

# Non Parametric Multi Template Registration\*

Alois Kneip<sup>†,‡,??</sup>, and Heiko Wagner<sup>§,??,??</sup>

*Address of the First and Second authors*  
Usually a few lines long  
e-mail: [first@somewhere.com](mailto:first@somewhere.com)

[second@somewhere.com](mailto:second@somewhere.com)

*Address of the Third author*  
Usually a few lines long  
Usually a few lines long  
e-mail: [third@somewhere.com](mailto:third@somewhere.com)  
url: <http://www.foo.com>

**Abstract:** The abstract should summarize the contents of the paper. It should be clear, descriptive, self-explanatory and not longer than 200 words. It should also be suitable for publication in abstracting services. Please avoid using math formulas as much as possible.

This is a sample input file. Comparing it with the output it generates can show you how to produce a simple document of your own.

**MSC 2010 subject classifications:** Primary 60K35, 60K35; secondary 60K35.

**Keywords and phrases:** sample, L<sup>A</sup>T<sub>E</sub>X 2 $\varepsilon$ .

## 1. Introduction

The data that we consider are a sample of  $n$  smooth random functions  $x_1, \dots, x_n$  defined over a closed interval on the real line, such as the curves displayed in the top panel of Figure 1. The functions share a common set of shape features, consisting a peak and a valley. The sizes of the features vary, and we refer to this as *amplitude variation*. The locations of the features also vary from curve to curve, which indicates the existence of *phase variation*. Generally speaking, registration deals with separating amplitude and phase variation in a statistically meaningful way. The aim is to search for a set of smooth strictly monotonic functions  $h_i$ , called *warping function*, which eliminate phase variation such that the registered functions  $y_i(t)$  of the form

$$y_i(t) = x_i[h_i(t)] = (x_i \circ h_i)(t) \quad (1.1)$$

---

\*Footnote to the title with the “thankstext” command.

†Some comment

‡First supporter of the project

§Second supporter of the project

represents amplitude variation. Since monotone transformations do not destroy shape features the registered functions will possess the same sequences of peaks and valleys as the original functions  $x_i$ .

Phase variation is present in many important applications. A classical example is human growth data, where substantially interpretable shape features like the pubertal growth spurt occur at different ages. Many other examples stem from experimental data where, for example, phase variation arises as a consequence of different individual reaction times. Phase variation poses severe problems for the application of functional versions of commonly used multivariate data analyses such as computing pointwise means, variances and correlations; principal components analysis and canonical correlation analyses (Ramsay and Silverman [19]; Silverman [23]).

Traditional literature on the registration problem aims to define warping functions in such a way that registered functions  $y_i$  have all shape features aligned. A widely used method *landmark registration* removes phase variation by monotonically transforming the domain for each curve so that points specifying the locations of shape features are aligned across curves. Landmark registration has been studied in depth by Bookstein [2, 3], Kneip and Gasser [10] and Gasser and Kneip [6].

Many other methods not using landmarks have also been developed, partly in response to situations where landmarks are not clearly identifiable in all curves. A common property of the most important methods proposed in this context is to determine warping functions  $h_i$  by minimizing a distance  $d(x_i \circ h_i, \gamma)$  between registered functions  $y_i(t) = x_i(h_i(t))$  and a template  $\gamma(t)$ . There is a considerable literature proposing algorithms which aim to minimize the distance  $d_2(x_i \circ h_i, \gamma) = \|x_i \circ h_i - \gamma\|_2$ , where  $\|\cdot\|_2$  denotes the  $L^2$ -distance, see, for example, Sakoe and Chiba [20], Ramsay [17], Ramsay and Li [18], or Kneip et al. [12]. Usually additional regularization techniques are applied. Well-known problems with these techniques have lead to the development of alternative distance measures based on more sophisticated semi-norms (see e.g. Ramsay and Silverman [19], Wang and Gasser [27, 28, 29], or Srivastava et al. [26]).

All these method share a common point of view. The success of a registration method is assessed in terms of how well it is able align visible features. Templates are often determined iteratively from the sample and their construction aims to establish a “structural mean” which possess all common shape features at mean locations and with mean amplitude. Hence, traditionally registration tends to concentrate on establishing a most informative mean curve summarizing the sample functions.

However, more recent work also tends to apply registration procedures in the context of more complex problems of statistical data exploration and inference. In functional data analysis the most frequently applied procedures are based on identifying **low dimensional linear subspaces of functions** that are able to provide accurate approximations of the observed functional data. An essential tool is functional principal component analysis (FPCA), where sample curves are approximated as elements of the linear space generated by a few leading functional principal components. For functions exhibiting a registration

problem, Ashton et al. use a norm based method for aligning functions, and then apply FPCA separately for registered curves and warping functions.

For clustering functions, Sangalli et al. [22] propose a procedure which is based on several templates instead of only one scaled mean. This k mean approach assumes, that every curve in the set belongs to one of  $k$  specific clusters. The method tries to determine the mean (template) of any cluster iteratively and then using for example one of the above mentioned methods to align the curves within the cluster.

In this paper we study registration with respect to a best subspace selection. The idea is to determine warping functions  $h_i$  such that for a suitable  $K = 1, 2, \dots$  we have

$$x_i(t) \approx \sum_{j=1}^K a_{ij} \gamma_j(h_i^{-1}(t)). \quad (1.2)$$

for some basis functions  $\gamma_1, \dots, \gamma_K$  and individually different coefficients  $a_{i1}, \dots, a_{iK}$ .

Our approach provides an extension of some ideas already presented in Kneip and Ramsay [11]. A basic insight when analyzing (1.2) is a tight connection between optimal warping and dimensionality  $K$  of the subspace spanned by  $\gamma_1, \dots, \gamma_K$ . To see the point consider a very simple example: Let  $x_i(t) = a_i \sin(t + b_i)$ . Obviously, using warping functions  $h_i^{-1}(t) = t - b_i$ , these functions can be represented by a  $K = 1$  dimensional space with  $\gamma(t) = \sin(t)$ . But at the same time we have  $x_i(t) = a_i \sin(t + b_i) = a_i \sin(b_i) \cos(t) + a_i \cos(b_i) \sin(t)$ . Therefore, *without applying any warping* ( $h_i(t) = t$ ) the functions lie in a  $K = 2$ -dimensional space with  $\gamma_1(t) = \cos(t)$ ,  $\gamma_2 = \sin(t)$ ,  $a_{i1} = a_i \sin(b_i)$ , and  $a_{i2} = a_i \cos(b_i)$ . Note that in many important applications the presence of phase variation is already imposed from a substantial point of view (different reaction times, etc.). But this simple example shows that variation in the locations of shape features (peaks and valleys) may either be explained by phase variation or by amplitude variation  $\sum_{j=1}^K \gamma_j(t) a_{ij}$  for  $K \geq 2$ . There is no unique, identifiable truth and the only useful criterion is statistical interpretability. It is immediately seen that for any sample  $\{x_i\}$  an accurate representation (1.2) can be achieved with  $h_i(t) \equiv t$  if  $K$  is chosen sufficiently large, but in the presence of considerable phase variation the resulting dimension  $K$  may be much too large to allow a useful analysis.

In this work we take a different point of view of the registration problem. Registration is not simply considered as an alignment problem, our aim is to find an optimal trade-off between the complexities of amplitude and phase variation in order to obtain an accurate low dimensional representation of the curves for further analysis. A general discussion of the qualitative model (1.2) is given in Section 2. It is shown that existing template-based registration procedures in tendency aim to find a solution of (1.2) with  $K = 1$ . But in many interesting application a higher dimensional subspace representation, e.g.  $K = 2$  or  $K = 3$ , may be more appropriate, and lead to better interpretable statistical results than either a FPCA of unregistered curves or an analysis with usual norm-based registration methods. The amount of warping required to obtain an optimal

approximation will depend on  $K$ , and usually less complex warping functions have to be used if  $K$  is large. The situation is also illustrated by Figure 1, which shows a simulated sample which can be exactly described by a  $K = 2$  dimensional model 1.2. A detailed description how this sample is generated is given Section (3.1). Here peak alignment (which is similar to registering to  $K = 1$ ) increases the complexity of the warping functions. But it can be seen from an FPCA decomposition that the data cannot be described by two components anymore. Like in the unregistered case more components are needed to describe the peak aligned curves<sup>1</sup>. An explanation for this unexpected behavior is given in Proposition 2d.

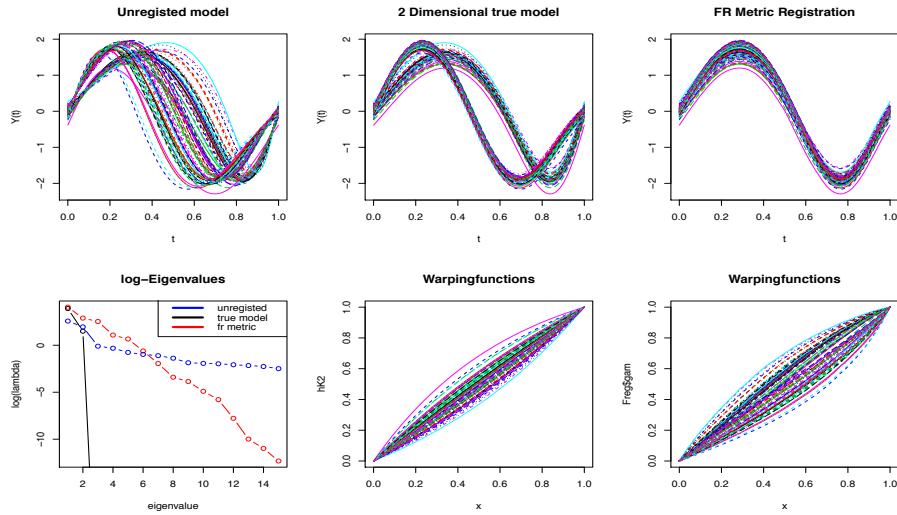


Fig 1: Example for curves generated by (1.2) with  $K = 2$ . The lower left Figure provides the log eigenvalues of an FPCA decomposition for the three types of registration given in the upper figures. The alignment of the peaks increases the model complexity (log-Eigenvalues) and the complexity of the warping functions.

Any suitable subspace registration should be based on a sophisticated algorithm which provides an effective solution to the fitting problem introduced by (1.2). In Section 3 we will describe an algorithm.

Defining such algorithms is not a trivial task. For example, Gervini and Gasser [7], as well as Claeskens, Silverman and Slaets [4] and Slaets, Claeskens and Hubert [24] present multi-resolution approaches to registration. Assuming on discretized observations, they define basis expansions for amplitude and phase variation and use algorithms designed for fitting mixed-effect models. Although this seems to work well in their applications, we will show in Section 2 that the

<sup>1</sup>We used the fdasrvf package for R for the alignment

machinery of mixed-effect models in general does not seem to be appropriate in a registration context. There is a considerable identification problem, since the structure of the covariance function of  $x_i(t)$  may not provide much information about an optimal level of registration.

The paper is organized as follows. In Section 2 we will study the qualitative model (1.2) and discuss resulting problems of identifiability. Section 3 contains a description of our algorithmic approach, while Section 4 provides a simulation study. Two applications, on human growth curves and gene expression data are presented in Section 5.

## 2. Registering to low dimensional linear spaces

### 2.1. Expressing data via phase and amplitude

We consider observations consisting of a sample of continuous random functions  $x_1, \dots, x_n$  defined on a common interval, that we may take as  $[0, 1]$  without losing generality.

There exists numerous ways to define the warping function  $h$ . For example with a composition of *warplet kernel functions* as introduced by Claeskens, Silverman and Slaets [4] or using *I-Splines* from Ramsay [16]. The only important restriction is that a warping function has to be an element of the space continuous and strictly increasing functions  $\mathcal{H}$ . Besides requiring that  $h(0) = 0$  and  $h(t) = t$  is often used. We follow the representation used in Ramsay and Silverman [19], where a warping function is defined as

$$h(t) = \frac{\int_0^t \exp(W(u))du}{\int_0^1 \exp(W(u))du}$$

with  $W(u) \in L^2[0, 1]$ .  $h$  is a bijection from  $L^2[0, 1] \rightarrow \mathcal{H}$ . This interpretation of  $h$  has the advantage that we can do further statistical analysis at the  $W$  functions. For example performing an FPCA which is not meaningful at  $h$ .

The functional inverse  $h^{-1}$  with the property  $h^{-1}[h(t)] = (h^{-1} \circ h)(t) = t$  for all  $t$  is uniquely defined, and the identity warping function  $\mathcal{I}$  given by  $\mathcal{I}(t) = t$  for all  $t$  acts as the unit element  $\mathcal{H}$  for functional composition. Note that  $h(0) = 0$  and  $h(1) = 1$  excludes simple shifts as in the sine-cosine example in the introduction. But common start and end points of  $h$  simplify the problem and are fairly natural in many applications. It is of course possible to modify this requirement in specific situations. We also want to emphasize that additional smoothness assumptions on  $x_i$  and  $h$  will be useful in many contexts, but they are not necessary for the general discussion in this section.

Following the introduction our qualitative approach can now be described as determining warping functions  $\{h_i\}_{i=1,\dots,n}$  as well as a  $K = 1, 2, 3, \dots$  and a corresponding linear subspace  $\mathcal{L}_K \subset L^2[0, 1]$  of continuous functions such that

for any suitable set of basis functions with  $\text{span}\{\gamma_1, \dots, \gamma_K\} = \mathcal{L}_K$  we have

$$y_i(t) := x_i(h_i(t)) = \sum_{j=1}^K a_{ij} \gamma_j(t), \quad i = 1, \dots, n \quad (2.1)$$

for some real-valued coefficients  $a_{i1}, \dots, a_{in}$ .

It will be shown below that for samples of functions possessing a common structure it is not a restriction to require that (2.1) holds exactly. But in practice there will of course exist random wiggles introduced by errors, and one then has to look for approximate solutions. The algorithm introduced in Section 3 will then estimate the solution with minimal  $L^2$ -approximation error.

Quite obviously the above qualitative model has many degrees of freedom and there are serious issues with identifiability. First of all the linear subspace  $\mathcal{L}_K$  can be identified, and the particular basis  $\gamma_1, \dots, \gamma_K$  could be chosen by the statistician. Our algorithmic solution (see Section 3) is based on using the eigenfunctions corresponding to the leading eigenvalues of the empirical second moment operator defined by  $M(x) = \frac{1}{n} \sum_{i=1}^n \langle x_i, x \rangle x_i$  for  $x \in L^2[0, 1]$ .

While this is not a real problem, also warping functions and the space  $\mathcal{L}_K$  may not be unique. A trivial non-identifiability consists in the fact that for an arbitrary function  $g \in \mathcal{H}$  (2.1) remains valid if  $h_i$  and  $\gamma_j$  are replaced by  $h_i^* = h_i \circ g$  and  $\gamma_j^* := \gamma_j \circ g$ ,  $j = 1, \dots, K$ . This effect can be eliminated by requiring that warping functions are **standardized** such that  $\bar{W}(u) = \frac{1}{n} \sum_{i=1}^n W_i(u) = 0$ . Note, that using  $W$  to standardize the warping function is different to the usual approach where the warping functions are standardized directly.

But standardizing can only be shown to solve problems of identifiability in the case  $K = 1$ . For  $K \geq 2$  there may exist different sets of normalized warping functions and different subspaces satisfying (2.1). If multiple solutions are present, the complexity of the phase variation depending on the observed curves varies for every solution. We require to choose the easiest phase variation which guarantees a low dimensional amplitude representation. Because we want to perform further analysis at  $W_i(u)$ , easy means in this context that the functions  $W_i(u)$  should be as close as possible to 0. This introduces an additional requirement for a suitable selection of  $W_i(u)$ :

- Under all possible sets  $\{W_i\}_{i=1, \dots, n}$  of warping functions  $\{h_i\}_{i=1, \dots, n}$  satisfying (2.1) and  $\bar{W}(u) = 0$  choose the solution with the smallest sample variance of  $W_i$  in the sense that

$$V(W) := \frac{1}{n} \sum_{i=1}^n \int_0^1 (W_i(u))^2 du$$

is minimal.

Registration is driven by the succession of local extrema (shape features) in each of the functions  $x_i$ . For any continuous function  $x$  one can determine locations  $0 < \tau_1^x < \tau_2^x < \dots < \tau_{q(x)}^x < 1$  and heights  $x(\tau_1^x), \dots, x(\tau_{q(x)}^x)$  of all isolated local extrema in the interior of  $[0, 1]$ . This means that for all  $j =$

$1, \dots, q(x)$  there exists an open neighborhood  $U(\tau_j^x)$  of  $\tau_j^x$  such that either  $x(t) \geq x(\tau_j^x)$  for all  $t \in U(\tau_j^x)$  or  $x(t) \leq x(\tau_j^x)$  for all  $t \in U(\tau_j^x)$ . If  $q(x) < \infty$  let  $p(x) = q(x) + 2$ , and let  $P(x) = (x(0), x(\tau_1^x), \dots, x(\tau_{q(x)}^x), x(1))^T \in \mathbb{R}^{p(x)}$  denote the corresponding  $p(x)$ -dimensional vector of heights of local extrema (including starting and end points).

We will say that a continuous function  $x \in L^2[0, 1]$  possesses a constant segment if there exists an interval  $[a, b] \subset [0, 1]$ ,  $a < b$ , such that  $x(t) = x(s)$  for all  $t, s \in [a, b]$ . Warping does not change monotone functions, and in general warping functions are not uniquely defined if functions have constant segments. Therefore, in the following discussion we will concentrate on curves which are strictly monotone between local extrema.

Note that for any continuous  $x$  and any warping function  $h$  the resulting function  $y = x \circ h$  has the **same vector**  $P(y) = P(x)$  of heights of local extrema. This means that the registered curves  $y_i$  in (2.1) will exhibit the same visual shape (in terms of the succession of local extrema) as the original functions  $x_i$ . But essentially  $P(x)$  is the **only structural feature** of  $x_i$  which is invariant against strictly monotone transformations. It is thus the driving force of identifiability of (2.1).

**Proposition 1.** *Assume that none of the curves  $x_1, \dots, x_n$  possesses a constant segment. Then for any  $i = 1, \dots, n$*

- a) *For any function  $y$  possessing no constant segments and satisfying  $p(y) = p(x_i)$  as well as  $P(y) = P(x_i)$  there exists a **unique**  $h \in \mathcal{H}$  such that  $x_i \circ h = y$ .*
- b) *Additionally assume (2.1) and suppose that there exists a unique  $y \in \mathcal{L}_K = \text{span}\{\gamma_1, \dots, \gamma_K\}$  such  $p(y) = p(x_i)$  as well as  $P(y) = P(x_i)$ . If  $y$  does not have a constant segment, the coefficients  $a_{ij}$ ,  $j = 1, \dots, K$ , as well as the warping function  $h_i$  in (2.1) are uniquely determined.*

Note that if  $K > 1$ , it is possible that the condition  $P(y) = P(x)$  does not have a unique solution  $y \in \mathcal{L}_K$ . In this case, there will exist several possible warping functions and corresponding coefficients. Among possible candidates the above additional condition will then lead to select the one with  $\frac{1}{n} \sum_{i=1}^n (W_i(u))^2 du = \min$ .

## 2.2. Registration and subspaces of functions

In practice a closer analysis of the sequence of peaks and values will provide information about the dimension  $K$  of a suitable subspace satisfying (2.1). Indeed, requiring (2.1) for some  $K < \infty$  is no restriction if all sample functions have a typical, common shape in terms of shape features.

For a set  $\{h_i\}$  of warping functions we will say that  $\dim\{x_1 \circ h_1, \dots, x_n \circ h_n\} = K$  if the registered functions  $y_i = x_i \circ h_i$  span a  $K$ -dimensional linear functions space, i.e. if (2.1) holds for some  $\gamma_1, \dots, \gamma_K$  and suitable coefficients  $a_{ij}$ ,  $i = 1, \dots, n$ ,  $j = 1, \dots, K$ .

**Proposition 2.** Assume that none of the curves  $x_1, \dots, x_n$  possesses a constant segment, and assume  $\mathbf{p} := \max_{i=1, \dots, n} p(x_i) < \infty$ .

- a) There then exists some set  $\{h_i\}$  of warping functions such that  $\dim\{x_1 \circ h_1, \dots, x_n \circ h_n\} \leq \mathbf{p}$ .
- b) Additionally assume that  $\mathbf{p} = p(x_1) = \dots = p(x_n)$  and that the empirical covariance matrix of the vectors  $P(x_1), \dots, P(x_n)$  has rank  $\mathbf{p}_0 \leq \mathbf{p}$ . There then exists some set  $\{h_i\}$  of warping functions such that  $\dim\{x_1 \circ h_1, \dots, x_n \circ h_n\} \leq \mathbf{p}_0$ .
- c) There exists a set  $\{h_i\}$  with  $\dim\{x_1 \circ h_1, \dots, x_n \circ h_n\} = 1$  if and only if  $\mathbf{p} = p(x_1) = \dots = p(x_n)$  as well as  $\mathbf{p}_0 = 1$ .
- d) If  $\dim\{x_1 \circ h_1, \dots, x_n \circ h_n\} = K$ , then there exists some set  $\{h_i^*\}$  of warping functions such that  $\dim\{x_1 \circ h_1^*, \dots, x_n \circ h_n^*\} = K+1$  and  $\frac{1}{n} \sum_{i=1}^n \int_0^1 (W_i^*(u))^2 du < \frac{1}{n} \sum_{i=1}^n \int_0^1 (W_i(u))^2 du$ .

Assertion a) of the proposition corresponds to Proposition 1 of Kneip and Ramsey (2008). The proposition tells us that already the number of peaks and valleys to be found in each curve  $x_i$  may provide an idea about an appropriate choice of  $K$ .

A consequence of Proposition 2c exact registration to  $K = 1$  requires  $\mathbf{p}_0 = 1$ , and it will also provide an alignment of peaks. The situation is, however, different for  $K \geq 2$ . Then Proposition 2b only provides a sufficient condition which is far from necessary. As has already been shown in Figure 1, a suitable registration to a higher dimensional space will not necessarily go along with an alignment of peaks.

Indeed, (2.1) may hold for some  $K \geq 2$  with  $K \ll \mathbf{p}_0$ , and it may even be true if  $p(x_i) \neq p(x_j)$  for some sample functions  $x_i, x_j$ . This is illustrated by Figure 2 which shows a  $K = 2$  dimensional example of sample functions generated by two Legendre Polynomials together with some amount of warping. Details can be found in Appendix B.1. In this case  $p(x_i)$  varies between 3 and 5. Using  $K = 2$ , the algorithm described in Section 3 is able to retrieve the correct amount of warping and to reconstruct the two-dimensional space of registered functions. In contrast, an attempt to align peaks and to register to  $K = 1$  delivers unreasonable results.

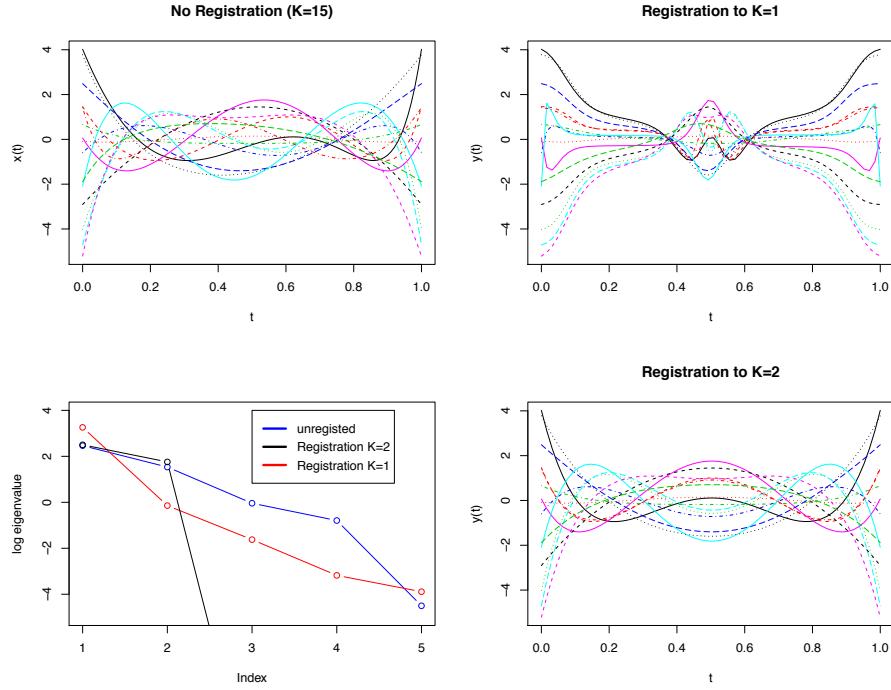


Fig 2: Registration of curves generated by Simulation B.1 using our algorithm from Section 3

Proposition 2d tells us that there will usually also exist a trade-off between dimensionality  $K$  and complexity of warping functions. Figure 3 provides an illustration. The generating process is described in Appendix B3. As seen in the upper right Figure the curves can be described with a  $K = 1$  model up to a negligible error. But when using our algorithm with  $K = 3$  (lower figure), the solution minimizing  $V(W)$  leads to a more complex amplitude model together with a drastic reduction of the complexity of warping functions.

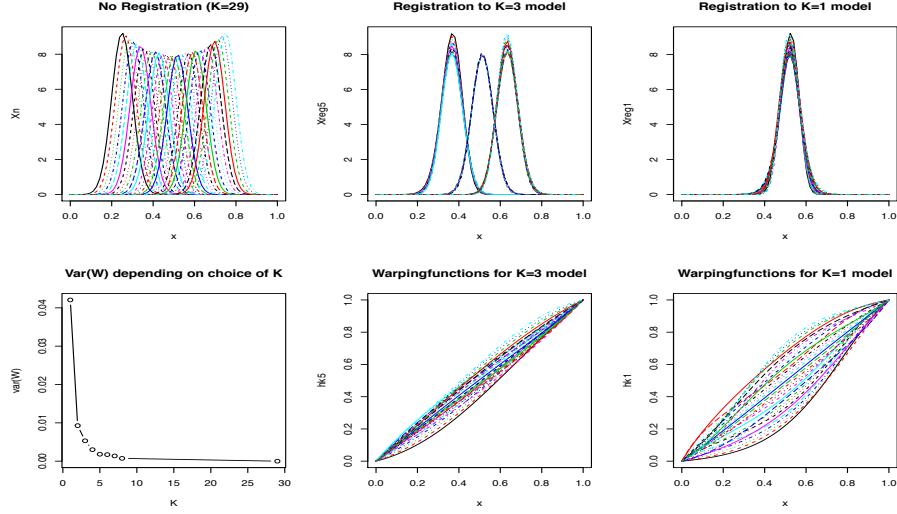


Fig 3: 29 curves generated by a gaussian kernel function with phase shift and amplitude variation. Registered using our algorithm from Section 3.

Proposition 2a the structure of underlying sample functions may provide information about a suitable dimension  $K$ . To illustrate this point consider a simple example motivated by our application on yeast genes discussed in detail in Section 4.2. Consider continuous periodic functions with period length equal to 1, and assume that in each period every curve just possesses one local maximum and one minimum.

When analyzing such functions  $x_i$  on the interval  $[0, 1]$ , periodicity just means that  $x_i(0) = x_i(1)$ ,  $i = 1, \dots, n$ . If each of the curves just has one maximum and one minimum, then  $p(x_1) = \dots = p(x_n) = 4$ , and by  $x_i(0) = x_i(1)$  Proposition 2b) implies the existence of a set  $\{h_i\}$  of warping functions such that  $\dim\{x_1 \circ h_1, \dots, x_n \circ h_n\} = 3$ .

It is indeed simple to construct such 3 dimensional space  $\mathcal{L}_3$  analytically. For example, let  $\gamma_1(t) \equiv 1$ ,  $\gamma_2(t) = \sin(2\pi t)$ ,  $\gamma_3(t) = \cos(2\pi t)$ , and  $\mathcal{L}_3 := \text{span}\{\gamma_1, \gamma_2, \gamma_3\}$ . Quite obviously, for any  $x_i$  there exists a unique element  $y_i \in \mathcal{L}_3$  with  $p(x_i) = p(y_i) = 4$  and  $P(x_i) = P(y_i)$ . By Proposition 1 we can thus conclude that there are unique warping functions  $h_i$  and unique coefficients  $a_{ij}$  such that

$$y_i(t) := x_i(h_i(t)) = a_{i1} + a_{i2} \sin(2\pi t) + a_{i3} \cos(2\pi t), \quad t \in [0, 1], \quad i = 1, \dots, n \quad (2.2)$$

Note that the functions  $y \in \mathcal{L}_3$  have their local extrema at different locations, depending on  $a_{i2}$  and  $a_{i3}$ . Registration to  $\mathcal{L}_3$  therefore does not lead to an alignment of shape features.

But  $\mathcal{L}_3$  is not the only possible candidate space. Consider the space  $\mathcal{L}_3^*$  of all polynomials  $y_{b_1, \dots, b_5}(t) = \sum_{j=1}^5 b_j t^{j-1}$  of order 5 satisfying the constraints

$y_{b_1, \dots, b_5}(0) = y_{b_1, \dots, b_5}(1)$  as well as  $y'_{b_1, \dots, b_5}(0) = y'_{b_1, \dots, b_5}(1)$ . This is again a three dimensional space of functions with identical starting and end points, while the  $y'_{b_1, \dots, b_5}(0) = y'_{b_1, \dots, b_5}(1) \neq 0$  generates functions with one local maximum and one minimum in the interior of  $[0, 1]$ . There thus exists a set of warping functions  $\{h_i^*\}$  such that  $x_i \circ h_i^* \in \mathcal{L}_3^*$ . The two spaces  $\mathcal{L}_3^*$  and  $\mathcal{L}_3$  are not identical.

Our side condition of functions  $W_i$  with minimal empirical variance will then select one of the candidate spaces in dependence of  $x_1, \dots, x_n$ . For example, in the special case where the analytical structure of  $x_i$  already corresponds to Fourier functions such that  $x_i \in \mathcal{L}_3$ , then of course the solution with minimal  $V(W)$  is  $W_i(u) = 0$  which corresponds to  $h_i(t) = t$  for all  $i$  (i.e. no warping at all). If the original (unregistered) sample itself is not low dimensional, then our approach is to determine the linear subspace where the least amount of warping is necessary. It is, of course, entirely possible that the structure of  $x_1, \dots, x_n$  is such that for suitable warping functions we have  $\dim\{x_1 \circ h_1, \dots, x_n \circ h_n\} = 2$  or even  $\dim\{x_1 \circ h_1, \dots, x_n \circ h_n\} = 1$ . We want to note, however, that we do not have a formal proof of whether or not the condition of minimal  $V(W)$  always leads to a unique solution.

### 2.3. Connection to existing norm-based registration procedures

As already mentioned in the introduction the most commonly used registration procedures tend to determine warping functions by minimizing a distance  $d(X_i(h_i), \gamma)$  between  $x_i \circ h_i$  and a single template  $\gamma$ . A template may be either a prior known or be determined, e.g. iteratively from the data. The oldest and most widely applied procedures rely on minimizing an  $L^2$ -distance

$$d^2(x_i(h_i), \gamma) = \int_0^1 [x_i(h_i) - \theta(t)]^2 dt,$$

possibly subject to some further regularity conditions. Note that  $d^2(x_i(h_i), \gamma) = 0$  for all  $i = 1, \dots, n$  if and only if

$$x_i(h_i(t)) = \gamma(t), \quad t \in [0, 1], i = 1, \dots, n,$$

which means that after registration all functions are identical. This can only be true if there are no amplitude differences, and in the context of (2.1) one may speak of a “ $K = 0$  dimensional approximation”. It is commonly known that this method does not work very well in the presence of substantial amplitude variation, since there is then a tendency towards warping functions that tend to pinch in the regions over which curves are non-zero (see for example Ramsay and Li [18]).

The problems of simple  $L^2$ -based registration have led to the development of more advanced distance measures based on semi-norms  $d$  with the property that  $d(x_i(h_i), \gamma) = 0$  if  $x_i \circ h_i = a_i \gamma$  for some  $a_i \in \mathbb{R}$ . Important methods in this context have been proposed, for example, by Ramsay and Silverman [19], Wang and Gasser [29], or Srivastava et al. [26].

If  $d(X_i(h_i), \gamma) \approx 0$  then  $x_i(h_i) \approx a_i\gamma$ , and (2.1) holds with  $K = 1$ . Such approaches may thus be interpreted as registration procedures which determine warping function such that registered curves  $y_i = x_i \circ h_i$  are as “close as possible” to the elements of a one dimensional linear functions space, where closeness is measured by the particular semi-norm employed.

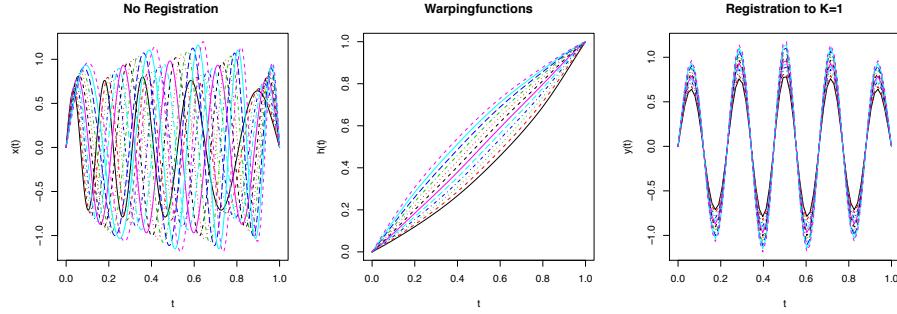


Fig 4: Registration of curves generated by Simulation B.2 to  $K=1$  using our algorithm from Section 3.

Different from  $K > 1$ , registration to a one dimensional subspace is closely connected to an alignment of shape features.  $x_i \circ h_i = a_i\gamma$ , then necessarily  $P(x_i) = P(a_i\gamma)$ , and for all local extrema we have  $\tau_j^{x_i \circ h_i} = \tau_j^\gamma$  as well as  $h_i(\tau_j^\gamma) = \tau_j^{x_i}$ . If the warping functions are standardized directly, then  $\frac{1}{n} \sum_{i=1}^n \tau_j^{x_i} = \frac{1}{n} \sum_{i=1}^n h_i(\tau_j^\gamma) = \tau_j^\gamma$ , and it is no restriction to require  $\gamma = \frac{1}{n} \sum_{i=1}^n x_i \circ h_i$ . In this case  $\tau$  represents a “structural mean” with the property that all shape features can be found at average locations and with average amplitude. This corresponds to the aim of landmark registration.

For samples  $x_1, \dots, x_n$  with complicated amplitude structure a one dimensional representation may not provide a good approximation, and  $d(x_i(h_i), \gamma) \gg 0$  for all possible choices of  $h$  and  $\gamma$ . In this situation the method of Srivastava et al. is of particular interest, since the specific properties of the Fisher-Rao metric employed will still tend to provide a solution where the main shape features are aligned. To our experience this procedure constitutes the best existing alignment method. But as shown above, norm-based registration may tend to explain too much by phase variation, leading to unnecessarily complex warping functions as well as in registered functions which cannot be well represented by a low dimensional subspace. While in the example given Figure 1 registration to  $K = 1$  may still be seen as a useful tool to define a “most representative mean curve” summarizing the sample, standard norm-based registration will fail in the example presented in Figure 2.

## 2.4. Registration and FPCA

When considering registered curves  $y_i = x_i \circ h_i$  representation (2.1) of course resembles a usual decomposition in terms of a truncated Karhunen-Loëve decomposition using the leading  $K$  principal components. A minor difference only consist in the fact the FPCA yields a decomposition of  $y_i - \frac{1}{n} \sum_{j=1}^n y_j$ , while in our approach we do not subtract sample means. This is motivated by the fact that, as explained in Section 2.2, the basic theme of registration is a modeling of shape features. But the visual shapes of the functions  $y_i$  and  $y_i - \frac{1}{n} \sum_{j=1}^n y_j$  may be very different.

As mentioned above, there will exist a trade-off between dimensionality  $K$  and complexity of warping functions. In principle, it will always be possible to find a sufficiently high dimension  $\kappa$  such that (2.1) approximately holds without applying any warping functions, i.e.

$$x_i(t) \approx \sum_{j=1}^{\kappa} a_{ij} g_j(t), \quad i = 1, \dots, n \quad (2.3)$$

Instead of using the empirical covariance operator, orthonormal functions  $g_j$  can be determined as eigenfunctions of the empirical second moment operator defined by  $M(x) = \frac{1}{n} \sum_{i=1}^n \langle x_i, x \rangle x_i$  for  $x \in L^2[0, 1]$ . If the sample mean is zero, then the above representation is equivalent to FPCA.

An important point is that even a small amount of registration may dramatically improve approximation by low dimensional subspaces. The effect is illustrated by Figure 5. We essentially refer to the example of Figure 5, but in order to define zero mean random functions the generating process is slightly modified by flipping the curves randomly. Details are explained in Appendix B3. Resulting sample curves are shown in the upper left part of Figure 5. Using a set of simple warping functions, similar to those presented to the lower middle part of Figure 3, the sample can almost exactly be registered to a  $K = 3$  dimensional space. On the other hand a  $\kappa = 3$  dimensional FPCA decomposition (2.3) does not provide any reasonable approximation of the unregistered sample curves, as seen in the the upper middle part of the figure.

Recall that for i.i.d random functions  $x_i(t)$ , up to sampling error the (truncated) Karhunen-Loëve decomposition (2.3) is to be determined from eigenvalues  $\lambda_j$  and eigenfunctions of the underlying covariance operator. The scores  $a_{ij}$  are uncorrelated random coefficients with  $\text{var}(a_{ij}) = \lambda_j$ . There is thus a crucial difference between (2.1) and (2.3). While registration based on (2.1) exploits common shape features of observed sample functions, many structurally different random processes may possess the same covariance function. This is shown in the lower left part of Figure 5, which provides sample functions from a Gaussian process possessing same covariance function as the original data generating process defined in Appendix B3. The only difference is that scores  $a_{ij}$  are now independent  $N(0, \lambda_j)$ -distributed random variables, while the original process goes along with an extremely non-Gaussian distribution of scores. The Gaussian sample shows functions which are structurally very different from the original

samples, and no simplification via registration is possible. At the same time, in contrast to original sample curves, a  $\kappa = 3$  dimensional FPCA decomposition (lower middle part of Figure 5) provides a structurally reasonable, although strongly “smoothed”, approximation of the Gaussian sample functions.

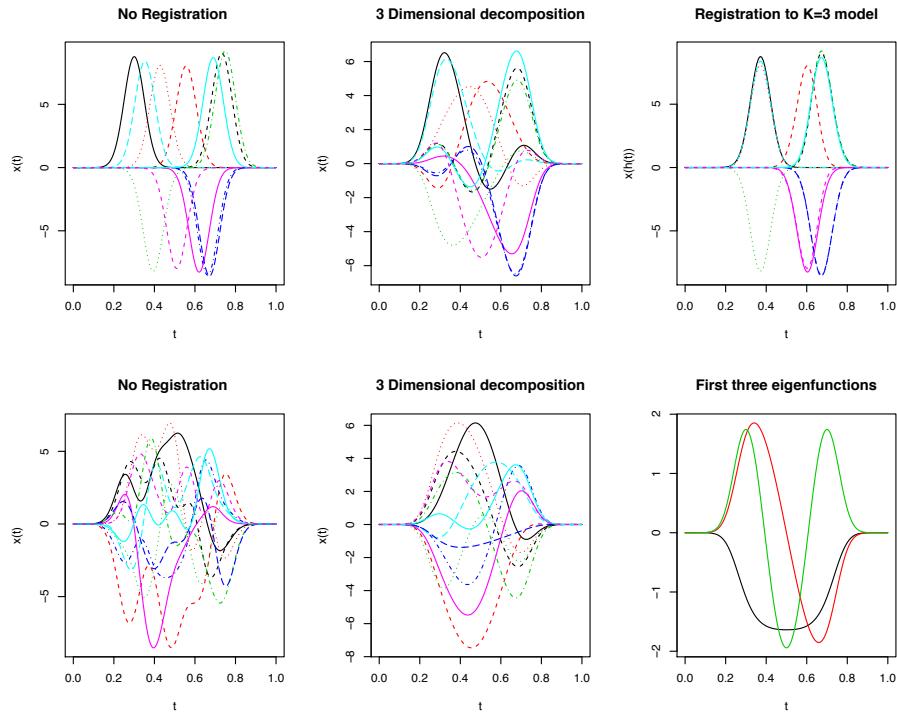


Fig 5: Each left picture shows 12 randomly sampled curves from the same space but with different eigenfunction scores. The lower left shows the first three eigenfunctions which are in common for both curves.

Further insight can be obtained from another simple, analytical example:

- 1) Let  $x_1^*, \dots, x_n^*$  be independent realizations of a stochastic process  $X^* \in L^2[0, 1]$ , where  $X^*$  is a standard Brownian motion on  $[0, 1]$ . This implies  $E(X^*(t)) = 0$  and  $E(X^*(t)X^*(s)) = \min\{t, s\}$ . Possible sample functions  $x_i$  are shown in the left part of Figure 6.
- 2) Let  $x_1, \dots, x_n$  be independent realizations of a stochastic process  $X \in L^2[0, 1]$  defined as follows: For two independent random variable  $T$  and  $A$ , where  $T \sim U(0, 1)$  and  $A$  is a binary variable with  $P(A = 1) = P(A = -1) = \frac{1}{2}$  let  $X(t) = 0$  for  $0 \leq t < T$  and  $X(t) = A$  for  $T \leq t \leq 1$ . This also implies  $E(X(t) = 0)$  and  $E(X(t)X(s)) = \min\{t, s\}$ . Possible sample functions  $x_i$  are shown in the right part of Figure 6.

The processes  $X^*$  and  $X$  thus have zero means and identical covariance functions  $\sigma(t, s) = \min\{t, s\}$ . Therefore on a population level also functional principal components are identical. It is well-known eigenvalues of the corresponding covariance operator are  $\lambda_r = \frac{1}{(r-0.5)^2\pi^2}$ ,  $r = 1, 2, \dots$ , while corresponding orthonormal eigenfunctions are given by  $\gamma_r(t) = \sin((r - 1/2)\pi t)$ ,  $r = 1, 2, \dots$ , and we thus obtain decompositions

$$X^*(t) = \sum_{j=1}^{\infty} a_j^* \gamma_j(t), \quad X(t) = \sum_{j=1}^{\infty} a_j \gamma_j(t),$$

where  $Var(a_j^*) = Var(a_j) = \lambda_j$ . Formally, the main difference between the two processes is that  $X^*$  has a Gaussian distribution, while  $X$  is “extremely” non-Gaussian. By the slow rate of decrease of  $\lambda_j$  a fairly large value  $\kappa$  will be required in order to obtain a useful approximation (2.3).

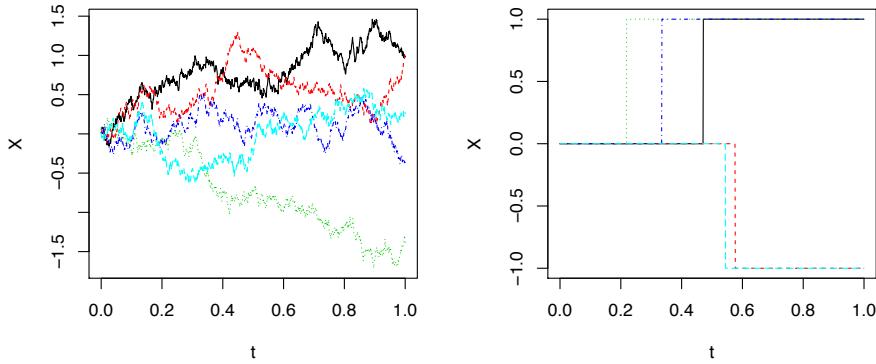


Fig 6: Sample of 5 random curves of process  $X^*$  left and the process  $X$  right

At the same time the resulting sample functions  $x_i^*$  and  $x_i$  are structurally very different. While it does not make any sense to speak of phase variation in the context of a Brownian motion  $x_i^*$ , phase variation obviously constitutes the main source of variability of  $x_i$ . In terms of (2.1) the sample functions can be represented by a  $K = 1$  dimensional model,  $x_i(h_i(t)) = a_i \gamma(t)$ ,  $i = 1, \dots, n$ , where  $a_i \in \{-1, 1\}$ ,  $\gamma(t) = 0$  for  $0 \leq t < 1/2$  and  $\gamma(t) = 1$  for  $1/2 \leq t \leq 1$ , while  $h_i(T_i) = 1/2$ . Here  $T_i$  denotes the realized value of the r.v.  $T$ . Since  $\gamma$  consist of two monotone segments, only the values  $h_i(0) = 0$ ,  $h_i(T_i) = 1/2$ , and  $h_i(1) = 1$  are fixed, all other values of  $h_i(t)$  are arbitrary. But for any reasonable interpolation scheme,  $\{h_i\}$  will be a simple, one-dimensional family of functions.

The examples show that registration should be done by methods which exploit *the functional shapes of observed functions*. They raise doubts about a general applicability of mixed-effect models in order to separate amplitude and

phase variation for the (discretized) data. Mixed-effect models rely on assuming a Gaussian distribution, and likelihood-based estimation relies on fitting the covariance structure of the observations. But this may run into serious identification problems, since, as shown above, the covariance function  $\sigma(t, s)$  of a sample of random functions *does not necessarily provide any information about a suitable amount of registration.*

### 3. Implementation

We want to construct an estimation procedure that automatically selects for a fixed  $K$  templates  $\gamma_j(t)$ ,  $j = 1, \dots, K$  which has to be orthonormal and provide for a given  $K$  the best possible least squares approximation. In most applications the curves  $x_i$ ,  $i = 1, \dots, N$  are only observed at discrete time points, in this case we use linear interpolations to get a continuous approximation of the curves. We define

$$S(h, K) = \min_{(\gamma_i): \langle \gamma_i, \gamma_j \rangle = \delta_{ij}} \frac{1}{N} \sum_{i=1}^N \|x_i(h_i(t)) - \sum_j \gamma_j(t) \langle \gamma_j, x_i(h_i) \rangle\|^2 \quad (3.1)$$

where  $\delta_{ij}$  is the Kronecker's function and  $\langle a, b \rangle = \int_0^1 a(t)b(t)dt$  the  $L^2$  inner product. For a given  $K$ , this function is only depend on the choice of  $h \in \mathcal{H}$ .

**Remark.** Suppose  $h_K$  minimizes (3.1) for  $K$  and  $h_K^*$  for  $K^*$  with  $K^* < K$ , then  $S(h_K, K) \leq S(h_K^*, K) \leq S(h_K^*, K^*)$ . This tells us that if we want to do a  $K$  low dimensional decomposition of the data for further analysis the registration with  $K$  templates is always the best choice. Therefore a method that does a registration with  $K = 1$  it is only meaningful to interpret a single template for example the scaled structural mean.

$V(W) = V(\log(h'))$  and  $S(h, K)$  are in interdependency with each other. The minimizing solution for  $V(W)$  is given by  $W_i(u) = 0$  for all  $i$  while in general this does not minimize  $S(h, K)$ . This kind of multi-objective minimization  $f(t) = \min_h(S(h, K), V(\log(h')))$  problems are important in many scientific fields, for example economics and engineering and well studied in the literature [5]. An important tool to define all meaningful solutions is the pareto concept. A sloppy definition of *pareto optimality* is that a solution is pareto optimal if there exists no other solution that dominates it. In our case we are interested in a pareto optimal solution which explains most of the variance in a low dimensional space while the reduction of phase variance is only secondary. This solutions are the at extremes values of the pareto optimal solution set. For a given  $K$  an optimality criteria for these points is defined by

$$S^* = \min_h S(h, K) \quad (3.2)$$

$$h^*(t) = \arg \min_h V(\log(h')) \text{ s.t. } S(h, K) = S^* \quad (3.3)$$

Because we do not know how to solve this equation analytical we stick to non-linear programming to determine a solution. Solving equation 3.2 is a very

complex task, to get an algorithm that performs this minimization in an adequate time we use several tricks to fasten up the computation. Every point at convex hull of pareto optimal solution set can be estimated by using *weighted sum scalarization*

$$S_P^*(K, w) := \min_h [w_1 \frac{S(h, K)}{S_0(h)} + w_2 V(\log(h'))], w > 0, w_1 + w_2 = 1 \quad (3.4)$$

to avoid scaling confusion we divide by  $S_0(h) := \frac{1}{N} \sum_{i=1}^N \|x_i(h_i(t))\|^2$ . Because we are interested in the extreme value,  $w_2$  has to be very small or to be precise converge to 0. The advantage of solving 3.4 instead of equation 3.2 directly is that minimization algorithm only has to solve an unconstrained minimization problem instead of two minimization problems where in addition one is constrained. Beside, this way of minimization has an additional advantage when facing real data. Mathematically 3.1 has no minimum but only an infimum if a strict factor model 1.2 is not present or  $K$  is chosen to small while 3.4 still has a minimum. Minimize equation 3.2 will thus result in an extreme warping. In that case  $w_2$  can be understood as an regulator to avoid extreme warping by explaining less with the registered  $K$  dimensional template model. How to set automatically an appropriate value for  $w_2$  is an open task for further research, at the current stage  $w_2$  has to be chosen by the statistician. In most cases a wrong choice of  $w_2$  is highly visible, an example can be seen in Figure 2, where warping to  $K = 1$  with a small  $w_2$  results in a strange registration outcome. From our experience the choice  $w_2$  very close to 0 usually delivers good results for further analysis, in our simulations and applications  $w_2 = 0.001$  was used.

Nonlinear programming algorithms are only applicable to minimize an objective function over a finite set of variables and not over functions. To overcome this issue we approximate  $W_i(u)$  with  $p$  orthonormal splines  $B_j(u)$ ,  $j = 1, \dots, p$  based on B-spline basis functions as described by Mason, Rodriguez and Seatzu [13]. The approximate representation  $W_i(u) \approx \sum_1^p B_j(u)c_{ij}$  leads to

$$h_i(t) \approx \tilde{h}(t, c_i) := \int_0^t \exp\left(\sum_1^p B_j(u)c_{ij}\right) du / \int_0^1 \exp\left(\sum_1^p B_j(u)c_{ij}\right) du \quad (3.5)$$

Because we require that  $B_j$  are orthonormal we get

$$V(W) \approx \tilde{V}(c) = \frac{1}{pN} \sum (c_{ij}^2).$$

To ensure that  $\bar{W}(u) = 0$  we always chose  $c_{ij}$  such that  $\frac{1}{N} \sum_{i=1}^N c_{ij} = 0$ ,  $\forall j = 1, \dots, p$ .

Recall that  $\frac{S(\tilde{h}(c), K)}{S_0(\tilde{h}(c))} = \frac{\sum_{i=K+1}^{\infty} \tilde{\lambda}_i(c)}{\sum_{i=1}^{\infty} \tilde{\lambda}_i(c)}$  with  $\tilde{\lambda}(c)$  being ordered eigenvalues of the second-moment operator  $M(x) = \frac{1}{n} \sum_{i=1}^n \langle x_i(\tilde{h}(c_i)), x \rangle x_i(\tilde{h}(c_i))$ . To estimate the eigenvalues we use the duality relation from Härdle and Simar [8] which usually can be computed very fast if  $N$  is small. The duality relation

states that the eigenvalues  $\tilde{\lambda}(c)$  corresponds to the eigenvalues of the  $N \times N$  matrix

$$D_{ij}(c) = \frac{1}{N} \int_0^1 x_i(\tilde{h}(t, c_i)) x_j(\tilde{h}(t, c_j)) dt$$

In our algorithm this integral is approximated by Riemann sums. The final minimization problem is given by

$$c^* = \arg \min_{c \in \mathbb{R}^{Np}} (1 - w_2) \frac{\sum_{i=K+1}^{\infty} \tilde{\lambda}_i(c)}{\sum_{i=1}^{\infty} \tilde{\lambda}_i(c)} + w_2 \tilde{V}(c) \quad (3.6)$$

where  $c$  is a vector of size  $Np$ , which has to be passed to a minimization algorithm. We decide to use the "newuo" algorithm developed by Powell [15] which has the advantages to be able to handle a large amount of variables in endurable time.

To evaluate optimal parameter  $c^*$  equation 3.6 with initial values  $c_{ij} = 0 \forall i, j$  which corresponds to  $h_i(t) = t \forall i$  is solved.

The optimal warping functions  $h^*$  are computed with 3.5 using  $c^*$ . To get the optimal template decomposition, let  $\theta_j$ ,  $j = 1, \dots, N$  be the eigenvectors of  $D(c^*)$  then

$$\begin{aligned} \gamma_j^*(t) &= \frac{1}{\sqrt{N \tilde{\lambda}_l(c^*)}} \sum_{i=1}^N \theta_{ij} x_i(h_i^*(t)) \\ a_{ij}^* &= \theta_{ij} \sqrt{N \tilde{\lambda}_l(c^*)} \end{aligned}$$

Until now consider that  $K$  is fixed in many situations  $K$  will be unknown. We recommend to use  $\mathbf{p}_0$  from Proposition 2b as an initial guess for  $\hat{K}$  and follow Algorithm 1 to estimate  $K$  and  $c^*$ . Using  $\hat{c}$  as the a new starting point for the minimization algorithm in each iteration fastens up the procedure, because usually  $\hat{c}$  is already close to  $c^*$ . Besides in case of a very difficult minimization problem this strategy overcomes a disadvantages of the nonlinear programming algorithm which rarely tend to stuck in a local minimum. The stop value  $\mathbf{d}$  depends on the model assumptions.

For example if we assume the curves  $x_i(t_k)$ ,  $i = 1, \dots, n$   $k = 1, \dots, T$  are observed at discrete equidistant time points  $t = t_1, \dots, t_T$  and a true factor model is given. Because linear interpolation us used to compute  $S(h, K)$  we have to take the interpolation error into account. We suggest using  $\mathbf{d} = \mathcal{O}(T^{-3})$  in this case. If we assume that the curves are additional contaminated by an error  $\epsilon_{ik} \sim N(0, \sigma^2)$  then  $\mathbf{d} = \hat{\sigma}^2$  is suggested. Where estimated  $\hat{\sigma}^2$  is some non parametric estimator applied at  $x_i(t_k)$ .

### 3.1. Shifted sinus functions

We also use our method to compute the introductory example given in Figure 1. This simulation is closely connected to the analysis of the growth curve in Section 4.1. Let  $t = [1, 2\pi]$ , we construct two groups where  $z_{1i}, z_{2i} \sim \mathbf{N}(0, 0.03)$ ,  $z_{3i}, z_{4i} \sim$

---

**Algorithm 1** recommend pseudo code if K is unknown

---

```

1: set  $\hat{K} = p_0$  and  $c = 0$ 
2: loop:
3: minimize 3.6 using  $\hat{K}$  and  $c$ .
4: get results  $\hat{c}$ .
5: if  $d \geq \sum_{l=\hat{K}+1}^{\infty} \tilde{\lambda}_l(\hat{c})$  then
6:   set  $c = \hat{c}$ ,  $\hat{K} = \hat{K} - 1$ 
7:   goto loop.
8: Output  $K = \hat{K} + 1, c^* = c$ 

```

---

$\mathbf{N}(0, 0.4)$   $i = 1, \dots, 90$

$$y_i(t) = \begin{cases} z_{4i}/4 + (1 + z_{1i})\sin(h_i(t)) + 1.05(1 + z_{2i})\sin(h_i(t)^2/(2\pi)) & \text{for } i = 1, \dots, 45 \\ z_{4i}/5 + (2 + z_{1i})\sin(h_i(t)) + (-0.5 + z_{2i})\sin(h_i(t)^2/(2\pi)) & \text{for } i = 46, \dots, 90 \end{cases}$$

where the warping functions are given by

$$h_i(t) = 2\pi \frac{\exp(z_{4i}(t+3)/6)}{\exp(z_{4i})}$$

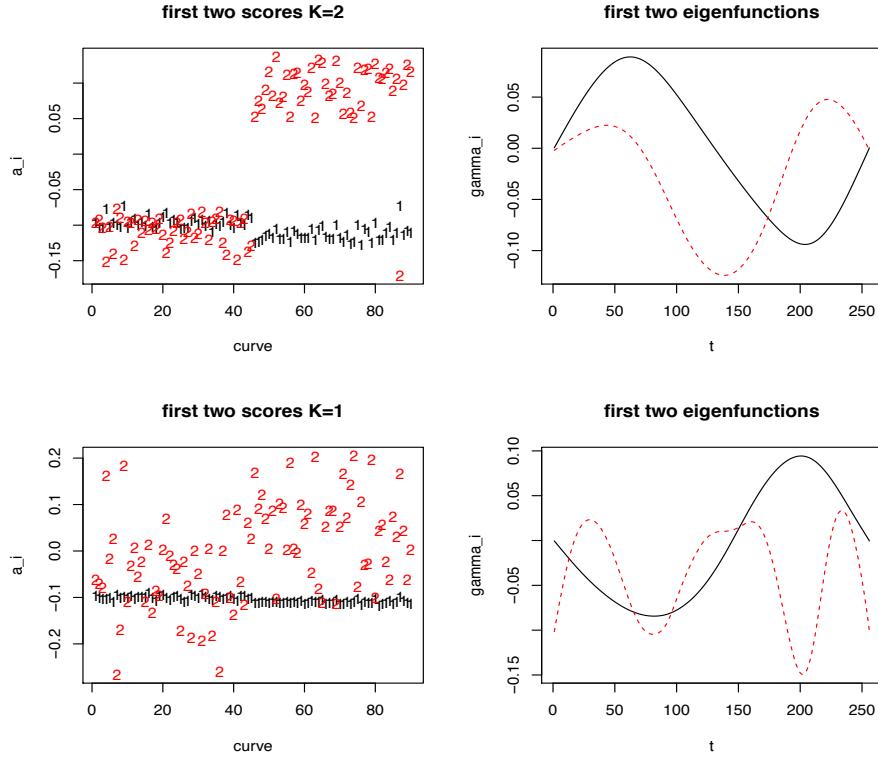


Fig 7: Numbers mark the score

To show the interdependency of phase and amplitude we did registration with  $K = 2$  and  $K = 1$ .  $K = 1$  can not deliver optimal representation by Proposition 2b. We can observe that with  $K = 2$  the scores of the first 46 curves are significantly different from the others. Obviously we are not able to recover  $z_1, z_2$  because our basis functions  $\gamma$  are different from the generating ones. Our automatic clustering and registration within this clusters corresponds pretty much to a similar simulation from Sangalli et al. [22] where K-mean clustering is used to identify the clusters. Our method can be considered as more flexible because no a priori clustering is needed. Table 1 shows how different choices of  $K$  effects the actual representation. Here  $K$  is the dimension used for registration and  $K^*$  the dimension used for decomposition. The impact of Proposition 3 and Proposition 2c is obvious. Registering to  $K = 1$  and then using  $K^* = 2$  to decompose the data is significantly worse than registering to  $K = 2$  from the start.

In this simulation  $z_{3i}$  is independent from  $i$ , as we will see in our analysis

	$V(W)$	$S(h, K)$
$K = 1, K^* = 1$	0.06758	0.006233
$K = 1, K^* = 2$	0.06758	0.001315
$K = 2, K^* = 1$	<b>0.03246</b>	0.3694242
$K = 2, K^* = 2$	<b>0.03246</b>	<b>0.000811</b>

TABLE 1  
*Model comparison*

of the growth curves it might be useful to take a closer look at the warping functions for clustering data or identifying time series trends as done by Poss and Wagner [14].

## 4. Applications

### 4.1. Berkley Growth Data

The growth curves were gained by measuring the growth of kids over a timespan of 18 years. What usually been analysed are not the growth curves directly but their smoothed velocity (first derivative). The dataset is not only interesting because there exists a registration problem, but because two groups are present. The dataset is popular to show the power of a registration method for example in Sangalli et al. [22] or Srivastava et al. [26]. We decided to use this well known real dataset to show exemplary how to perform statistics using our method. In particular we will present a method for group identification. From biology we know, growth features of boys and girls are different. Not only in a temporal sense but also in structure of the growth. These preconditions are very similar to simulation 3.1. For comparison we perform a registration with  $K = 1$  and  $K = 2$ .

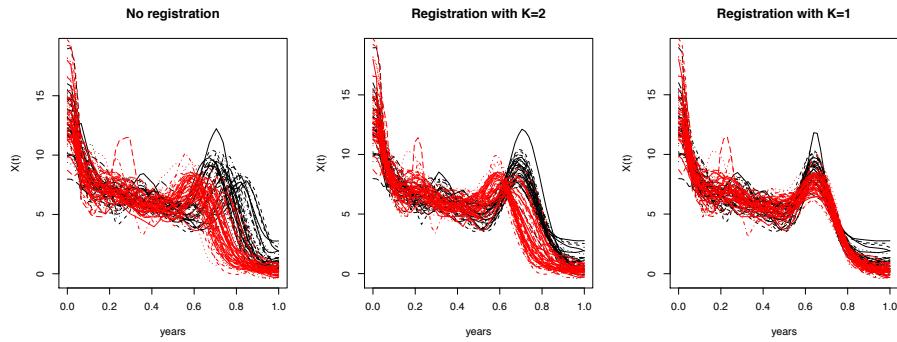


Fig 8: Girls are coloured red, boys back

When choosing  $K = 1$  boys and girls are matched at one single peak. Most of the variance is explained with warping as seen in table X. More interesting is the alignment with  $K = 2$  in simulation 3.1 we saw that our method is capable to identify groups if they have a different structure. In this application the templates are automatically chosen such that the girls are matched with only a few errors to one group and the boys to the other.

From the introductory sinus example we know that there is a interdependency between the representation of information in the warping and the synchronized space. To take this into account we decide to embed  $W(t)$  in our analyze for a fair comparison.

The procedure will work as follows. We apply our method with  $K = 1$  and  $K = 2$  to the first derivative of the growth curves  $x_i(t)$  to get an estimator for  $W_i(u)$  and  $y_i(t)$ . We follow these steps to apply the logit model introduced by Berkson [1]

- Code outcome binary,  $Z_i = 1$  (girls)  $Z_i = 0$  (boys)
- Template decomposition  $y_i(t) \approx \sum_{j=1}^K a_{ij}\gamma_j(t)$
- FPCA on  $\hat{W}_i(u) \approx \sum_{j=1}^P \vartheta_{ij}\varphi_j(u)$
- Construct logit model

$$P(Z_i = 1) = \frac{\exp(f(i))}{1 + \exp(f(i))}, \quad f(i) = \theta_0 + \sum_{j=1}^K a_{ij}\theta_j + \sum_{j=K+1}^{K+P} \vartheta_{i(j-K)}\theta_j$$

We use respectively three components for our analysis. For the unregistered curves we choose  $K = 3$  to reflect that the phase is modeled by additional basis functions. For the registration to  $K = 2$  we also use two template scores as well as one warping score ( $K = 2, P = 1$ ). When we register to  $K = 1$  the information moved from the synchronized space into the phase space therefore we use one template and two warping scores ( $K = 1, P = 2$ ) for the logit model.

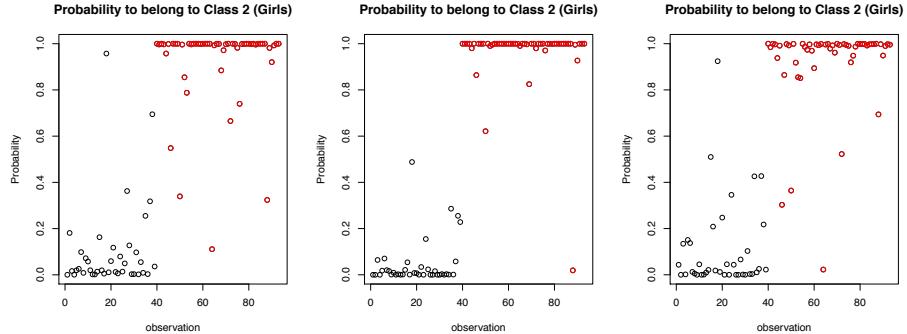


Fig 9: Unregistered,  $K=2$ ,  $K=1$ . Accuracy-rate: 0.94, 0.98, 0.94

For further analysis  $K = 2$  delivers the best performance. This is reflected by a higher accuracy level and also by a clearer separation of the two groups as seen in Figure 9.

#### 4.2. Yeast Genes

A yeast cell contains approximately 6000 genes. To save energy depending on the actual task only a few genes are active, Spellman et al. [25] suggest that only approximately 600 genes are active during the cell division. It is an important scientific task to identify these active genes. An interesting approach which uses functional data analysis was done by Zhao, Marron and Wells [30]. We will roughly follow the idea of their work, but use our registration procedure to improve the identification of active curve.

#### 4.2.1. The experiment

Spellman et al. [25] measured the gene expression of all 6178 genes of a yeast cell during two cell cycles. The observation last 2 hours where 18 equidistant observations were made. Two cycles were observed because this allows to identify the active genes due to periodicity.

Zhao, Marron and Wells [30] reduced due to technical issues the examined genes to 4489 basically all observations with missing values were dropped. The authors suggest that only approximately 200 genes were active.

#### 4.2.2. The setup

Simplified we can summarize the idea of Zhao, Marron and Wells [30] by fitting the data to periodic template functions that are intended to measure the periodicity. We stick to the most popular periodic curve, the sinus function. The activation of a gene did not take place at the same time, therefore a representation for periodic functions that allows for shifted is required. With  $t = (0, 4\pi)$  such a representation is given by our example from the introduction with

$$x_i(t) \approx \alpha_i \sin(t + s_i) = \alpha_i \cos(s_i) \sin(t) + \alpha_i \sin(s_i) \cos(t)$$

the scores are given by  $a_i = (\alpha_i \cos(s_i), \alpha_i \sin(s_i))$ . The idea is, that if  $\|a_i\|$  is close to zero there is no periodicity present. To visualize the selection we perform a k-medoids cluster algorithm by Kaufman and Rousseeuw [9] with 4 clusters at the scores  $a_i$  of the 400 curves where  $\|a_i\|$  is biggest.

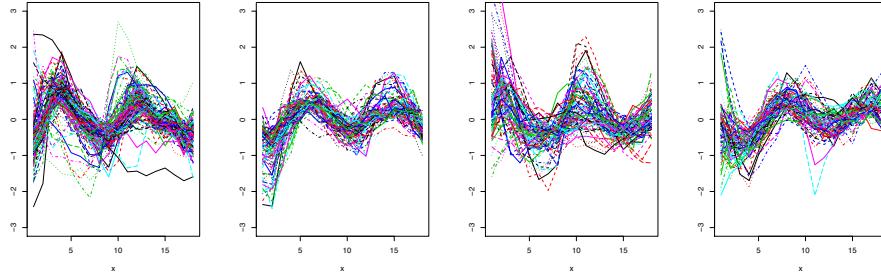


Fig 10: The first 400 curves where  $\|a_i\|$  is biggest clustered in 4 groups

In Figure 10 we can see, that the clustering basically separates curves due to their phase. The reason is that the scores are directly connected to the phase variation, because the angle from  $\alpha_i \cos(s_i) = 0$  represents the shift. In addition some non-periodic curves are distributed over all clusters. This is not surprising, because the used strategy has some disadvantages. The idea to identify curves due to the size of the coefficients bears the risk to select the wrong curves,

because high coefficients can have several origins. For example curves which are big, but not necessarily periodic will tend to get high coefficients. Another problem is caused by using pre-specified template basis functions. Only a curves which is an exact shifted sinus function is captured correctly. Other periodic functions are problematic and will generate scores that are not necessarily related to their periodicity.

In our approach warping was introduced to the problem, we use the 400 curves from the previous analysis as a pre-selection which we refined by using registration. In contrast to the previous approach we do not pre-specify the template basis functions. From Section 2.2 we knew, that every periodic function with two peaks can be represented with  $K = 3$  and additional warping. Therefore we use  $K = 3$  to determine the warping and the best template using our method. After the registration, some phase variation has been taken out of the synchronized space. As a result the scores do no longer basically represent the phase of the curves like in the original approach, but also the shape of the curves. This allows us to use the scores for a cluster analysis that separate periodic from non periodic functions which was not possible in the original approach. To cluster the scores we use again the k-medoids algorithm with 4 clusters, the result can be seen in figure 11.

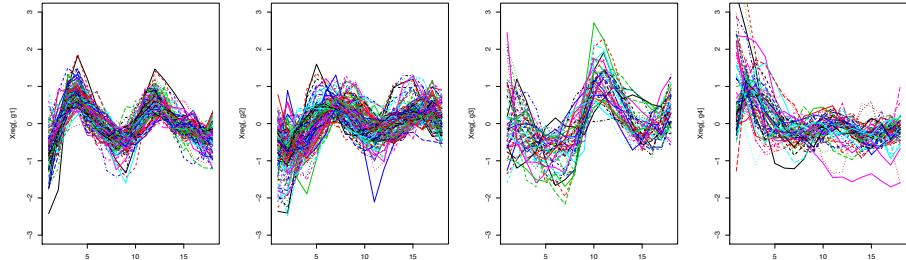


Fig 11: Clustering results with warping

While the curves in cluster 4 are no doubt not periodic, the curves in cluster 3 are controversial. We decide to follow a conservative path and drop the curves in cluster 3 and 4. After dropping these curves we are left with 278 curves. For a better visualization of our final results in Figure 12 we perform an additional clustering with 4 clusters and colored each cluster with a different color.

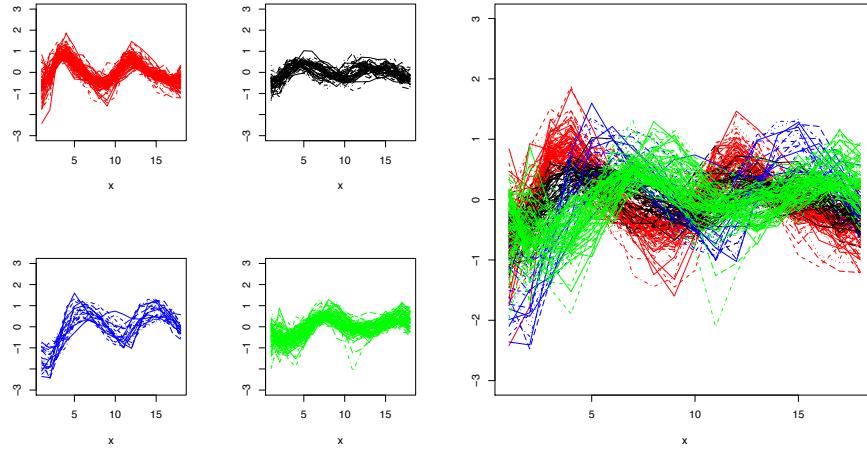


Fig 12: Curves that where finally labeled as periodic colored for better visualization. The left figures shows the clusters, the right all clusters joined together.

#### 4.3. Aneurisk Data

The last application is about applying our method to the aneurisk data set. The aneurisk data set contains the centerline of the Internal Carotid Artery (ICA) of 65 Patients. Depending where the aneurism is observed the the left or right ICA is stored. One aim of the anuerisk project is to explore the role of vessel morphology to determine the pathogenesis of cerebral aneurysms. To detect the differences in the vessel morphology a registration is needed, because length and also the shape of blood vessels differ from person to person. Attempts in this direction where done by Sangalli et al. [21] with using a  $K = 1$  method and by Sangalli et al. [22] using the k mean approach with  $k = 2$ . Our aim concerning the data is that we want to rely the position of the aneurysm to the vascular geometry of the ICA. In particular if the aneurysm is at the terminal bifurcation along the ICA or above. For the analysis already preprocessed data of the first directional derivative which was smoothed by Sangalli et al. [21] is used. To correct for different length of the ICA differs from person to person, we scale the data to  $t = (0, 1)$ . In most approaches the data is additionally adjusted if the left or right ICA is observed by flipping the sign of the  $x$  coordinate. In our approach such an adjustment was not done because we want to use the data as raw as possible. We suspect that the structural difference of left and right ICA might not only affect a simple coordinate flip but also other structural features. After testing the registration to different with different choiches of  $K$ , we find out that  $K = 3$  delivers the best result for our purposes. The data can be interpreted as a function  $x_i : \mathbb{R} \rightarrow \mathbb{R}^3$   $i = 1 \dots, 65$ . Analog to 1.2 a  $K$

dimensional representation of the registered curves are given by

$$x_i(h_i(t)) = y_i(t) := [y_{i,x}(t), y_{i,y}(t), y_{i,z}(t)] \approx \sum_{j=1}^K [\gamma_{x,j}(t), \gamma_{y,j}(t), \gamma_{z,j}(t)] a_{ij}$$

where  $a_{ij} = \int_0^1 \gamma_{y,j}(u) y_{i,x}(u) + \gamma_{y,j}(u) y_{i,y}(u) + \gamma_{z,j}(u) y_{i,z}(u) du$ . For this application 3.1 was modified to work with three spatial coordinates

$$S(h, K) = \min_{(\gamma_i): (\gamma_i, \gamma_j) = \delta_{ij}} \sum_{i=1}^N \left\| \sum_{m=(x,y,z)} X_{m,i}(h_i(t)) - \sum_{j=1}^K \gamma_{m,j} \left( \sum_{m=(x,y,z)} \langle \gamma_{m,j}, X_{m,i}(h_i) \rangle \right) \right\|^2$$

The results of the registration are shown in Figure 13.

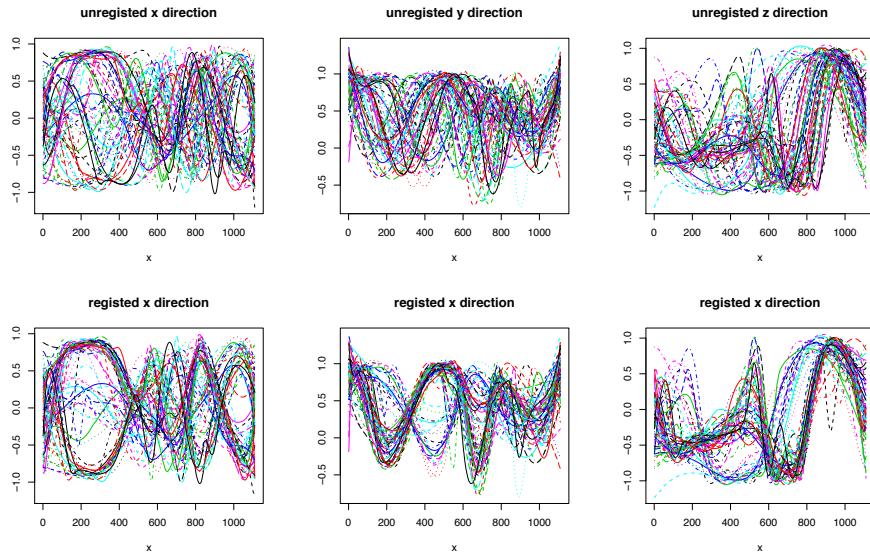


Fig 13: Upper pictures unregistered, Lower registration to  $K = 3$

#### 4.3.1. Results

The patients were divided into three groups: upper ICA, lower ICA and no aneurysm. To check if the registered  $K = 3$  dimensional representation can be used to determine differences between the groups the scores  $a$  were analyzed. The Boxplots 14 gives a hint that the third component is responsible for the position of the aneurism.

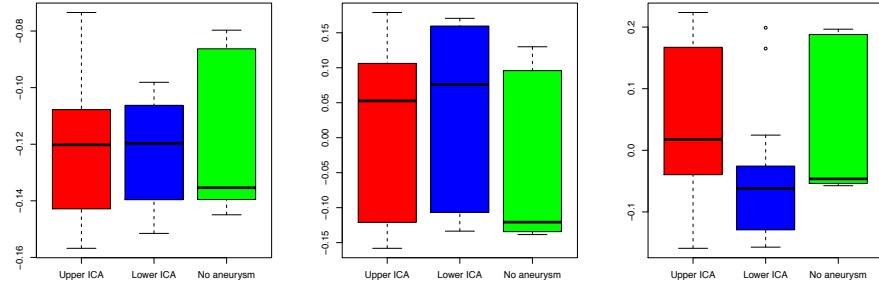


Fig 14: Patients grouped by aneurism position

To figure out what is been represented by the first two scores the patients where additionally divided into a group where the right ICA is examined and second group where the left ICA is observed. Boxplot 15 suggests that the first two components reflects which ICA is observed. The issue that two components are needed to represent the orientation of the ICA is an indicator that the differences between left an right ICA should not be handled purely by flipping the  $x$  coordinate.

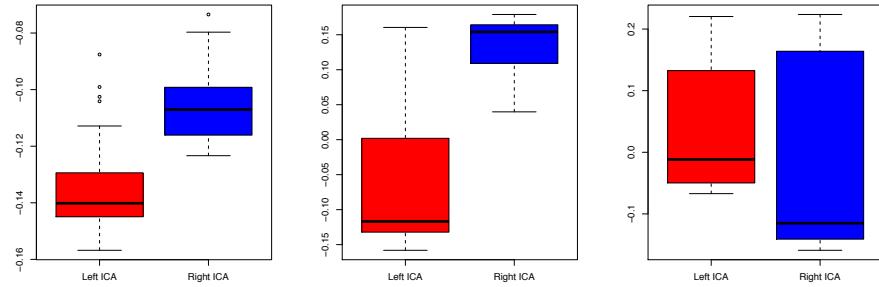


Fig 15: Patients grouped by left or right ICA

To support our findings we additionally perform a *Welch Two Sample t-test* where we compare the upper and lower ICA in Table 2 as well as left and right ICA in Table 3.

In real clinical situations an important task is to identify the aneurysm position automatically, because the "upper aneurysm" is the most dangerous one. The analyze in Section 4.1 concerning the growth data provide an excellent tool for an automatic detection procedure. For the analysis one group was formed by

	first score	second score	third score
p-value	0.8838	0.591	0.000296

TABLE 2

*Welch Two Sample t-test via groups before and after ICA*

	first score	second score	third score
p-value	$3.871e^{-11}$	$2.2e^{-16}$	0.1251

TABLE 3

*Welch Two Sample t-test via groups left and right artery*

the upper aneurysm patients and the other one by joining the lower and the no aneurysm patients. The previous analyze determined, that only the third component is connected the the aneurism position. This leads to a logit regression using only the third template score as well as the fist score from the warping functions which was able to achieved an accuracy rate of more than 70 percent.

#### 4.3.2. Conclusion

Our analysis supports the findings of Sangalli et al. [21], we can confirm that the position of the aneurism has influence on structure of the artery. In our analysis the score of  $\gamma_3$  represents the structure that which is connected to the position of the aneurysm. We can also identify if we face left or right via the first two scores. One advantage of our procedure that we did not preselect the template functions, therefore it may possible to get new insides analyzing the structure of  $\gamma_3$ .

## Appendix A: Proofs

### A.1. Proof of Proposition 1

- a. A warping from  $x_i$  to  $y$  can be constructed such that  $h_i(\tau_l^{x_i}) = \tau_l^y$  which maps all peaks together. Let further  $t_{il} := [\tau_l^{x_i}, \tau_{l+1}^{x_i}]$   $l = 1, \dots, p(x_i) - 1$  be open intervals, then for all  $l$ ,  $h_i(t_{il}) = x_i^{-1}(y(t_{il}))$  which is unique due to the uniqueness of the inverse.
- b. Assume wlog, that  $\gamma_1, \dots, \gamma_K$  are orthonormal. From a) we know that there exists a unique warping  $h_i(t)$ . Therefore  $a_{ij} = \int_0^1 \gamma_j(t) x_i(h_i(t)) dt = \int_0^1 \gamma_j(t) y(t) dt$ .

### A.2. Proof of Proposition 2

- a. see Kneip and Ramsay [11] Proposition 1a
- b. Let wlog  $x_1, \dots, x_{p_0}$  be a subset of  $x$  such that the rank of the covariance  $P(x_1), \dots, P(x_{p_0})$  is  $p_0$ . For any warping this subset can always be represented with  $p_0$  basisfunctions. Let  $h_i$  be any warpingfunction where  $x_i(h_i(\tau_k^{x_i})) = x_l(h_l(\tau_k^{x_l})) \forall k \in 1, \dots, p, i, l \in (1, \dots, p_0)$ . For any curve  $j > p_0$  there exists some  $b$  such that  $P(x_j) = \sum_{i=1}^{p_0} b_{ij} P(x_i)$ . Then we can construct a  $p_0$  dimensional representation with

$$y_i(t) = \begin{cases} x_i(h_i(t)) & \text{for } i = 1, \dots, p_0 \\ \sum_{j=1}^{p_0} b_{ij} x_j(h_j(t)) & \text{for } i > p_0 \end{cases}$$

From 1a we know that there exists  $h_i$  which warps  $x_i$  to  $y_i$ .

- c. " $\Rightarrow$ "  $p_0 = 1 \Rightarrow \exists a_{ij}, P(x_i) = a_{ij} P(x_j) \Rightarrow$  from 1a  $\forall i \exists$  unique warping  $h_i(t)$  from  $\frac{1}{a_{ij}} x_i(t)$  to  $x_j(t)$  where  $x_j(t)$  is chosen such that  $h_j(t) = t \Rightarrow \dim(x_i \circ h_i, \dots, x_n \circ h_n) = 1$   
 $\Leftarrow \dim(x_i \circ h_i, \dots, x_n \circ h_n) = 1 \Rightarrow x_i \circ h_i = a_{ij} (x_j \circ h_j) \Rightarrow P(x_i \circ h_i) = P(a_{ij} (x_j \circ h_j)) = a_{ij} P((x_j \circ h_j)) \Rightarrow P(x_i) = a_{ij} P(x_j) \Rightarrow p_0 = 1$
- d. Let  $\dim\{x_1 \circ h_1, \dots, x_n \circ h_n\} = K$ ,  $h^*$  can always be constructed: For one  $j$  with  $h_j(t) \neq t$  set  $h_j^*(t) = t$  and  $h_i^*(t) = h_i(t)$   $i \neq j$ .

### A.3. Choice of $\mathbf{d}$ intuition

In this section we do not want to retrieve asymptotic results for our estimator but just give some slight intuition for  $\mathbf{d}$ . Therefore we treat  $\gamma_j$  as a known parameter here, which will not make any sense when doing asymptotics because obviously  $\gamma$  is estimated during the algorithm and thus depends on the data.

### A.3.1. Factor model no error

Curves  $x_i$ ,  $i = 1, \dots, n$  are observed at equidistant timepoints  $t_1, \dots, t_T$ . Let  $h$  fulfill

$$x_i(t_k) = \sum_{j=1}^K a_{ij} \gamma_j((h_i)^{-1}(t_k))$$

suppose wlog  $t_k \leq h_i(t) \leq t_{k+1}$  then the error approximating from linear interpolation with  $\hat{x}_i(h_i(t))$  is given by

$$x_i(h_i(t)) - \hat{x}_i(h_i(t)) = \frac{1}{2}(h_i(t) - t_{k+1})(h_i(t) - t_k)x''_i(t_I), \quad t_I \in [t_k, t_{k+1}] \quad (\text{A.1})$$

therefore for all  $h_i(t) \in [0, 1]$

$$|\hat{x}_i(h_i(t)) - \sum_{j=1}^K a_{ij} \gamma_j(t)| \leq \frac{1}{8T^2} \max_{u \in [0, 1]} |x''_i(u)| = \mathcal{O}\left(\frac{1}{T^2}\right) \quad (\text{A.2})$$

and

$$(\hat{x}_i(h_i(t)) - \sum_{j=1}^K a_{ij} \gamma_j(t))^2 = \mathcal{O}\left(\frac{1}{T^4}\right) \quad (\text{A.3})$$

$$\frac{1}{n} \sum_{i=1}^n \int_0^1 (\hat{x}_i(h_i(t)) - \sum_{j=1}^K a_{ij} \gamma_j(t))^2 dt = \mathcal{O}\left(\frac{1}{T^3}\right) \quad (\text{A.4})$$

### A.3.2. Factor model with iid. Gaussian noise

Curves  $y_i$ ,  $i = 1, \dots, n$  are observed at equidistant timepoints  $t_1, \dots, t_T$ . Let  $h$  fulfill

$$y_i(t_k) = \sum_{j=1}^K \underbrace{a_{ij} \gamma_j((h_i)^{-1}(t_k))}_{x_i(t_k)} + \epsilon_{ik}.$$

let wlog  $t_k \leq h_i(t) \leq t_{k+1}$  then a linear interpolation of  $y_i$  gives

$$\hat{y}_i(h_i(t)) = \frac{h_i(t) - t_{k+1}}{t_k - t_{k+1}}(x_i(t_k) + \epsilon_{ik}) + \frac{h_i(t) - t_k}{t_k - t_{k+1}}(x_i(t_{k+1}) + \epsilon_{ik+1}) \quad (\text{A.5})$$

$$= \hat{x}_i(h_i(t)) + \underbrace{\frac{h_i(t) - t_{k+1}}{t_k - t_{k+1}} \epsilon_{ik} + \frac{h_i(t) - t_k}{t_k - t_{k+1}} \epsilon_{ik+1}}_{e_i(h_i(t))} \quad (\text{A.6})$$

this leads to

$$E_\epsilon \left( \sum_{i=1}^n \int_0^1 (\hat{y}_i(h_i(t)) - x_i(h_i(t)))^2 \right) = \\ (A.7)$$

$$\sum_{i=1}^n \int_0^1 (\hat{x}_i(h_i(t)) - x_i(h_i(t)))^2 - E_\epsilon(2x_i(h_i(t))e_i(h_i(t)) + 2\hat{x}_i(h_i(t))e_i(h_i(t)) + e_i(h_i(t))^2 dt) \\ (A.8)$$

looking at the segments gives

$$E_\epsilon \left( \int_0^1 x_i(h_i(t))e_i(h_i(t))dt \right) = \int_0^1 x_i(h_i(t))E_\epsilon(e_i(h_i(t)))dt = 0 \\ (A.9)$$

$$E_\epsilon \left( \int_0^1 \hat{x}_i(h_i(t))e_i(h_i(t))dt \right) = \int_0^1 \hat{x}_i(h_i(t))E_\epsilon(e_i(h_i(t)))dt = 0 \\ (A.10)$$

$$E_\epsilon \left( \int_0^1 e_i(h_i(t))^2 dt \right) < \sigma^2 \\ (A.11)$$

## Appendix B: Simulations

### B.1. Legendre Polynomials

A two dimensional factor model using the second and fourth Legendre-polynomial was simulated, i.e. we generated  $i = 1, \dots, 15$  curves over the interval  $[-1, 1]$  by:

$$x_i(t) = a_{i1} \frac{1}{2}(3h_i(t)^2 - 1) + a_{i2} \frac{1}{8}(35h_i(t)^3 - 30h_i(t)^2 + 3),$$

the warping functions  $h_i$  are given by

$$h_i(t) = 2 \frac{\exp(z_i(t+1)/2) - 1}{\exp(z_i) - 1} - 1, \quad a_i \neq 0 \quad \text{and } t \text{ otherwise}$$

with  $a_{i1}, a_{i2}, z_i$  are iid.  $\mathcal{N}(0, 1)$ .

Here the generated curves don't share the same peaks. Hence, any registration procedure which does not register the given sample to a two dimensional space, but instead tries to register to  $K = 1$ , will necessarily fail to give an exact low dimensional representation.

### B.2. 1 dimensional

We simulate 12 curves over the interval  $[0, 9]$  with  $a = (0.8, 0.8333, \dots, 1.2)$  by

$$x_i(t) = a_i(1 - (h_i(t)/9 - 0.5)^2) \sin(\pi h_i(t)),$$

where the warping functions are given by  $z = (-1.2, -1, \dots, 1, 1.2)$

$$h_i(t) = 9 \frac{e^{z_i t / 9} - 1}{e^{z_i} - 1}, \quad a_i \neq 0 \quad \text{and } t \text{ otherwise}$$

Even this simulation only has one component it is very demanding for most algorithms as Srivastava et al. [26] shows with a similar simulation. The problem here is that the peaks overlap, algorithms that use a local registration approach will likely stuck in a local minima.

### B.3. Shifted normal densitys

Let  $s(m) = 0.75 - 0.5m$ ,  $z(m) = 5 \frac{(s(m) - 0.5)^2}{2} + 1$ ,  $m \in [0, 1]$ , we generate a common covariance for both samples with  $t, s \in [0, 1]$

$$\Gamma(t, s) = \int_0^1 \frac{z(m)^2}{2\pi} \exp\left(\frac{(s(m) - t)^2 + (s(m) - s)^2}{2}\right) dm$$

let  $\phi_i(t)$  being the ordered eigenfunctions and eigenvalues  $\lambda_i$  the eigenvalues of  $\Gamma$ , then  $x_j^*(t)$  is generated by sample  $a_{ij}^* \sim N(0, \lambda_i)$ ,  $j = 1, \dots, N$  while we generate  $x_j(t)$  such that  $a_{ij} = (2m_i - 1) \int_0^1 \phi_i(t) \frac{z(k_j)}{\sqrt{2\pi}} \exp\left(\frac{(s(k_j) - t)^2}{2}\right) dt$ ,  $k_j \sim U(0, 1)$ ,  $m_j \sim B(1, 0.5)$ ,  $j = 1, \dots, N$

$$x_j^*(t) = \sum_{i=1}^{\infty} a_{ij}^* \phi_i(t), \quad x_j(t) = \sum_{i=1}^{\infty} a_{ij} \phi_i(t)$$

## References

- [1] BERKSON, J. (1944). Application of the Logistic Function to Bio-Assay. *Journal of the American Statistical Association* **39** 357–365.
- [2] BOOKSTEIN, F. L. (1978). *The Measurement of Biological Shape and Shape Change*. Springer.
- [3] BOOKSTEIN, F. L. (1997). *Morphometric Tools for Landmark Data: Geometry and Biology. Geometry and Biology*. Cambridge University Press.
- [4] CLAESKENS, G., SILVERMAN, B. W. and SLAETS, L. (2010). A multiresolution approach to time warping achieved by a Bayesian prior posterior transfer fitting strategy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72** 673–694.
- [5] EHRGOTT, M. (2000). *Multicriteria Optimization. Lecture notes in economics and mathematical systems*. Springer.
- [6] GASSER, T. and KNEIP, A. (1995). Searching for Structure in Curve Sample. *Journal of the American Statistical Association* **90** pp. 1179–1188.
- [7] GERVINI, D. and GASSER, T. (2005). Nonparametric maximum likelihood estimation of the structural mean of a sample of curves. *Biometrika* **92** 801–820.
- [8] HÄRDLE, W. K. and SIMAR, L. (2012). *Applied Multivariate Statistical Analysis*. Springer.
- [9] KAUFMAN, L. and ROUSSEEUW, P. (1987). *Clustering by Means of Medoids. Reports of the Faculty of Mathematics and Informatics*. Faculty of Mathematics and Informatics.
- [10] KNEIP, A. and GASSER, T. (1992). Statistical Tools to Analyze Data Representing a Sample of Curves. *The Annals of Statistics* **20** 1266–1305.
- [11] KNEIP, A. and RAMSAY, J. O. (2008). Combining Registration and Fitting for Functional Models. *Journal of the American Statistical Association* **103** 1155–1165.
- [12] KNEIP, A., LI, X., MACGIBBON, K. B. and RAMSAY, J. O. (2000). Curve registration by local regression. *Canadian Journal of Statistics* **28** 19–29.
- [13] MASON, J. C., RODRIGUEZ, G. and SEATZU, S. (1993). Orthogonal splines based on B-splines with applications to least squares, smoothing and regularisation problems. *Numerical Algorithms* **5** 25–40.
- [14] POSS, D. and WAGNER, H. (2014). Analysis of Juggling Data: Registerung Data to Principal Components to explain Amplitude Variation. *Electronic Journal of Statistic* **8** 1825–1834.
- [15] POWELL, M. J. D. (2006). The NEWUOA software for unconstrained optimization without derivatives. In *Large-Scale Nonlinear Optimization*, (G. Di Pillo and M. Roma, eds.). *Nonconvex Optimization and Its Applications* **83** 255–297. Springer US.
- [16] RAMSAY, J. O. (1988). Monotone Regression Splines in Action. *Statistical Science* **3** 425–441.
- [17] RAMSAY, J. O. (1998). Estimating smooth monotone functions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60** 365–375.

- [18] RAMSAY, J. O. and LI, X. (1998). Curve registration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60** 351–363.
- [19] RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. *Springer Series in Statistics*. Springer.
- [20] SAKOE, H. and CHIBA, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on* **26** 43–49.
- [21] SANGALLI, L. M., SECCHI, P., VANTINI, S. and VENEZIANI, A. (2009). A Case Study in Exploratory Functional Data Analysis: Geometrical Features of the Internal Carotid Artery. *Journal of the American Statistical Association* **104** 37-48.
- [22] SANGALLI, L. M., SECCHI, P., VANTINI, S. and VITELLI, V. (2010). k-mean alignment for curve clustering. *Computational Statistics & Data Analysis* **54** 1219-1233.
- [23] SILVERMAN, B. W. (1995). Incorporating Parametric Effects into Functional Principal Components Analysis. *Journal of the Royal Statistical Society* **57** 673–689.
- [24] SLAETS, L., CLAESKENS, G. and HUBERT, M. (2012). Phase and amplitude-based clustering for functional data. *Computational Statistics & Data Analysis* **56** 2360 - 2374.
- [25] SPELLMAN, P. T., SHERLOCK, G., ZHANG, M. Q., IYER, V. R., ANDERS, K., EISEN, M. B., BROWN, P. O., BOTSTEIN, D. and FUTCHER, B. (1998). Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization. *Mol. Biol. of the Cell* **9** 3273–3297.
- [26] SRIVASTAVA, A., WU, W., KURTEK, S., KLASSEN, E. and MARRON, J. S. (2011). Registration of Functional Data Using Fisher-Rao Metric.
- [27] WANG, K. and GASSER, T. (1997). Alignment of Curves by Dynamic Time Warping. *Annals of Statistics* **25** 1251-1276.
- [28] WANG, K. and GASSER, T. (1998). Asymptotic and bootstrap confidence bounds for the structural average of curves. *The Annals of Statistics* **26** 972–991.
- [29] WANG, K. and GASSER, T. (1999). Synchronizing sample curves nonparametrically. *The Annals of Statistics* **27** 439–460.
- [30] ZHAO, X., MARRON, J. S. and WELLS, M. T. (2004). The Functional Data Analysis View of Longitudinal Data. *Statistica Sinica* **14** 789-808.