boldsymbol{x}) + b(\boldsymbol{x})\| \, \mathrm{d}x + \frac{1}{2} \int_{\mathscr{I}} \|v(\boldsymbol{x}) - u(\boldsymbol{x})\|^2 \, \mathrm{d}\boldsymbol{x} \tag{A.6}$$

where $A_{\boldsymbol{x}} : \mathbb{R}^d \to \mathbb{R}^q$. The minimization of (A.6) boils down to pointwise minimization of a strictly convex cost function of the form

$$f(\boldsymbol{v}) = \lambda \|A\boldsymbol{v} + \boldsymbol{b}\| + \frac{1}{2}\|\boldsymbol{v} - \boldsymbol{u}\|^2. \tag{A.7}$$

In the following we present the tools used for solving the minimization problem (A.7). We start by recalling some elements of convex analysis, the reader can refer to Ekeland & Teman (1999) for a complete introduction to convex analysis in both finite and infinite dimension. Here we will restrict ourselves to finite dimensional problems.

A function $f : \mathbb{R}^d \to \mathbb{R}$ is one-homogeneous if $f(\lambda \boldsymbol{x}) = \lambda f(\boldsymbol{x})$, for all $\lambda > 0$. For a one-homogeneous function, it is easily shown that its Legendre-Fenchel transform

$$f^*(\boldsymbol{x}^*) = \sup_{\boldsymbol{x} \in \mathbb{R}^d} \{\langle \boldsymbol{x}, \boldsymbol{x}^* \rangle - f(\boldsymbol{x})\} \tag{A.8}$$

is the characteristic function of a closed convex set $\mathfrak{C}$ of $\mathbb{R}^d$,

$$d_{\mathfrak{C}}(\boldsymbol{x}^*) := f^*(\boldsymbol{x}^*) = \begin{cases} 0 & \text{if } \boldsymbol{x}^* \in \mathfrak{C}, \\ +\infty & \text{otherwise.} \end{cases} \tag{A.9}$$

The one-homogeneous functions that will interest us here are of the form $f(\boldsymbol{x}) = \|A\boldsymbol{x}\|$ where $A : \mathbb{R}^d \to \mathbb{R}^q$ is linear, and $\|\cdot\|$ is the usual Euclidean norm of $\mathbb{R}^q$.

The following lemma characterizes the Legendre-Fenchel transform of $f(\boldsymbol{x})$.

Lemma A.1. The Legendre-Fenchel tranform of $\boldsymbol{x} \mapsto \|A\boldsymbol{x}\|$ is the characteristic function $d_{\mathfrak{C}}$ of the elliptic ball $\mathfrak{C}$ given by the set of $\boldsymbol{x}$'s in $\mathbb{R}^d$ that satisfy the following conditions

$$A^\dagger A\boldsymbol{x} = \boldsymbol{x} \tag{A.10}$$

$$\boldsymbol{x}^\top A^\dagger A^{\dagger\top} \boldsymbol{x} \leq 1, \tag{A.11}$$

where $A^\dagger$ denotes the Moore-Penrose pseudoinverse of A.

It is well known that $A^\dagger A$ is the orthogonal projection onto $\operatorname{Ker} A^\perp$, so the equality $\boldsymbol{x} = A^\dagger A\boldsymbol{x}$ means that $\boldsymbol{x}$ belongs to $\operatorname{Ker} A^\perp$. On this subspace, $A^\dagger A^{\dagger\top}$ is positive definite and the inequality thus defines an elliptic ball.

The lemma will not be proven here, but we indicate how it can be done. In the case where A is the identity $\mathbb{I}_d$ of $\mathbb{R}^d$, it is easily shown that $\mathfrak{C}$ is the unit sphere of $\mathbb{R}^d$. The case where A is invertible follows easily, while the general case follows from the latter using the structure of pseudoinverse (see Golub & van Loan 1989 for instance).

The following proposition gives the minimizer of the energy (A.6).

Proposition A.1. The minimizer of the function

$$f(\boldsymbol{v}) = \lambda\|A\boldsymbol{v} + \boldsymbol{b}\| + \frac{1}{2}\|\boldsymbol{v} - \boldsymbol{u}\|^2$$

is given as follows.

(i) In the case $\boldsymbol{b} \notin \operatorname{Im} A$, f is smooth. It can be minimized by usual methods.

(ii) In the case $\boldsymbol{b} \in \operatorname{Im} A$, f, which is not smooth for $\boldsymbol{v} \in \operatorname{Ker} A + A^\dagger b$, reaches its unique minimum at

$$\boldsymbol{v} = \boldsymbol{u} - \pi_{\lambda\mathfrak{C}}\left(\boldsymbol{u} + A^\dagger\boldsymbol{b}\right) \tag{A.12}$$

where $\pi_{\lambda\mathfrak{C}}$ is the projection onto the convex set $\lambda\mathfrak{C} = \{\lambda\boldsymbol{x},\, \boldsymbol{x} \in \mathfrak{C}\}$, with $\mathfrak{C}$ as described in Lemma A.1.

Proof. To see (i), write $\boldsymbol{b}$ as $A\boldsymbol{b}_0 + \boldsymbol{b}_1$, with $\boldsymbol{b}_0 = A^\dagger\boldsymbol{b}$, $A\boldsymbol{b}_0$ being the orthogonal projection of $\boldsymbol{b}$ onto $\operatorname{Im} A$, while $\boldsymbol{b}_1$ is the residual of the projection. The assumption of (i) implies that $\boldsymbol{b}_1 \neq \boldsymbol{0}$ is orthogonal to the image of A. One can then write

$$\|A\boldsymbol{v} + \boldsymbol{b}\| = \|A(\boldsymbol{v} + \boldsymbol{b}_0) + \boldsymbol{b}_1\| = \sqrt{\|A(\boldsymbol{v} + \boldsymbol{b}_0)\|^2 + \|\boldsymbol{b}_1\|^2} \tag{A.13}$$

which is always strictly positive as $\|\boldsymbol{b}_1\|^2 > 0$, and smoothness follows.

In the case of (ii), since $\boldsymbol{b} \in \operatorname{Im} A$, we can substitute $\boldsymbol{v} \leftarrow \boldsymbol{v} + A^\dagger\boldsymbol{b}$ in function (A.7) and the resulting function has the same form as a number of functions found in Chambolle (2004) and Chambolle & Pock (2011). We refer to these works for the computation of minimizers in terms of the Legendre-Fenchel transform. □

Example A.1. Consider the minimization problem

$$\arg\min_{\boldsymbol{v}} \left(\lambda\|A\boldsymbol{v} + \boldsymbol{b}\| + \frac{1}{2}\|\boldsymbol{v} - \boldsymbol{u}\|^2\right), \qquad \lambda > 0. \tag{A.14}$$

where $A \in \mathbb{R}^{q\times 2}$ and $\boldsymbol{b} \in \operatorname{Im} A$. If A has maximal rank (i.e. 2), then is is well known that the 2×2 matrix $C = A^\dagger A^{\dagger\top}$ is symmetric and positive definite (Golub & van Loan

1989). The set $\mathfrak{C}$ is then an elliptic disc determined by the eigenvectors and eigenvalues of C. The projection can be computed by the efficient algorithm described in Example A.3, which has much better properties than the method originally suggested in Rakêt et al. (2011).

When the matrix has two linearly dependent columns $\boldsymbol{a} \neq \mathbf{0}$ and $c\boldsymbol{a}$, a series of straightforward calculations give

$$\operatorname{Ker} A = \mathbb{R}\boldsymbol{y}, \quad \operatorname{Ker} A^{\perp} = \mathbb{R}\boldsymbol{x}, \quad \operatorname{Im} A = \mathbb{R}\boldsymbol{a} \tag{A.15}$$

with $\boldsymbol{x} = \frac{1}{1+c^2}(1, c)^\top$ and $\boldsymbol{y} = \frac{1}{1+c^2}(-c, 1)^\top$, and

$$A^\dagger A^{\dagger\top} = \frac{1}{(1+c^2)^2\|\boldsymbol{a}\|^2}\begin{pmatrix} 1 & c \\ c & c^2 \end{pmatrix}. \tag{A.16}$$

If $c = 0$, the inequality (A.11) from Lemma A.1, just amounts to

$$-\|\boldsymbol{a}\| \leq u_1 \leq \|\boldsymbol{a}\|, \qquad \boldsymbol{u} = (u_1, u_2)^\top, \tag{A.17}$$

while equality (A.10) in Lemma A.1 simply says that $u_2 = 0$, thus $\mathfrak{C}$ is the line segment

$$[-\|\boldsymbol{a}\|\boldsymbol{x}, \|\boldsymbol{a}\|\boldsymbol{x}] \subset \mathbb{R}^2. \tag{A.18}$$

The case where $c \neq 0$ is identical, and obtained for instance by rotating the natural basis of $\mathbb{R}^2$ to the basis $(\boldsymbol{x}, \boldsymbol{y})$. ○

Example A.2. Consider again the minimization problem (A.14), but this time assuming that $\boldsymbol{b} \notin \operatorname{Im} A$. Using (A.13) we can rewrite the minimization problem as

$$\arg\min_{\boldsymbol{v}} \left(\lambda\sqrt{\|A(\boldsymbol{v} + \boldsymbol{b}_0)\|^2 + \|\boldsymbol{b}_1\|^2} + \frac{1}{2}\|\boldsymbol{v} - \boldsymbol{u}\|^2 \right), \qquad \lambda > 0. \tag{A.19}$$

The minimizing $\boldsymbol{v}$ is found by solving the equation

$$\lambda\frac{A^\top A(\boldsymbol{v} + \boldsymbol{b}_0)}{\|A\boldsymbol{v} + \boldsymbol{b}\|} + \boldsymbol{v} - \boldsymbol{u} = 0$$

which may be done by gradient descent or a (quasi-)Newton method. ○

Example A.3. Consider the problem of projecting a point $\boldsymbol{x}_0$ onto the ellipsoid given by

$$\mathfrak{C} = \{\boldsymbol{x} \in \mathbb{R}^q \,|\, \boldsymbol{x}^\top C\boldsymbol{x} \leq 1\},$$

where $\boldsymbol{x}_0 \notin \mathfrak{C}$. The projected point $\hat{\boldsymbol{x}}$ can be found as

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}\in\mathfrak{C}} \|\boldsymbol{x} - \boldsymbol{x}_0\|^2.$$

This problem can be reformulated by introducing a Lagrange multiplier ξ, giving the objective function

$$f(\boldsymbol{x}, \lambda) = \|\boldsymbol{x} - \boldsymbol{x}_0\|^2 + \xi(\boldsymbol{x}^\top C\boldsymbol{x} - 1).$$

From the condition that

$$\frac{\partial}{\partial \boldsymbol{x}} f(\boldsymbol{x}, \xi) = 2(\boldsymbol{x} - \boldsymbol{x}_0) + 2\xi C\boldsymbol{x} = \boldsymbol{0},$$

we get that

$$\hat{\boldsymbol{x}} = (\xi C + \mathbb{I})^{-1}\boldsymbol{x}_0.$$

However we need to determine the value of the Lagrange multiplier ξ. Since we assumed that $\boldsymbol{x}_0$ was outside the ellipsoid, we know that the projected point will lie on the boundary of the ellipsoid $\partial\mathfrak{C}$, which means that ξ is a root of

$$G(\xi) = ((\xi C + \mathbb{I})^{-1}\boldsymbol{x}_0)^\top C(\xi C + \mathbb{I})^{-1}\boldsymbol{x}_0 - 1. \tag{A.20}$$

We can use the following theorem due to Kiseliov (1994) to determine the correct value of ξ.

Theorem A.1. The root ξ^* of the function (A.20) is unique and can be found by the iterative Newton process

$$\xi_0 = 0, \qquad \xi_{n+1} = \xi_n - \frac{G(\xi_n)}{G'(\xi_n)},$$

where $\xi_k \uparrow \xi^*$. The rate of convergence is quadratic.

Kiseliov (1994) in addition gives a nonlinear version of the Newton process described in the above theorem, which is even more efficient. Compared to the added complexity of the implementation, the overall gain of using such an algorithm is limited, and we will recommend the process described here. ∘

1.3 Algorithm

In this section we describe a general algorithmic framework for estimating the warping function from energies on the form (A.3). The duality-based approach has good computational properties, because the solutions to the two sub-energies can often be done in parallel. This makes the algorithm well-suited for massively parallel processors.

The structure of the algorithm is depicted in Algorithm A.1. The different choices made in the algorithm build on a foundation of practices that has been shown to improve accuracy in optical flow estimation (Sun et al. 2010). The steps of the algorithm are described below.

Data: Two images I_0 and I_1
Result: The warping function v
for $\ell = \ell_{\text{max}}$ *to* 0 **do**
    `// Pyramid levels`
    Downsample the images I_0 and I_1 to current pyramid level
    for $w = 0$ *to* w_{max} **do**
        `// Warping`
        Compute v as the minimizer of E_1 (A.4)
        for $i = 0$ *to* i_{max} **do**
            `// Inner iterations`
            Compute u as the minimizer of E_2 (A.5)
        end
        Upscale v and u to next pyramid level
    end
end

Algorithm A.1: Computation of warping function.

Pyramid An image pyramid is built, where on each level, prior to downsampling to the next pyramid level, the images are smoothed with a Gaussian kernel of standard deviation σ. The downsampling is done by means of linear interpolation. Evaluation at non-pixel positions in images is done by bicubic interpolation.

Warping At the beginning of each warp, the image I_1 is warped according to the current estimate of the flow v.

Upscaling Flows are upscaled using linear interpolation, and their values are divided by the downscale factor of the pyramid, in order for vector lengths to match the current image size. This is followed by an application of a 3×3 median filter which slightly increase convergence (Sun et al. 2010).

1.4 Registration of 2D chromatograms

Chromatography is a process for separating mixtures. One use of chromatography is measuring relative proportions of analytes in a number of mixtures, to determine differences. An example of a 2D chromatogram is shown in Figure A.1. The chromatograms we are considering have been generated using ultra-high-performing liquid chromatography with diode-array-detection (Petersen et al. 2011). The chromatograms consists of 209 wavelengths each measured at 24,000 retention times. The subject of the analysis is rapeseed seedlings having been exposed to different levels glyphosate (commonly known

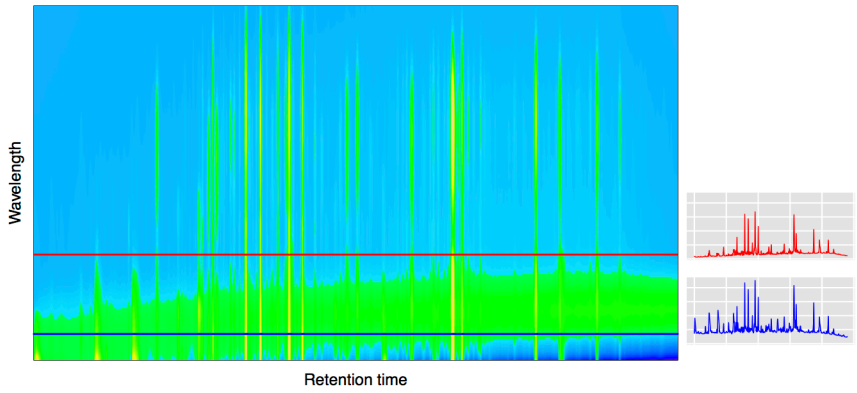


Figure A.1: Example of a chromatogram along with the absorbance (A.U.) curves corresponding to two fixed wavelengths.

as Roundup®).

The images arising from this procedure will have shifts in retention time, but because of the experimental setup, no such shifts occur in the wavelength dimension. This means that we have a one-dimensional registration problem for a two-dimensional image.

Figure A.2 depicts the absorbance as a function of retention time for a single wavelength in four chromatograms. The retention time shifts are clearly visible. Furthermore there seem to be a varying detector sensitivity, resulting in some of the curves consistently having higher peaks that others. Finally there are small variations that cannot be explained by the mentioned issues, and which can be ascribed to serially correlated effects and noise.

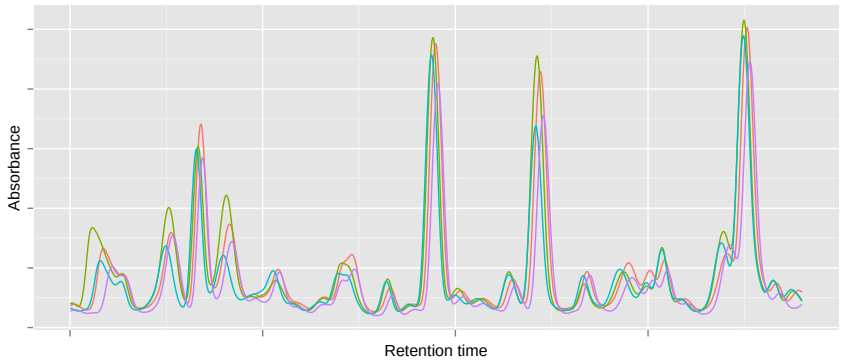


Figure A.2: A range of retention times for a single wavelength for four chromatograms.

1.4.1 Registration algorithm

Given two chromatograms $I_0, I_1 : \mathscr{T} \to \mathbb{R}$ of size $n_1 \times n_2$, where $\mathscr{T} = \mathscr{T}_w \times \mathscr{T}_t$, consider the problem of estimating the disparity $v : \mathscr{T}_t \to \mathscr{T}_t$ such that $I_1(w, t + v(t))$ is properly registered to $I_0(w, t)$.

From the point of view of Section A the natural formulation of the data term is as a vector valued problem. Let

$$\boldsymbol{I}_i(t) = \begin{pmatrix} I_i(w_1, t) \\ \vdots \\ I_i(w_{n_1}, t) \end{pmatrix}.$$

The warping constraint may then be written as

$$\boldsymbol{I}_1(t + v(t)) - \boldsymbol{I}_0(t) \approx \mathbf{0}.$$

Linearizing this around a given estimate v_0, we get the following system of equations

$$\underbrace{\partial_t \boldsymbol{I}_1(t + v_0)}_{\boldsymbol{a}} v(t) \underbrace{-\partial_t \boldsymbol{I}_1(t + v_0)v_0 + \boldsymbol{I}_1(t + v_0) - \boldsymbol{I}_0(t)}_{\boldsymbol{b}} = \mathbf{0}.$$

Considering an L^1-norm of this linearization of this data term, we see that the case (ii) of Proposition A.1 is very easily calculated, however, it seems unlikely that it will ever be the case that $\boldsymbol{b} \in \operatorname{Im} \boldsymbol{a}$ for just a moderate number of wavelengths. This means that we will almost surely be in the less attractive case (i) where we have to minimize by some iterative procedure.

Instead we will consider an alternative registration method for this dataset. The idea is to treat the one-dimensional vector valued registration problem as a two-dimensional problem, and couple the different vector channels through the regularization rather than through the data term. The method is generally applicable, and works by posing a d dimensional registration problem with data taking values in a q dimensional space, as a one-dimensional registration problem on a $d + 1$ dimensional domain. This is done by treating the vector channels as an added dimension to the domain. This way the regularization will be $d + 1$ dimensional, and by enforcing strong (or increasing) weight on the regularity across this new dimension, information is propagated between the different channels of the image to produce a warping function that is homogeneous along the new dimension.

As described above, we start out by estimating disparities for each wavelength. In the given example we are interested in a robust L^1-norm for the data term. The robustness is important because of the varying detector sensitivity and serially correlated effects, where for example an L^2-norm may cause problems in relation to outliers. For regularization, we are interested in a term that, in addition to imposing regularity on the estimated

disparities, regularize across wavelengths. Since one would expect drifts in retention time to be continuous, the warp should be smooth, and therefore we will regularize using the squared gradient magnitude. The energy to be minimized looks as follows

$$E(v) = \lambda \int_{\mathscr{T}} \|I_1(w, t + v_w(t)) - I_0(w, t)\| \,\mathrm{d}w \,\mathrm{d}t + \int_{\mathscr{T}} \|\nabla_{w,t} v_w(t)\|^2 \,\mathrm{d}t.$$

This functional is minimized as described in the previous sections, where the data term is iteratively approximated by its first-order Taylor approximation around the given estimate v_0^w

$$\rho(v)(w, t) = \partial_t I_1(w, t + v_0^w(t))(v(t) - v_0^w(t)) + I_1(w, t + v_0^w(t)) - I_0(w, t).$$

Furthermore data fidelity and regularization are decoupled by means of a quadratic proximity term

$$\begin{aligned} E(\boldsymbol{v}, \boldsymbol{u}') = \lambda \int_{\mathscr{T}} \|\rho(v_w)(w, t)\| \,\mathrm{d}w \,\mathrm{d}t \\ + \frac{1}{2\theta} \int_{\mathscr{T}} \|v_w(t) - u_w(t)\|^2 \,\mathrm{d}t + \int_{\mathscr{T}} \|\nabla_{w,t} u_w(t)\|^2 \,\mathrm{d}t \end{aligned}$$

where θ is sufficiently small. Using Proposition A.1, the pointwise solution in v_w is found to be

$$v_w(t) = u_w(t) - \lambda\theta \begin{cases} -\partial_t I_1(w, t + v_0^w(t)) & \text{if } \frac{\rho(v)(w,t)}{\lambda\theta} < -|\partial_t I_1(w, t + v_0^w(t))|^2 \\ \partial_t I_1(w, t + v_0^w(t)) & \text{if } \frac{\rho(v)(w,t)}{\lambda\theta} > |\partial_t I_1(w, t + v_0^w(t))|^2 \\ \frac{\rho(v)(w,t)}{\partial_t I_1(w, t + v_0^w(t))} & \text{if } \frac{|\rho(v)(w,t)|}{\lambda\theta} \leq |\partial_t I_1(w, t + v_0^w(t))|^2 \end{cases}.$$

The problem in u_w is a Tikhonov regularization problem that can be solved using standard methods. The weighting of the finite difference approximation of the derivative in the wavelength dimension is used to control the homogeneity of the solution across wavelengths. In the presented setup we found that a weighting factor of 10 produced good results. E is minimized iteratively in a coarse-to-fine manner, where the input images and the corresponding disparities are gradually upsampled in the retention time dimension, but the wavelength dimension is kept at its original size. Following Algorithm A.1 we use $\ell_{\text{max}} = 160$ pyramid levels and a scaling factor between levels of 0.97, yielding a downsampling factor at the coarsest level of approximately 130. The standard deviation of the Gaussian kernel used for smoothing was set to $\sigma = \frac{\sqrt{2}}{4}$. $w_{\text{max}} = 100$ warps are performed at each level using $i_{\text{max}} = 20$ inner iterations, and λ was set to 60, while θ was fixed at 0.1.

The algorithm has been implemented in CUDA C in order to take advantage of the thousands of cores on modern GPUs.

Figure A.3 shows the individual wavelength warping curves (gray) of the described method with the average plotted on top (red) for two 2D chromatograms. To illustrate the effect of the weighting of the finite differences in the wavelength dimension the result of a weighting factor of 0 (i.e. registering each wavelength independently) and 10 are shown. As can be seen, the higher weighting factor results in successful propagation of information across wavelengths, producing a uniform result along the wavelength dimension. The independent registration produces much more variable results, which results in a smoother average registration curve with fewer details.

Figure A.3: Warping functions v_w for all individual wave lengths w (gray) with the average registration plotted on top (red).

The average registration is used as the final single disparity $v : t \to t$. The registration was then done by warping the chromatograms according to v for each wavelength. The result of the registration procedure on the data in Figure A.2 can be found in Figure A.4. This plot suggests that data is very well aligned after registration.

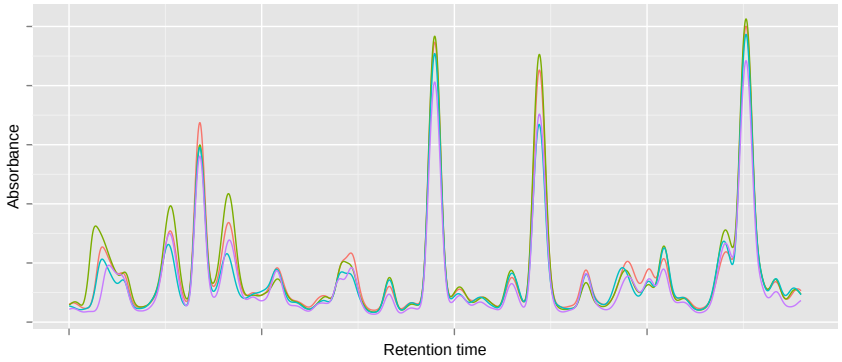


Figure A.4: The chromatograms from Figure A.2 registrated along retention time.

Bibliography

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997), 'Gapped BLAST and PSI-BLAST: a new generation of protein database search programs', *Nucleic Acids Research* **25**(17), 3389–3402.

Antoniadis, A. & Sapatinas, T. (2007), 'Estimation and inference in functional mixed-effects models', *Computational statistics & data analysis* **51**(10), 4793–4813.

Bolin, D. & Lindgren, F. (2011), 'Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping', *The Annals of Applied Statistics* pp. 523–550.

Boularan, J., Ferré, L. & Vieu, P. (1994), 'Growth curves: a two-stage nonparametric approach', *Journal of Statistical Planning and Inference* **38**(3), 327–350.

Brox, T., Bruhn, A., Papenberg, N. & Weickert, J. (2004), High accuracy optical flow estimation based on a theory for warping, *in* 'ECCV', Vol. 4, pp. 25–36.

Cambanis, S., Huang, S. & Simons, G. (1981), 'On the theory of elliptically contoured distributions', *Journal of Multivariate Analysis* **11**(3), 368–385.

Chambolle, A. (2004), 'An algorithm for total variation minimization and applications', *Journal of Mathematical Imaging and Vision* **20**, 89–97.

Chambolle, A. & Pock, T. (2011), 'A first-order primal-dual algorithm for convex problems with applications to imaging', *Journal of Mathematical Imaging and Vision* **40**, 120–145.

Charon, N. (2013), Analysis of geometric and functional shapes with extensions of currents. Application to registration and atlas estimation., PhD thesis, Ecole normale supérieure de Cachan.

Chen, H. & Wang, Y. (2011), 'A penalized spline approach to functional mixed effects model analysis', *Biometrics* **67**(3), 861–870.

Cox, D. D. (1984), 'Multivariate smoothing spline functions', *SIAM Journal on Numerical Analysis* **21**(4), 789–813.

Cuevas, A., Febrero, M. & Fraiman, R. (2004), 'An anova test for functional data', *Computational statistics & data analysis* **47**(1), 111–122.

Darkner, S., Larsen, R. & Paulsen, R. (2007), Analysis of deformation of the human ear and canal caused by mandibular movement, *in* N. Ayache, S. Ourselin & A. Maeder, eds, 'Medical Image Computing and Computer-Assisted Intervention – MICCAI 2007', Vol. 4792 of *Lecture Notes in Computer Science*, Springer, pp. 801–808.

Demetz, O., Hafner, D. & Weickert, J. (2013), The complete rank transform: A tool for accurate and morphologically invariant matching of structures, *in* 'Proc. 2013 British Machine Vision Conference, Bristol, UK'.

Dobler, G., Fassnacht, C., Treu, T., Marshall, P. J., Liao, K., Hojjati, A., Linder, E. & Rumbaugh, N. (2013), 'Strong lens time delay challenge: I. experimental design', *arXiv preprint arXiv:1310.4830* .

Dryden, I. L. & Mardia, K. V. (1998), *Statistical Shape Analysis*, John Wiley & Sons New York.

Duchamp, T. & Stuetzle, W. (2003), 'Spline smoothing on surfaces', *Journal of Computational and Graphical Statistics* **12**(2), 354–381.

Ekeland, I. & Teman, R. (1999), *Convex Analysis and Variational Problems*, SIAM.

Ferraty, F. & Vieu, P. (2006), *Nonparametric Functional Data Analysis*, Springer.

Fisher, N. I. (1993), *Statistical analysis of spherical data*, Cambridge University Press.

Glasbey, C. A. & Mardia, K. V. (1998), 'A review of image-warping methods', *Journal of applied statistics* **25**(2), 155–171.

Golub, G. & van Loan, C. (1989), *Matrix Computations*, The John Hopkins University Press, Baltimore, Maryland.

Grenander, U. & Miller, M. I. (1998), 'Computational anatomy: An emerging discipline', *Quarterly of applied mathematics* **56**(4), 617–694.

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

Hastie, T., Tibshirani, R. & Friedman, J. (2009), Basis expansions and regularization, *in* 'The Elements of Statistical Learning', Springer Series in Statistics, Springer New York, pp. 139–189.

Hauberg, S., Sommer, S. & Pedersen, K. S. (2012), 'Natural metrics and least-committed priors for articulated tracking', *Image and Vision Computing* **30**(6-7), 453–461.

Hobolth, A. & Jensen, E. B. V. (2000), 'Modelling stochastic changes in curve shape, with an application to cancer diagnostics', *Advances in Applied Probability* **32**(2), 344–362.

Horn, B. K. P. & Schunck, B. G. (1981), 'Determining optical flow', *Artificial Intelligence* **17**, 185–203.

Horváth, L. & Kokoszka, P. (2012), *Inference for Functional Data with Applications*, Springer.

Joshi, S. C. & Miller, M. I. (2000), 'Landmark matching via large deformation diffeomorphisms', *Image Processing, IEEE Transactions on* **9**(8), 1357–1370.

Kiseliov, Y. (1994), 'Algorithms of projection of a point onto an ellipsoid', *Lithuanian Mathematical Journal* **34**, 141–159.

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

Kotz, S., Kozubowski, T. & Podgorski, K. (2001), *The Laplace Distribution and Generalizations: A Revisit With Applications to Communications, Exonomics, Engineering, and Finance*, number 183 *in* 'Birkhäuser Applied Probability and Statistics', Springer.

Leadbetter, M. R., Lindgren, G. & Rootzén, H. (1982), *Extremes and Related Properties of Random Sequences and Processes*, Springer Berlin Heidelberg New York.

Lindeberg, T. (1993), *Scale-space theory in computer vision*, Springer.

Lindgren, F., Rue, H. & Lindström, J. (2011), 'An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**(4), 423–498.

Liu, Z. & Guo, W. (2012), 'Functional mixed effects models', *Wiley Interdisciplinary Reviews: Computational Statistics* **4**(6), 527–534.

Lucas, B. D. & Kanade, T. (1981), An iterative image registration technique with an application to stereo vision, *in* 'Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI '81)', pp. 674–679.

Luong, H. V., Raket, L. L. & Forchhammer, S. (2014), 'Re-estimation of motion and reconstruction for distributed video coding', *Image Processing, IEEE Transactions on* **23**(7), 2804–2819.

Markussen, B. (2007), 'Large deformation diffeomorphisms with application to optic flow', *Computer Vision and Image Understanding* **106**(1), 97 – 105.

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

Marron, J. S. & Alonso, A. M. (2014), 'Overview of object oriented data analysis', *Biometrical Journal*. (to appear).
URL: *http://dx.doi.org/10.1002/bimj.201300072*

McCall, C., Reddy, K. K. & Shah, M. (2012), Macro-class selection for hierarchical k-NN classification of inertial sensor data., *in* 'PECCS', pp. 106–114.

Nielsen, M., Johansen, P., Jackson, A. D., Lautrup, B. & Hauberg, S. (2008), 'Brownian warps for non-rigid registration', *Journal of Mathematical Imaging and Vision* **31**(2-3), 221–231.

Olsen, O. F. & Nielsen, M. (1997), Multi-scale gradient magnitude watershed segmentation, *in* A. Bimbo, ed., 'Image Analysis and Processing', Vol. 1310 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 6–13.

Olshen, R. A., Biden, E. N., Wyatt, M. P. & Sutherland, D. H. (1989), 'Gait analysis and the bootstrap', *The Annals of Statistics* **17**(4), 1419–1440.

Paragios, N., Chen, Y. & Faugeras, O. D. (2006), *Handbook of Mathematical Models in Computer Vision*, Springer.

Petersen, I. L., Tomasi, G., Sørensen, H., Boll, E. S., Hansen, H. C. B. & Christensen, J. H. (2011), 'The use of environmental metabolomics to determine glyphosate level of exposure in rapeseed (Brassica napus L.) seedlings', *Environmental Pollution* **159**(10), 3071 – 3077.

Pizer, S. M., Fletcher, P. T., Joshi, S., Gash, A. G., Stough, J., Thall, A., Tracton, G. & Chaney, E. L. (2005), 'A method and software for segmentation of anatomic object ensembles by deformable m-reps', *Medical Physics* **32**(5), 1335–1345.

Pizer, S. M., Jung, S., Goswami, D., Vicory, J., Zhao, X., Chaudhuri, R., Damon, J. N., Huckemann, S. & Marron, J. (2013), Nested sphere statistics of skeletal models, *in* 'Innovations for Shape Analysis', Springer, pp. 93–115.

Pottmann, H. & Hofer, M. (2005), 'A variational approach to spline curves on surfaces', *Computer Aided Geometric Design* **22**(7), 693–709.

Rakêt, L. L. (2013), 'Duality based optical flow algorithms with applications', University of Copenhagen prize thesis in Computer Science, Copenhagen University Library.

Raket, L. L., Grimme, B., Markussen, B., Schöner, G. & Igel, C. (2014), Statistical analysis of human arm movements using timing and motion separation, *in* 'Advances in Neural Information Processing Systems'. (submitted).

Rakêt, L. L. & Markussen, B. (2014), 'Approximate inference for spatial functional data on massively parallel processors', *Computational Statistics & Data Analysis* **72**, 227 – 240.

Rakêt, L. L., Roholm, L., Nielsen, M. & Lauze, F. (2011), TV-L^1 optical flow for vector valued images, *in* Y. Boykov, F. Kahl, V. Lempitsky & F. Schmidt, eds, 'Energy Minimization Methods in Computer Vision and Pattern Recognition', Vol. 6819 of *Lecture Notes in Computer Science*, Springer, pp. 329–343.

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

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

Sangalli, L. M., Ramsay, J. O. & Ramsay, T. O. (2013), 'Spatial spline regression models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**(4), 681–703.

Scherzer, O., Grasmair, M., Grossauer, H., Haltmeier, M. & Lenzen, F. (2008), *Variational Methods in Imaging*, Vol. 167, Springer.

Simpson, D., Lindgren, F. & Rue, H. (2012*a*), 'In order to make spatial statistics computationally feasible, we need to forget about the covariance function', *Environmetrics* **23**(1), 65–74.

Simpson, D., Lindgren, F. & Rue, H. (2012*b*), 'Think continuous: Markovian Gaussian models in spatial statistics', *Spatial Statistics* **1**, 16–29.

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

Srivastava, A., Wu, W., Kurtek, S., Klassen, E. & Marron, J. (2011), 'Registration of functional data using Fisher-Rao metric', *arXiv preprint arXiv:1103.3817* .

Steinbrücker, F., Pock, T. & Cremers, D. (2009), Large displacement optical flow computation without warping, *in* 'Computer Vision, 2009 IEEE 12th International Conference on', pp. 1609 –1614.

Sun, D., Roth, S. & Black, M. (2010), Secrets of optical flow estimation and their principles, *in* 'Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on', pp. 2432–2439.

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

Verzelen, N., Tao, W. & Müller, H.-G. (2012), 'Inferring stochastic dynamics from functional data', *Biometrika* **99**(3), 533–550.

Vialard, F.-X. (2013), 'Extension to infinite dimensions of a stochastic second-order model associated with shape splines', *Stochastic Processes and their Applications* **123**(6), 2110–2157.

Wahba, G. (1975), 'Smoothing noisy data with spline functions', *Numerische Mathematik* **24**(5), 383–393.

Wahba, G. (1981), 'Spline interpolation and smoothing on the sphere', *SIAM Journal on Scientific and Statistical Computing* **2**(1), 5–16.

Wahba, G. (1990), *Spline Models for Observational Data*, Society for Industrial and Applied Mathematics.

Wang, H. & Marron, J. (2007), 'Object oriented data analysis: Sets of trees', *The Annals of Statistics* **35**(5), 1849–1873.

Wang, Y. (1998), 'Mixed effects smoothing spline analysis of variance', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60**(1), 159–174.

Worsley, K. J. & Friston, K. J. (1995), 'Analysis of fMRI time-series revisited—again', *NeuroImage* **2**(3), 173–181.

Zach, C., Pock, T. & Bischof, H. (2007), A duality based approach for realtime TV-L^1 optical flow, *in* F. Hamprecht, C. Schnörr & B. Jähne, eds, 'Pattern Recognition', Vol. 4713 of *Lecture Notes in Computer Science*, Springer, pp. 214–223.