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

Lars Lau Rakêt[1], Stefan Sommer[1], Bo Markussen[2]
[1]Department of Computer Science
[2]Department of Mathematical Sciences
University of Copenhagen, Denmark

December 10, 2013

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 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.

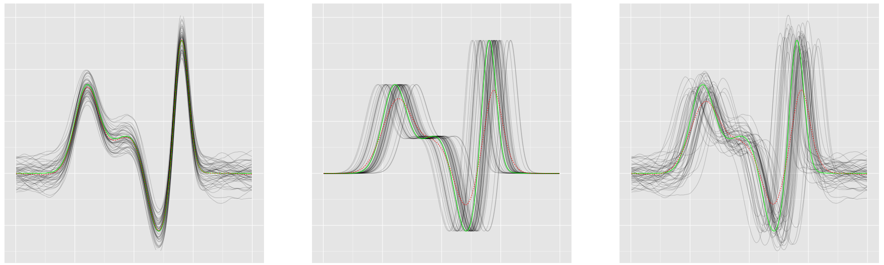
The mixed effects are assumed to be observations of Gaussian processes, and the resulting calculations are carried out by iteratively linearizing the model and estimating parameters from the resulting likelihood function. In addition to allowing estimation of the optimal combination of smoothing and 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.



(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.

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 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 labor curve data, where amplitude variation is modeled parametrically, with random additive and multiplicative effects.

Another application 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

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

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

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

where $\boldsymbol{r} = \boldsymbol{y} - \hat{\boldsymbol{\theta}}^{\boldsymbol{w}^0} - Z(\mathrm{E}\left[\,\boldsymbol{w} \,|\, \boldsymbol{y}\,\right] - \boldsymbol{w}^0) - \mathrm{E}\left[\,\boldsymbol{x} \,|\, \boldsymbol{y}\,\right]$. 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)

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

and

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

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)

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

where

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

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 $\mathrm{E}\left[\,\boldsymbol{x} \,|\, \boldsymbol{w}, \boldsymbol{y}\,\right]$ 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.

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
```

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 procedure 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

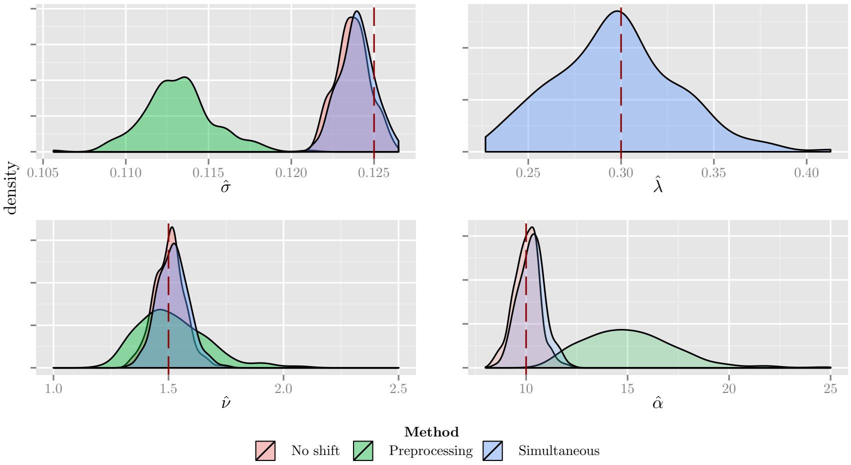


Figure 2: Density plots of variance parameter estimates from 200 independent realizations of the model (10). Seven outliers have been removed in the bottom left plot (4 *Simultaneous*, 3 *Preprocessing*)

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 $\sigma^2\mathcal{S}$

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

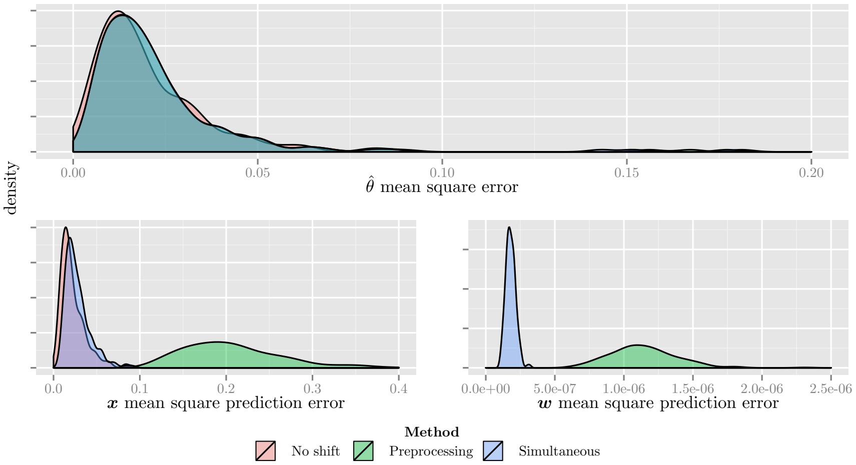


Figure 3: Density plots of mean square errors of $\hat{\theta}$ (top) and predictions of the serially correlated effects x (buttom left) and the warping parameters w (buttom right) from 200 independent realizations of the model (10). Ten outliers have been removed in the buttom left plot (4 *Simultaneous*,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

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 correlated 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.

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

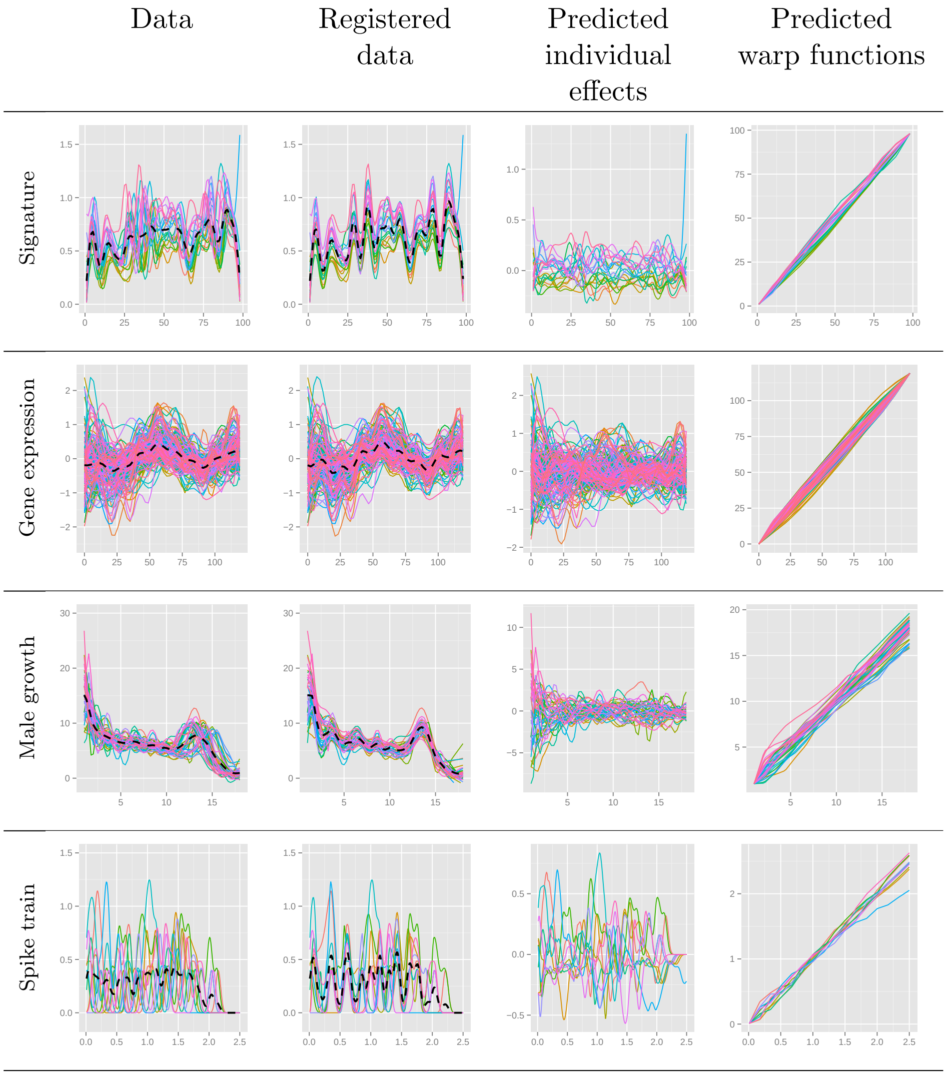


Figure 4: Results of Algorithm 1 on four datasets. Black dashed curves show the mean curve.

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 $t'_1, \dots, t'_{n_w}$, i.e. the covariance matrix C_0 of $\boldsymbol{w}_i = (w_{i1}, \dots, w_{in_w})$ 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.

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

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	$1.96 \cdot 10^{-4}$	230	$4.33 \cdot 10^5$	1.65	0.19	0.59	1.07	0.26
Gene expression	$2.03 \cdot 10^{-4}$	282	$2.12 \cdot 10^5$	2.98	0.05	0.94	1.19	0.81
Male growth	$1.41 \cdot 10^{-4}$	751	$2.47 \cdot 10^5$	2.86	0.35	0.77	1.11	0.42
Spike train	$1.67 \cdot 10^{-4}$	536	$1.04 \cdot 10^5$	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.

Acknowledgement

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.