Bayesian Analysis (2017) **12**, Number 2, pp. 557–582

Variational Bayes for Functional Data Registration, Smoothing, and Prediction

Cecilia Earls* and Giles Hooker†

Abstract. We propose a model for functional data registration that extends current inferential capabilities for unregistered data by providing a flexible probabilistic framework that 1) allows for functional prediction in the context of registration and 2) can be adapted to include smoothing and registration in one model. The proposed inferential framework is a Bayesian hierarchical model where the registered functions are modeled as Gaussian processes. To address the computational demands of inference in high-dimensional Bayesian models, we propose an adapted form of the variational Bayes algorithm for approximate inference that performs similarly to Markov Chain Monte Carlo (MCMC) sampling methods for well-defined problems. The efficiency of the adapted variational Bayes (AVB) algorithm allows variability in a predicted registered, warping, and unregistered function to be depicted separately via bootstrapping. Temperature data related to the El-Niño phenomenon is used to demonstrate the unique inferential capabilities for prediction provided by this model.

Keywords: Bayesian modeling, functional data, functional prediction, registration, smoothing, variational Bayes.

1 Introduction

This paper introduces a novel approach to functional data registration within a Bayesian hierarchical model. In this model, both registered and warping functions are modeled in terms of Gaussian processes with registration described by restrictions on the covariance of the registered functions. This enables both Markov Chain Monte Carlo (MCMC) and variational Bayes methods to be employed for estimation and prediction. Our model for registration extends current inferential procedures to include both 1) functional prediction in the context of registration and 2) registration and smoothing in one model.

The primary advantage to our proposed registration model in comparison to current registration methods is that it provides a probabilistic framework in which new observations can be considered. Assuming a new unregistered function has been partially recorded, with this framework, we can obtain estimates of the registered partial function and also the corresponding partial warping function. Using these estimates, the complete registered function, the complete warping function, and the complete unregistered function can be predicted. Details of the prediction model can be found in Section 5. To the authors' knowledge this is the first time functional prediction is considered in the context of registration.

*Department of Biological Statistics and Computational Biology, 1198 Comstock Hall, Ithaca, NY 14853, cae79@cornell.edu

†Department of Biological Statistics and Computational Biology, 1198 Comstock Hall, Ithaca, NY 14853, gjh27@cornell.edu

 DOI: 10.1214/16-BA1013

Additionally, our model can be extended to allow for noisy observations. Most current registration methods consider functional regularization as a pre-processing step with the exception of Raket et al. (2014). However, in the paper by Raket et al. (2014), the authors' maximum likelihood approach to registering latent functions does not provide an easy way to quantify variability in the estimates of the registered functions. In Appendix C.3 (Earls and Hooker, 2016), we provide an illustration of how smoothing functions in a pre-processing step can significantly underestimate the variability in the estimates of the registered functions.

The simplest definition of functional data registration is any algorithm that aligns functions in a way that eliminates all phase variability between functions (Ramsay and Silverman (2005)). Without registration, basic summary statistics such as the sample mean and covariance are less interpretable as time variation between significant features in the functions tends to dampen the amplitude variation in these features. Furthermore, the average timing of significant features may also be of interest and is difficult to obtain under traditional methods of analyzing functional data. There has been much recent interest in proper ways to define and measure registration as well as in developing registration methods with desirable statistical properties.

The evolution of registration dates back to Sakoe and Chiba (1978) where the authors use a dynamic programming algorithm for landmark registration. Landmark registration was again considered by Kneip and Gasser (1992, 1995). In 1997, a new cost function for functional registration was introduced by Wang and Gasser (1997). A significant advancement in registration literature can be traced to Silverman (1995) and Ramsay and Li (1998) where the authors introduce global registration procedures, and Ramsey considers the use of a flexible family of monotone warping functions. Parametric and B-spline base warping functions are considered by Brumback and Lindstrom (2004) and Gervini and Gasser (2004), respectively. Nonparametric maximum likelihood approaches to registration are considered by both Ronn (2001) and Gervini and Gasser (2005). A moments based approach to registration is introduced by James (2007). Tang and Muller (2008) propose pairwise curve synchronization. The first Bayesian approach to registration can be found in Telesca and Inoue (2007). Registration to principal components is considered by Kneip and Ramsay (2008). Finally, with regard to improvements in registration, the recent work by Srivastava et al. (2011) offers the most comprehensive framework for registration to date.

Much of the focus in combining registration with other types of inference in one model has been in the area of functional data clustering and registration. Current work in this area can be found in Liu and Yang (2009), Sangalli et al. (2010), and also a Bayesian approach in Zhang and Telesca (2014). Recent work by Raket et al. (2014) includes a model for functional smoothing and registration. While these extensions to registration procedures offer additional tools for functional data analysis, they tend to focus less on high-quality registration.

In this paper, we develop Bayesian hierarchical models that address both areas of development in registration procedures. First, a model is proposed to register functional data that gives estimates that compare favorably with those from the best current registration methods available, notably, Srivastava et al. (2011). Then, we demonstrate

how this model can be extended to incorporate other inferential procedures. The two examples provided in this paper are extensions for both a functional prediction model and a model for simultaneous registration and smoothing.

This paper also addresses the computational issues associated with high-dimensional Bayesian hierarchical models. To this end, we propose an alternative algorithm to variational Bayes approximation that can be used for models in which the full conditional distributions of a subset of the parameters are not from a known parametric family. To distinguish our algorithm from pure variational Bayes, in this paper we will refer to this approximate inference procedure as Adapted Variational Bayes (AVB).

This paper is organized as follows. Section 2 presents our basic registration model. The Adapted Variational Bayes algorithm is discussed in detail in Section 3. A comparison of results from our model to current methods can be found in Section 4.1. Additionally, a comparison of results obtained using AVB and those given by MCMC can be found in 4.2. The prediction model is presented in Section 5. In Section 5.4, the prediction model is used to forecast the future trajectory of sea-surface temperatures that are associated with the El-Niño phenomenon. An adaptation to our model that allows for noisy data is found in Appendix C (Earls and Hooker, 2016). Finally, a discussion is found in Section 6.

2 Gaussian Process Models for Registration

The functional registration models proposed in this paper are foremost designed to extend and improve on the minimum eigenvalue registration criterion for continuous registration first introduced by Ramsay and Li (1998). Accordingly, we will consider two functions perfectly registered if the variation between the two functions can be described entirely in terms of one functional direction – the *target function*. Our method of registration improves on Ramsay and Li's Procrustes method, Ramsay and Li (1998), by implicitly accounting for vertical shifts between registered functions and by allowing the target curve to evolve throughout the registration procedure. In Section 4, we will demonstrate how using the minimum eigenvalue criterion under these conditions provides a more complete curve registration. Our results are comparable to those of Srivastava et al. (2011).

The theoretical basis for modeling functional data as Gaussian processes in a hierarchical Bayesian environment is established in Earls and Hooker (2014). In our registration model, each registered function, $X_i(h_i(t))$, $i = 1, \ldots, N$, is the composition of an observed unregistered function, $X_i(t)$, with an unknown warping function, $h_i(t)$, over some fixed time domain $\mathcal{T} = [t_1, t_p]$. The function $h_i(t)$ is represented as $h_i(t) = t_1 + \int_{t_1}^{t} \exp(w_i(s))ds$ to enforce its monotonicity where $w_i(t)$ is specified as having a Gaussian process distribution, resulting in a functional random effect. In this paper, we will refer to $w_i(t)$ as the *base function* associated with warping function $h_i(t)$. The base functions are non-parametrically specified for optimal registration. We, however, impose the following restrictions on the warping functions:

1. $h(t_1) = t_1$,
2. $h(t_p) = t_p$, and
3. if $t_k > t_j$, then $h(t_k) > h(t_j)$ for all $t_k, t_j \in \mathcal{T}$.

Restrictions (1) and (3) are built into the definition of $h_i(t)$. Restriction (2) is imposed through the characteristic function in the expression for the prior defined for each base function, $w_i(t)$. Note that $w_i(t) = 0$ corresponds to the identity warping, $h_i(t) = t$. An important feature of our definition of the warping functions is that it defines an identifiable relationship between $h_i(t_j)$ and $w_i(t)$ which is necessary for predicting future outcomes of curves that are only partially observed. In Section 5 is a more thorough discussion of the prediction model.

We also model each registered function, $X_i(h_i(t))$, $i = 1 \dots N$, as a Gaussian process such that

$$X_i(h_i(t)) \mid z_{0i}, z_{1i}, f(t) \sim GP(z_{0i} + z_{1i} f(t), \gamma_R^{-1}\Sigma(s,t)), \quad s,t \in \mathcal{T}. \tag{1}$$

Here, $f(t)$ is the *target function*. The target function serves as the primary functional direction in which the registered functions vary. Accordingly, the above covariance function, $\gamma_R^{-1}\Sigma(s,t)$, is defined to penalize all variation in the registered functions beyond a scaling and vertical shifting of the target function (the function-specific mean). In these models we will define γ_R as a registration parameter that determines the severity of this penalty.

Given a sample of unregistered functions, $X_i(t)$, $i = 1 \dots N$, defined over the interval $\mathcal{T} = [t_1, t_p]$, we are interested in estimating the warping functions, $h_i(t)$, the shifting and scaling parameters, z_{0i} and z_{1i}, the target curve, $f(t)$, and the registered functions.

For now, we will assume the functions are recorded without noise. If the functions are recorded with noise, it is common practice in the current literature to first perform a pre-processing smoothing step. An undesirable result of this pre-processing step is that the subsequent inference procedure is unable to capture the extra variability associated with the smoothing process. This phenomenon is illustrated using the Berkeley boys growth velocity data in Appendix C.3 (Earls and Hooker, 2016). Appendices C.1 and C.2 (Earls and Hooker, 2016) detail how our basic registration model can be modified to both smooth and register functions.

Inference is accomplished through a Bayesian hierarchical model. The distributional assumptions and prior specifications for this model are

$$X_i(h_i(t)) \mid z_{0i}, z_{1i}, f(t) \sim GP(z_{0i} + z_{1i} f(t), \gamma_R^{-1}\Sigma(s,t)), \quad s,t \in \mathcal{T}, \quad i = 1, \dots, N, \tag{2}$$

$$\Sigma(s,t) = P_1(s,t) + P_2(s,t), \tag{3}$$

$$h_i(t) = t_1 + \int_{t_1}^{t} \exp(w_i(s))ds, \quad t \in \mathcal{T}, \quad i = 1, \dots, N,$$

$$w_i(t) \propto GP(0, \gamma_w^{-1}\Sigma(s,t) + \lambda_w^{-1} P_w(s,t))\mathbf{1}\{t_1 + \int_{t_1}^{t_p} \exp(w_i(s))ds = t_p\},$$

$$s,t \in \mathcal{T}, \quad i = 1,\ldots,N, \tag{4}$$

$$P_w(s,t) = P_2(s,t), \tag{5}$$

$$z_{0i} \mid \sigma_{z0}^2 \sim N(0,\sigma_{z0}^2), \quad i = 1,\ldots,(N-1), \quad z_{0N} = -\sum_{i=1}^{N-1} z_{0i},$$

$$\sigma_{z0}^2 \sim IG(a,b),$$

$$z_{1i} \mid \sigma_{z1}^2 \sim N(1,\sigma_{z1}^2), \quad i = 1,\ldots,N, \tag{6}$$

$$\sigma_{z1}^2 \sim IG(a,b), \tag{7}$$

$$f(t) \mid \eta_f, \lambda_f \sim GP(0,\Sigma_f(s,t)), \quad s,t \in \mathcal{T}, \tag{8}$$

$$\Sigma_f(s,t) = \eta_f^{-1} P_1(s,t) + \lambda_f^{-1} P_2(s,t), \tag{9}$$

$$\eta_f \sim G(c,d), \text{ and}$$

$$\lambda_f \sim G(c,d).$$

For this model, the parameters, $z_{0i}, i = 1,\ldots,N$, allow the registered functions to vary by vertical shifts from a scaling of the target function, $f(t)$. The constraint, $z_{0N} = -\sum_{i=1}^{N-1} z_{0i}$, ensures the average vertical shift is estimated to be 0. The parameters, $z_{1i}, i = 1,\ldots,N$, quantify amplitude variation in the registered functions. Note, the Gaussian distribution on $\mathbf{z}_1 = (z_{11}\ldots z_{1N})'$ can be replaced by a Dirichlet distribution on $\mathbf{z}_1/N$. The result is a slightly more complicated model that has the nice effect of scaling the target function to the empirical mean of the estimated registered functions. Priors (6) and (7) would then be omitted.

In the above model specifications, all covariance functions are composed of a linear combinations of two bi-variate functions, $P_1(s,t)$ and $P_2(s,t)$. $P_1(s,t)$ penalizes variation in constant and linear functions and $P_2(s,t)$ penalizes function variability in all other directions. Together they define a proper covariance function. For each covariance function above, the specification of the registration and smoothing parameters indicate the extent the two different types of variability should be penalized for each function. For example, for both the registered functions and the base functions, we want to penalize variation in *any* direction other than that of the mean function. The covariance specifications of $\gamma_R^{-1}\Sigma(s,t)$ and $\gamma_w^{-1}\Sigma(s,t)$ reflect these penalties, where the magnitude of the penalty is controlled by registration parameters, γ_R and γ_w, (distributional assumptions 2,3, and 4). We can use $P_2(s,t)$ to penalize roughness in a given function. Here we would like both the target function and the base functions to be smooth. This is achieved by the inclusion of $\lambda_f^{-1}P_2(s,t)$ and $\lambda_w^{-1}P_2(s,t)$ in the priors for these functions (distributional assumptions 8, 9, 4, and 5) where the level of the penalty is controlled by the smoothing parameters λ_f and λ_w. For the exact definitions of $P_1(s,t)$ and $P_2(s,t)$, see Earls and Hooker (2014).

Given the above model, in practice we will proceed by using finite approximations to each function and functional distribution. In Earls and Hooker (2014) we establish some theoretical properties of these types of approximations. The following finite-dimensional distributions are used in the final model in lieu of their infinite dimensional counterparts

above:

$$\mathbf{X}_i(\mathbf{h}_i) \mid z_{0i}, z_{1i}, \mathbf{f} \sim N_p(z_{0i}\mathbf{1} + z_{1i}\mathbf{f}, \gamma_R^{-1}\boldsymbol{\Sigma}), \quad i = 1 \dots N, \tag{10}$$

$$\mathbf{h}_i(t_j) = t_1 + \sum_{k=2}^{j}(t_k - t_{k-1})\exp(w_i(t_{k-1})), \quad i = 1 \dots N, \quad j = 1 \dots p,$$

$$\mathbf{w}_i \propto N_{p-1}(\mathbf{0}, \gamma_w^{-1}\boldsymbol{\Sigma} + \lambda_w^{-1}\mathbf{P}_w)\mathbf{1}\{t_1 + \sum_{k=2}^{p}(t_k - t_{k-1})\exp(w_i(t_{k-1})) = t_p\},$$

$$i = 1 \dots N, \tag{11}$$

$$\mathbf{f} \mid \eta_f, \lambda_f \sim N_p(0, \boldsymbol{\Sigma}_f), \text{ and}$$

$$\boldsymbol{\Sigma}_f = \eta_f^{-1}\mathbf{P}_1 + \lambda_f^{-1}\mathbf{P}_2.$$

Section 4 provides several examples that illustrate how allowing the target function to be estimated within the model results in a more complete functional registration in comparison to the Procrustes method, Ramsay and Li (1998). However, the Gaussian process model does not constrain the timing of a feature in the target function to occur at the average time of the corresponding unregistered features. Although the model for the $w_i(t)$ is centered on zero, it is still possible that the average of the estimated warped time points, $\overline{h_\cdot(t_1)}, \dots, \overline{h_\cdot(t_p)}$, does not correspond to the original time points. Shifting these by an additional registration so that the warped times average to the original time does not affect our prediction model, but it will then allow an explicit comparison of $h_i(t_j)$ to t_j to tell us whether the process is running ahead or behind "standard" time. Srivastava et al. (2011) use a similar "correction" to determine their target function. Details on how to perform this final registration can be found in Appendix A.3 (Earls and Hooker, 2016).

2.1 Registration and Warping Parameter Selection

Registration in this model is controlled by three parameters, γ_R, γ_w, and λ_w. The parameter γ_R determines the extent the registered functions will be penalized for varying from a shifting and scaling of the target function. This penalty for lack of registration is tempered by penalties for roughness in the warping functions. The parameter γ_w determines how far the warping functions can deviate from the identity warping while λ_w controls the smoothness of the warping functions. This model can also be adapted to allow for function specific warping penalties. In Section 5.4, we will give an example where function specific penalties for the base functions have been utilized to preserve significant covariance relationships in the estimated registered functions.

For a given statistical analysis, the registration parameters are chosen by the user. This is because variation in registered functions outside of shifts and rescaling $f(t)$ (controlled by γ_R) is nearly non-identifiable from variation in warping functions, as regulated by γ_w and λ_w. That is, there are linear directions of variation in the $w(t)$ that result in nearly linear directions of variation in the resulting $X_i(t)$ (a simplified example is the identity $\sin(t+\delta) = \cos(\delta)\sin(t) + \sin(\delta)\cos(t)$ which confounds horizontal

shifts with vertical variation in periodic functions). An exact description of this form of confounding is beyond the scope of this paper, but it yields unstable estimates when these parameters are not fixed and we therefore choose these by hand. In this model a large registration penalty, γ_R, in comparison to the penalty on warping, γ_w, will result in registered functions that no longer retain significant features in the data. Alternatively, a registration parameter that is too small will not properly align features. Desirable values of these parameters can be determined using short runs of the adapted variational Bayes algorithm described in Section 3.1. In practice, we have found these penalties should be adjusted by powers of ten to see a significant change in estimates of the registered functions. Once determined, γ_R, γ_w, and λ_w are fixed and can be used with the adapted variational Bayes estimates to initialize an MCMC sampler.

3 Variational Approximation

3.1 Variational Bayes

For our registration model, it is appropriate to use Markov Chain Monte Carlo (MCMC) methods to sample from the joint posterior distribution of all unknown parameters. However, for most applications, the dimensionality of the parameter space will require exceptionally long chains that are impractical and expensive to obtain. Here, we suggest a variational Bayes alternative to MCMC sampling to at the very least obtain good starting values for a MCMC sampler. Alternatively, we will show in Section 4.2 that differences in the estimated parameters obtained through adapted variational Bayes and MCMC sampling tend to be small, and estimation via adapted variational Bayes alone is likely sufficient for many inferential procedures.

The variational Bayes procedure described here is based on the variational methods proposed by Omerod and Wand (2010) and Bishop (2006). Their proposed method optimizes a lower bound of the marginal likelihood which results in finding an approximate joint posterior density that has the smallest Kullback–Leibler (KL) distance from the true joint posterior density. Both fixed form and nonparametric forms of variational Bayes algorithms are currently available. The variational Bayes algorithm that we propose is most closely related to fixed form variational Bayes. A clear explanation of fixed form variational Bayes can be found in Goldsmith et al. (2011) where the authors utilize variational Bayes for a functional regression model.

Suppose $q(\boldsymbol{\theta})$ is the approximated posterior joint distribution. The fixed form variational Bayes algorithm assumes for some partition of $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d\}$, $q(\boldsymbol{\theta}) = \prod_{k=1}^{d} q_k(\boldsymbol{\theta}_k)$, where each distribution q_k is of a known parametric form. Traditionally this requirement is satisfied through the use of conditionally conjugate priors.

In our model, the Gaussian process priors for the base functions, $w_i(t)$, $i = 1, \ldots, N$, are not conditionally conjugate to the likelihood function. Therefore, the fixed form variational Bayes optimization method does not apply directly since $q_k(\mathbf{w}_i)$, $i = 1, \ldots, N$ are not known parametric distributions.

3.2 Adapted Variational Bayes (AVB)

Suppose we order the parameter vector, $\boldsymbol{\theta}$, so that, $\boldsymbol{\theta} = \{\mathbf{w}_1, \ldots, \mathbf{w}_N, \boldsymbol{\theta}_{N+1}, \ldots, \boldsymbol{\theta}_d\}$, for $k = \{(N+1), \ldots, d\}$. While the q distributions on the approximated base functions, $\mathbf{w}_i$, $i = 1, \ldots, N$ are not of known parametric forms, each $q_k(\boldsymbol{\theta}_k)$, $k = \{(N+1), \ldots, d\}$, is a known parametric distribution that can be estimated using the standard fixed form variational Bayes algorithm. The following variational Bayes algorithm is adapted to include estimation for parameters without conditionally conjugate priors in addition to all other parameters that typically can be estimated using fixed form variational Bayes. This adaptation is similar to the variational approximation to the EM algorithm described by Tzikas et al. (2008) which performs approximate inference based on the EM algorithm for models where the posterior distribution of the latent variables is of an unknown form.

The Adapted Variational Bayes Algorithm

Define $f(\mathbf{X}, \mathbf{w}, \boldsymbol{\theta})$ as the joint distribution of the data, $\mathbf{X}$, and parameters of a Bayesian hierarchical model. Suppose an approximate joint posterior distribution of the parameters, $\boldsymbol{\theta} = \{\mathbf{w}_1, \ldots, \mathbf{w}_N, \boldsymbol{\theta}_{N+1}, \ldots, \boldsymbol{\theta}_d\}$, is of the form $\prod_{j=1}^{N} \prod_{k=N+1}^{d} q_j(\mathbf{w_j}) q_k(\boldsymbol{\theta_k})$ where each $q_j(\mathbf{w_j}), j = 1, \ldots, N$, is known only up to a constant of proportionality and the distributions, $q_k(\boldsymbol{\theta}_k), k = N+1, \ldots, d$, are of known parametric forms. We will define the following estimation procedure for $\boldsymbol{\theta}$ as the adapted variational Bayes algorithm.

1. Initialize $\boldsymbol{\theta}$.
2. For each iteration, m, and each k, $k = 1, \ldots, N$, update the estimate for $\mathbf{w}_k$ so that $\mathbf{w}_k^{(m)} = \sup_{\mathbf{w}_k} q_k(\mathbf{w}_k \mid \boldsymbol{\theta}_j^{(m-1)}, j = (N+1), \ldots, d)$. This is equivalent to setting $\mathbf{w}^{(m)} = \sup_{\mathbf{w}} f(\mathbf{X}, \mathbf{w} \mid \boldsymbol{\theta}_j^{(m-1)}, j = (N+1), \ldots, d)$.
3. For each iteration, m, and each k, $k = (N+1), \ldots, d$, update q_k so that $q_k^{(m)} \propto \exp[E_{(\boldsymbol{\theta}_{-k})}(\log f(\boldsymbol{\theta}_k \mid rest)]$, where the expectation is taken with respect to the distributions $q_j^{(m-1)}(\boldsymbol{\theta}_j)$, $j = 1, \ldots, d$, $j \neq k$.
4. Repeat steps (2) and (3) until the desired convergence criterion is met.

Here the notation, $E_{(\theta_{-k})}$, denotes the expected value over all parameters except θ_k. In the next section, we will drop the subscript k, and $E_{(\theta_{-\theta_k})}$ will represent the expectation over all parameters except for θ_k (e.g. $E_{(\theta_{-\eta_f})}$ will represent the expectation taken over all parameters except for η_f).

Theorem. *The adapted variational Bayes algorithm converges to estimated parameters, $\hat{\boldsymbol{\theta}}$, that minimize the Kullback–Leibler distance between the approximate posterior distribution, $q(\boldsymbol{\theta}_{-\mathbf{w}})$, and the posterior distribution, $f(\boldsymbol{\theta}_{-\mathbf{w}} \mid \mathbf{X}, \mathbf{w})$, for a local optimization of $f(\mathbf{X}, \mathbf{w} \mid \boldsymbol{\theta}_{-\mathbf{w}})$ in $\mathbf{w}$.*

Proof. Assume $\mathbf{w}$ is known. Goldsmith et al. (2011) demonstrate that minimizing the K-L distance between $q(\boldsymbol{\theta}_{-\mathbf{w}})$ and $f(\boldsymbol{\theta}_{-\mathbf{w}} \mid \mathbf{X}, \mathbf{w})$ is equivalent to maximizing the

following log of a q-specific lower bound of the joint marginal distribution of $\mathbf{X}$ and $\mathbf{w}$ in q.

$$\begin{aligned}\log(\mathbf{X}, \mathbf{w}; q) &= \int q(\boldsymbol{\theta}_{-\mathbf{w}}) \log\left\{\frac{f(\mathbf{X}, \mathbf{w}, \boldsymbol{\theta}_{-\mathbf{w}})}{q(\boldsymbol{\theta}_{-\mathbf{w}})}\right\} d\boldsymbol{\theta}_{-\mathbf{w}} \\ &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[\log[f(\mathbf{X}, \mathbf{w}, \boldsymbol{\theta}_{-\mathbf{w}})] - \log[q(\boldsymbol{\theta}_{-\mathbf{w}})]]\end{aligned}$$

The adapted variational Bayes algorithm alternates between: 1) maximizing $f(\mathbf{X}, \mathbf{w} \mid \boldsymbol{\theta}_{-\mathbf{w}})$ in $\mathbf{w}$ (possibly locally), and 2) fixing $\mathbf{w}$ at the value determined by the previous step and using traditional variational Bayes to maximize $f(\mathbf{X}, \mathbf{w}; q)$. Here we demonstrate this process results in a monotonic increasing sequence in $\log f(\mathbf{X}, \mathbf{w}; q)$ which guarantees the convergence of this algorithm.

For each iteration, m of our adapted variational Bayes algorithm,

$$\begin{aligned}\log f(\mathbf{X}, \mathbf{w}^{(m)}; q^{(m)}) &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[\log[f(\mathbf{X}, \mathbf{w}^{(m)}, \boldsymbol{\theta}_{-\mathbf{w}}^{(m)})] - \log[q(\boldsymbol{\theta}_{-\mathbf{w}}^{(m)})]] \\ &\leq E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[\log[f(\mathbf{X}, \mathbf{w}^{(m+1)}, \boldsymbol{\theta}_{-\mathbf{w}}^{(m)})] - \log[q(\boldsymbol{\theta}_{-\mathbf{w}}^{(m)})]] \quad (12) \\ &\leq E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[\log[f(\mathbf{X}, \mathbf{w}^{(m+1)}, \boldsymbol{\theta}_{-\mathbf{w}}^{(m+1)})] - \log[q(\boldsymbol{\theta}_{-\mathbf{w}}^{(m+1)})]] \quad (13) \\ &= \log f(\mathbf{X}, \mathbf{w}^{(m+1)}; q^{(m+1)}).\end{aligned}$$

The inequality in (12) is guaranteed by step 2 of the adapted variational Bayes algorithm, and the inequality in (13) is the result of using the traditional variational Bayes algorithm with $\mathbf{w}$ considered known (step 3). □

The lower bound of the marginal distribution of $\mathbf{X}$ and $\mathbf{w}$ can be monitored until changes in this function are under some threshold. The specific form of this function can be found in Appendix B.2 (Earls and Hooker, 2016). However, as the algorithm is guaranteed to converge, it is in practice more prudent to instead monitor changes in the parameter estimates from iteration to iteration and stop the algorithm when these changes are below a specified threshold.

Convergence of the AVB algorithm is guaranteed. However, convergence to a global maximum is not guaranteed. In the maximization step of the AVB algorithm, a function proportional to the approximate posterior for the base functions is maximized in the base functions. This function can be multimodal and occasionally the estimated base functions reflect a local maximum of the approximated posterior. To circumvent this problem, in practice it is sometimes necessary to adjust the registration and warping penalties as the functions become registered. An unregistered function that requires a substantial amount of warping can cause convergence to a local maximum due to the small penalty on warping. The flexibility in warping allowed by this small penalty can cause the function to deform rather than register. This can be remedied in two ways. The first option is to perform a simple initial warping for this function that prevents the optimization from falling into a local mode. The second option is to adjust the registration and warping parameters over time. Initially a stronger warping penalty is employed to prevent function deformation. Then, as the functions register, the warping

Method	Sim	BGV	RBGV
ME	176	1636	NA
F-R	1.3	2.9	NA
AVB	1820	36540	28692
MCMC	222	271296	46037

Table 1: Computational Time (seconds).

penalty can be reduced to allow for a more complete registration. When initializing an MCMC sampler, the final penalties on warping and registration from the adapted variational Bayes algorithm should be used.

While AVB estimates are often close to MCMC estimates, often it will still be desirable to characterize the posterior distributions of all parameters through MCMC sampling. This is particularly prudent when noise is present in the observed unregistered functions which introduces a significant amount of variability in the posterior samples (see Appendix C.3, Earls and Hooker (2016)). We can express the computational savings due to using the AVB algorithm as the amount of time saved in burn-in iterations. When applying the adapted variational Bayes algorithm to "real" data such as the *Berkeley Boys Growth Velocity Data*, we have found the adapted variational Bayes algorithm saves a significant amount of computational time in the burn-in period. For simulated data, there may not be any savings in computational time due to the following: 1) MCMC samples move very quickly towards the optimal solution when the registration problem is perfectly defined and 2) the maximization step of the AVB algorithm is inefficient, especially in the first few iterations. In Table 1, we compare the computational cost of obtaining estimates through MCMC sampling, AVB, the F-R algorithm, Srivastava et al. (2011), and the ME algorithm, Ramsay and Li (1998), for simulated data, the original *Berkeley Boys Growth Velocity Data* (BGV), and the boys growth velocity data corrupted with noise (RBGV). Full descriptions of these datasets and all proposed registration methods are described in Sections 4 and Appendix C.3 (Earls and Hooker, 2016). Since the ME, F-R and AVB algorithms do not include variability estimates, for this comparison the MCMC sampler is run long enough to obtain estimates but not necessarily long enough to characterize the posterior distributions. Note: since the ME and F-R algorithms treat smoothing as a pre-processing step, for the noisy growth velocity data we only compare the computational time needed to burn-in an MCMC sampler to the computational time required to obtain AVB estimates (that can then be used to initialize an MCMC sampler).

4 Comparison to Current Methods

4.1 Comparison to Other Registration Procedures

Our registration criterion minimizes all variation in the warped functions that is not in the direction of the target function (allowing for vertical shifts). In this respect, the underlying registration principle driving our model is similar to that proposed by

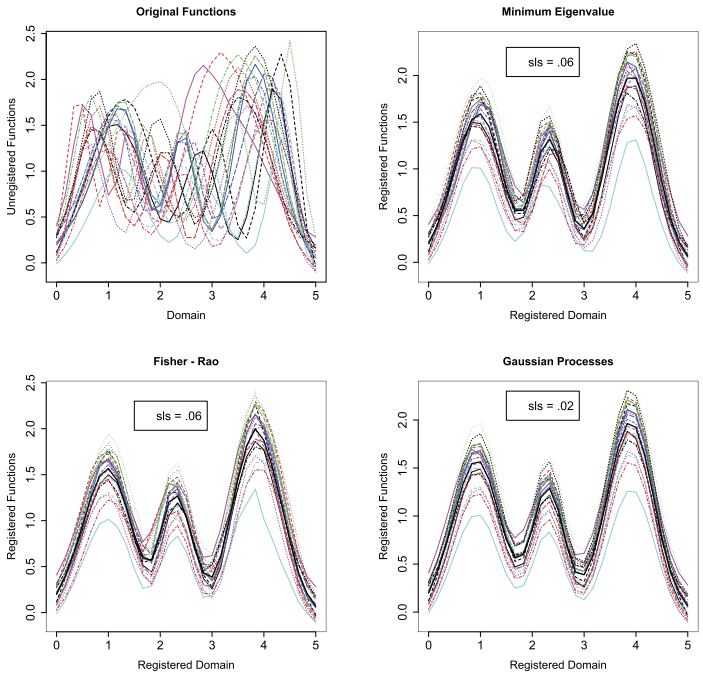

Figure 1: Simulated Data Set 2. **Top Left** Unregistered functions. **Top Right** Registered functions using the minimum eigenvalue criteria (R package 'fda'). **Lower Left** Functions registered by F-R (R package 'fdasrvf'). **Lower Right** Functions registered by the GP model.

Ramsay and Li (1998). Here we will compare our registration results to those using Ramsey's method as well as the registration procedure proposed by Srivastava et al. (2011). Srivastava et al. propose a geometric framework for functional data registration using the Fisher–Rao Riemannian metric, Rao (1945). In this paper we will refer to Ramsey and Li's registration procedure as ME (minimum eigenvalue, Ramsay and Li (1998)) and Srivastava's procedure as F-R (Fisher–Rao, Srivastava et al. (2011)), and the model proposed here as GP (Gaussian Processes). In the paper by Srivastava et al., several comparisons of registration under the F-R framework to the registration methods proposed by Gervini and Gasser (2004), James (2007), Liu and Muller (2004), and Tang and Muller (2008) are provided. In all cases, F-R appears to provide the most complete registration of the given set of functions. In light of this illustration, we will consider their method as the current frontrunner in registration procedures and use it as the standard for our comparisons.

Figures 1 and 2 contain the datasets used for this analysis. Each figure includes the original unregistered data along with plots of the functions registered using the three proposed methods. For all three registration methods, a range of parameter values were explored for optimal registration.

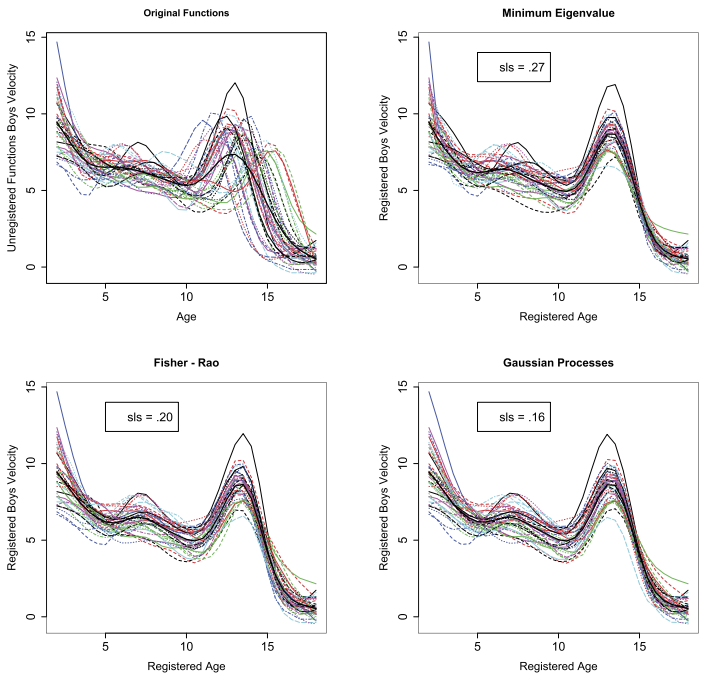

Figure 2: Registered Boys Growth Velocity. **Top Left** Original unregistered boys velocity data functions. **Top Right** Boys velocity functions registered using the minimum eigenvalue criteria (R package 'fda'). **Lower Left** Boys velocity functions registered by F-R (R package 'fdasrvf'). **Lower Right** Boys velocity functions registered by the GP model.

We have chosen to use the Sobolev Least Squares (*sls*) criterion to compare the three registration methods for each dataset as advocated by Srivastava et al. (2011). The Sobolev Least Squares criterion compares the total cross-sectional variance of the first derivatives of the registered functions to that of the original functions. Explicitly,

$$sls = \frac{\sum_{i=1}^{N} \int (X_i'(h_i(t)) - \frac{1}{N}\sum_{j=1}^{N} X_j'(h_j(t)))^2 dt}{\sum_{i=1}^{N} \int (X_i'(t) - \frac{1}{N}\sum_{j=1}^{N} X_j'(t))^2 dt}. \tag{14}$$

Lower values of *sls* correspond to better function alignment.

Simulated Data Set Figure 1 contains the functions of a simulated data set. These data consist of 20 unregistered scaled mixtures of three Gaussian probability density functions. All three registration procedures result in similar alignments. However, the GP method does a better job of recovering the original shape of the functions and results in the lowest *sls*. Note: The ME registered functions are based on 5 complete runs of the ME algorithm where in each run the previous runs results were used as the 'unregistered' functions.

Berkeley Boys Growth Velocity Data Figure 2 contains 39 velocity of growth functions for boys from the Berkeley Growth Study, Tuddenham and Snyder (1954). For this analysis, the original data are slightly changed to eliminate some erratic behavior at the beginning of each function. Here, GP and F-R yield similar registration results. However, the GP algorithm results in the lowest *sls*. ME registers the most significant peak in growth velocity but does not align lesser features as well as GP. Note: The ME registered functions are based on 2 complete runs of the ME algorithm where in each run the previous runs results were used as the 'unregistered' functions. Running this algorithm more than twice resulted in function distortion due to over-warping and a larger *sls*.

While the GP and F-R methods result in a similar alignment of functions, these results are achieved in very different environments that are specialized to satisfy specific inferential preferences. The F-R registration method is convenient (using R package 'fdasrvf') and provides fast high-quality estimates. On the other hand, while providing comparable registration results, our method expands inferential capability by providing 1) variability estimates for all unknown parameters and 2) a probability framework in which future partially observed unregistered functions are considered. In contrast to traditional functional prediction methods, our model not only provides an estimate of the complete unregistered function, but also estimates the complete warping function and the complete registered function. Details of the prediction model are found in Section 5.

4.2 Comparison to MCMC Results

To establish the utility of the adapted variational Bayes algorithm, here we compare the estimates of registered functions using adapted variational Bayes versus those obtained through MCMC sampling. For this exposition, the simulated data set and the Boys Growth Velocity data set described in Section 4.1 are used to look at the discrepancies between the estimated registered functions from MCMC sampling versus those determined by the AVB algorithm.

The squared L^2 norm of the difference between the AVB and MCMC estimate of a registered function is used to quantify the differences between these estimates. Figure 3 illustrates for the simulated data set how closely the AVB estimates follow the MCMC estimates. Even the largest squared L^2 norms of the differences between these two estimates correspond to minor changes in the estimates. These simulated data represent rather ideal conditions for registration where there is almost no variation in the registered functions beyond a scaling and vertical shift of the target function. Consequently, as we might expect, the MCMC and AVB estimation procedures are primarily in agreement. Figure 4 is a more realistic look at the differences between the MCMC and AVB registration results for data that has significant variation in the registered functions beyond a scaling and vertical shift of the target function. However, even here we see the AVB algorithm performs well. Of the 39 observations, in only 2 or 3 are there notable discrepancies between the AVB and MCMC estimated registered functions.

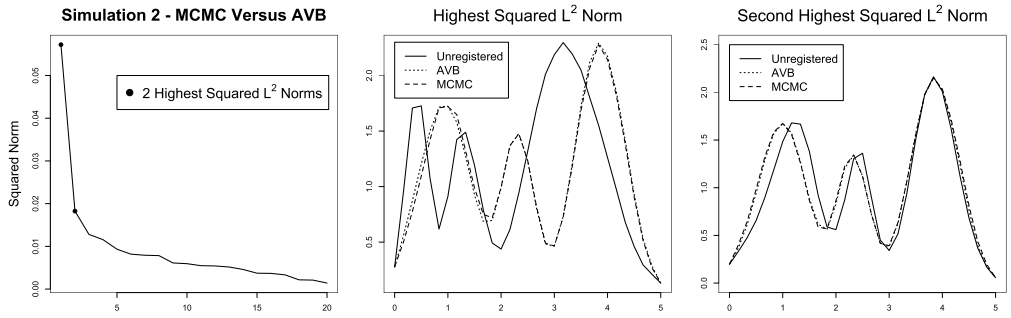

Figure 3: Simulated Data – Difference Between MCMC and AVB Estimates. **Left** Plot of the squared L^2 norm of the difference between the MCMC and AVB estimates for each observation in decreasing order of magnitude. **Center and Right** The original unregistered function plotted with the MCMC and AVB estimates of the registered functions for the observations with the two largest discrepancies between the MCMC and AVB estimates.

These examples show that while ideally AVB estimates are used to initialize a MCMC sampler, they often deviate only in minor ways from the posterior mean estimates obtained from MCMC sampling. The next logical step is to compare the variance of the approximated posterior distributions to that present in the MCMC samples. The AVB algorithm does not include approximate posterior distributions for the base functions or the registered functions. However, we can compare the estimated credible intervals obtained through AVB and MCMC sampling respectively for the target function. Here we will provide this comparison for the target functions associated with 1) the original (noiseless) Berkeley Boys Growth Velocity data and 2) the Berkeley Boys Growth Velocity data corrupted with Gaussian noise (see Appendix C.3, Earls and Hooker (2016), for more information on these data).

Generally the concern with using variational Bayes to approximate the posterior distributions is that variability is underestimated in the approximated posteriors, Wang and Titterington (2005) and Bishop (2006). In the appendices (Earls and Hooker, 2016), Figure C.1 contains credible bands determined through MCMC sampling for two of the registered functions from the noiseless growth velocity data. As can be seen in this figure, for noiseless observations, once the registration parameters have been set, the result of specifying a highly informative prior on the registered functions is that most of the variability from the mean is eliminated resulting in very narrow credible bands. Of course, if the credible bands associated with the registered functions are narrow, it makes sense that credible bands for the target function are even narrower. Thus, for the original boys growth velocity data, very little variability is exhibited in the MCMC samples of the target function and the differences between the estimated 95% credible band determined from the q distribution of the target function and that determined from the quantiles of the posterior MCMC sample are so small that on a graph they are indistinguishable. In the plot on the left in Figure 5 is a comparison of the width of the respective credible intervals for the target function at each time point. Here

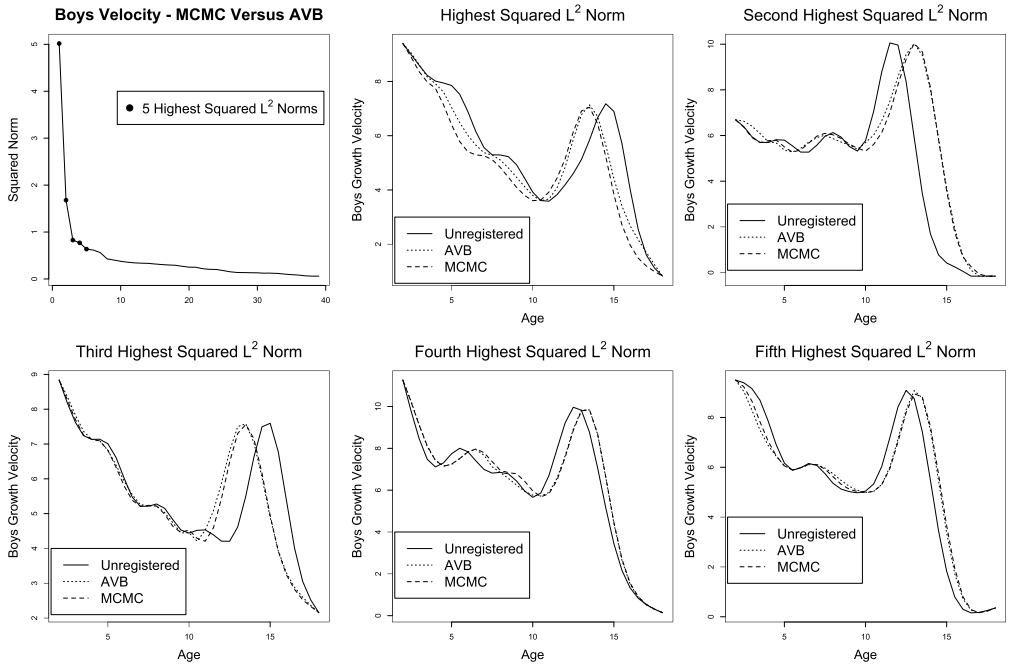

Figure 4: Registered Boys Growth Velocity – Differences Between MCMC and AVB Estimates. **Top Left** Plot of the squared L^2 norm of the difference between the MCMC and AVB estimates for each observation in decreasing order of magnitude. **Top Center and Right** The original unregistered function plotted with the MCMC and AVB estimates of the registered functions for the observations with the first two largest discrepancies between the MCMC and AVB estimates. **Lower** Plots of the next three observations with the highest squared L^2 norms of the difference between the MCMC and AVB estimates. The squared L^2 norm associated with the lower right plot is about .64. As can be seen in this illustration, at this level there are only small differences between the MCMC and AVB estimates.

it can be seen that not only are the credible intervals very narrow, but there are only minor differences in the width of the credible intervals determined from the approximate posterior distribution (AVB) and the MCMC sample.

As can be seen in Figure C.2 of the appendices (Earls and Hooker, 2016), if noisy data are recorded, the variability due to the noise process results in wider credible intervals for the registered functions (and hence the target function). On the right hand side of Figure 5 is a comparison of the width of the respective credible intervals for the target function at each time point for the model that both smooths and registers the noisy boys growth velocity data. In this illustration, we can see the variability present in the posterior MCMC sample for the target function is not captured well by the q distribution associated with the target function. This is not surprising and is an example of when performing MCMC sampling with an AVB initialization may be preferred to using AVB alone. Note: for these analyses both samplers were initialized

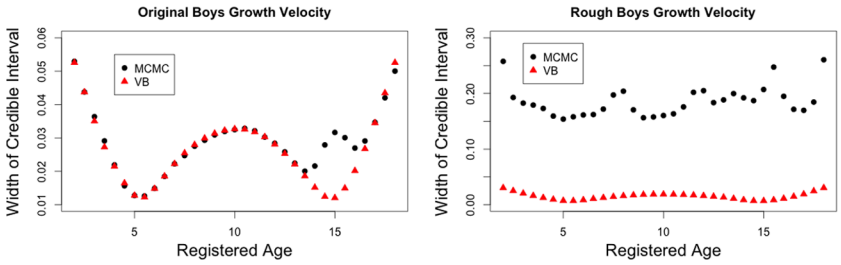

Figure 5: Pointwise Comparison of Credible Interval Width for the Target Function. Both plots contain the width of the credible interval based on the approximate posterior distribution for the target function obtained via AVB (triangles) and the width of the estimated credible intervals determined by the empirical quantiles of the posterior sample of the target function obtained through MCMC sampling (circles) at each time point. **Left** This is a comparison of the credible intervals for the target function when the data are recorded without noise. For most time points in this plot, the width of the credible interval at that time point using the approximate posterior is almost identical to that obtained through MCMC sampling. The biggest difference can be seen at ages 14–16 where the MCMC credible interval is slightly wider. **Right** This is a comparison of the credible intervals for the target function when the data are recorded with noise. Here it can be seen that the approximated posterior distribution of the target function significantly underrepresents the variability in the true posterior distribution.

with AVB estimates which made burn-in unnecessary. Post sampling analysis showed low autocorrelation in the final 999 samples of the target function (after thinning) for both the noiseless and noisy data.

5 Variational Approximation for Functional Prediction

5.1 Functional Prediction Algorithm

The probabilistic framework of our registration model provides a natural structure in which we can consider new observations. Functional prediction has previously been considered by Ferraty and Vieu (2006). Here we extend current methods by taking into account the phase variability of a partially observed function.

We will make the following assumptions.

1. We have a sample of approximated unregistered functions, $\mathbf{X}_i = (X_i(t_1), \ldots, X_i(t_p))'$, $i = 1, \ldots, N$.

2. $\mathbf{X}_i = (X_i(t_1), \ldots, X_i(t_p))'$, $i = 1, \ldots, N$ are registered using the registration method outlined in Section 2 via a MCMC sampler or adapted variational Bayes.

3. From (2) we have obtained estimates for the target function, $f(t)$, the registered

functions, $X_i(h_i(t))$, $i = 1, \ldots, N$, the warping functions, $h_i(w_i(t))$, $i = 1, \ldots, N$, σ^2_{z0}, and σ^2_{z1}.

4. A new function, $X_{N+1}(t)$ has been observed at the time points $(t_1, \ldots, t_r)'$, $r < p$.

5. $(X(h(t_1)), \ldots X(h(t_p)))' \approx N_p(\hat{\boldsymbol{\mu}}_{\mathbf{X(h)}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{X(h)}})$, the distribution of the registered functions can be approximated by a multivariate normal distribution using the sample mean, $\hat{\boldsymbol{\mu}}_{\mathbf{X(h)}}$, and sample covariance matrix, $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X(h)}}$, of the estimated registered functions obtained in (2).

6. $(w(t_1), \ldots w(t_{p-1}))' \approx N_{p-1}(\hat{\boldsymbol{\mu}}_{\mathbf{w}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{w}})$, the distribution of the base functions can be approximated by a multivariate normal distribution using the sample mean, $\hat{\boldsymbol{\mu}}_{\mathbf{w}}$, and sample covariance matrix, $\widehat{\boldsymbol{\Sigma}}_{\mathbf{w}}$, of the estimated base functions obtained in (2).

Under these assumptions, we will proceed as follows.

1. Register the partially observed function, $\mathbf{X^P}_{N+1} = (X_{N+1}(t_1), \ldots, X_{N+1}(t_r))'$ to the estimated target function, $\hat{f}(t)$, truncated to an appropriate registration time, t_f, $f \in \{1, \ldots, p\}$, so that $h_{N+1}(t_f) = t_r$.

2. Using the distributions from assumptions (5) and (6) above, the estimate of the partial registered function, $\mathbf{X^P}_{N+1}(\hat{\mathbf{h}}_{N+1}) = (X^P_{N+1}(\hat{h}_{N+1}(t_1)), \ldots, X^P_{N+1}(\hat{h}_{N+1}(t_f)))'$ and the estimate of the partial base function, $\widehat{\mathbf{w^P}}_{N+1} = (\widehat{w^P}_{N+1}(t_1), \ldots \widehat{w^P}_{N+1}(t_{f-1}))'$, estimate the registered and base functions to time t_p and t_{p-1} respectively using the conditional expectation of the multivariate normal distribution. Accordingly, denoting future registered observations and future warping function values, $\mathbf{X^F}_{N+1}(\mathbf{h}_{N+1})$ and $\mathbf{w^F}_{N+1}$, respectively, the estimates of these future values are $\widehat{\mathbf{X^F}}_{N+1}(\hat{\mathbf{h}}_{N+1}) = E(\mathbf{X^F}(\mathbf{h})|\mathbf{X^P}_{N+1}(\hat{\mathbf{h}}_{N+1}), \hat{\boldsymbol{\mu}}_{\mathbf{X(h)}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{X(h)}})$ and $\widehat{\mathbf{w^F}}_{N+1} = E(\mathbf{w^F}|\widehat{\mathbf{w^P}}_{N+1}, \hat{\boldsymbol{\mu}}_{\mathbf{w}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{w}})$.

3. Estimate the complete unregistered function, $X_{N+1}(t)$, using the inverse of the estimated warping function and the estimated registered function.

5.2 Determining the Last Registered Time

An additional random element in the prediction model is the last registered time of the truncated target function, t_f, used to register the partial observation. To obtain the best possible registration of the partial observation, a range of final registration times are considered over a finer domain. The efficiency of the adapted variational Bayes algorithm makes it possible to consider several possible partial registrations as follows.

1. For each of the time points $t_j, j \in \{m, \ldots, (m+k-1)\}$, $t_{m+k-1} < t_p$, the partially observed function, $X^P_{N+1}(t)$, is registered to the estimated target function, $\hat{f}(t)$,

truncated to time, t_j, so that $\hat{h}_{N+1(j)}(t_j) = t_r$, where $\hat{h}_{N+1(j)}(t)$ is the estimated warping function determined by registering $X^P_{N+1}(t)$ to the proposed final registration time t_j. Note, the first and last times considered in this interval are chosen by plotting the partial unregistered function and the target function together and determining a generous interval that contains the appropriate final registration time. This interval is subsequently made finer to allow this time to fall between two of the original time points.

2. Calculate $d_{t_j} = ||\mathbf{X^P}_{N+1} - (\hat{z}_{0(j)}\mathbf{1} + \hat{z}_{1(j)}\mathbf{f^U}_{(j)})||_2$ for each t_j, $j \in \{m, \ldots, m+k-1\}$ where $\mathbf{f^U}_{(j)} = (\hat{f}(t_1), \hat{f}(\hat{h}^{-1}_{N+1(j)}(t_2)), \ldots, \hat{f}(\hat{h}^{-1}_{N+1(j)}(t_r) = t_j))'$.
3. Set $t_f = \underset{t_j, j \in \{m,\ldots,m+k-1\}}{\arg\min} d_{t_j}$.

This algorithm determines the final registered time, t_f, that results in the minimum L^2 norm between the partially recorded unregistered function and the target function evaluated at the inverse of the warping function estimated using that final time. Note, for all j, $\mathbf{f^U}_{(j)}$ shares the same domain as the partially recorded unregistered function, $\mathbf{X^P}_{N+1}$.

5.3 Confidence Intervals for Predicted Functions

The efficiency of adapted variational Bayes for prediction also makes it possible to characterize variability in the estimates of the complete registered function, unregistered function, and base function via bootstrapping. To capture variability in the predicted functions, for each of $m = 1, \ldots, M$ iterations, perform the following steps.

1. Draw a new sample of N registered functions from $N_p(\hat{\boldsymbol{\mu}}_{\mathbf{X(h)}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{X(h)}})$.
2. Draw a new sample of N base functions from $N_{p-1}(\hat{\boldsymbol{\mu}}_{\mathbf{w}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{w}})$.
3. From the bootstrapped samples determined in (1) and (2), compute the sample mean and covariance matrix for the approximated registered functions, $\hat{\boldsymbol{\mu}}^{(m)}_{\mathbf{X(h)}}$ and $\widehat{\boldsymbol{\Sigma}}^{(m)}_{\mathbf{X(h)}}$ and also for the approximated base functions, $\hat{\boldsymbol{\mu}}^{(m)}_{\mathbf{w}}$ and $\widehat{\boldsymbol{\Sigma}}^{(m)}_{\mathbf{w}}$.
4. Register the partially observed function as in prediction step (1) in Section 5.1 above by setting the approximated target function, $\hat{\mathbf{f}}$, equal to $\hat{\boldsymbol{\mu}}^{(m)}_{\mathbf{X(h)}}$.
5. Draw a sample of size S from the distribution of $\mathbf{X^F(h)}|\mathbf{X^P}_{N+1}(\hat{\mathbf{h}}^{(m)}_{N+1})$, $\hat{\boldsymbol{\mu}}^{(m)}_{\mathbf{X(h)}}$, $\widehat{\boldsymbol{\Sigma}}^{(m)}_{\mathbf{X(h)}}$. Combine each of these samples of future values of the registered function with the estimated partially registered function determined in (4) to get a sample of S estimated registered functions.
6. Draw a sample of size S from the distribution of $\mathbf{w^F}|\widehat{\mathbf{w^P}}^{(m)}_{N+1}$, $\hat{\boldsymbol{\mu}}^{(m)}_{\mathbf{w}}, \widehat{\boldsymbol{\Sigma}}^{(m)}_{\mathbf{w}}$. Combine each of these samples of future values of the base function with the estimated

partial base function determined in (4) to get a sample of S estimated base functions.

7. Determine the S warping functions that result from step (6).
8. Determine the S unregistered functions that result from combining each of the S registered functions drawn in (5) with the corresponding inverse warping function from (7).

This process results in $M \times S$ bootstrapped samples of the registered function, warping function and unregistered function. From these samples, point wise bootstrapped confidence intervals can be determined for each function.

Here we have approximated the distributions of the base and registered functions by fitting a multivariate normal distribution. However, given the small sample size, approximating these distributions by a multivariate t distribution as suggested in Lange (1989) may provide confidence intervals with better coverage properties. Beran (1990) also provides a method for constructing robust confidence intervals in the context of univariate prediction problems.

5.4 Functional Prediction – El-Niño Data

The El-Niño data consist of weekly readings of sea surface temperature with the first observation in June of 1950. Complete data can be found at NOAA's Climate Prediction Center website (http://www.cpc.ncep.noaa.gov/data/indices/). The data that we are using for this analysis are found through Professor Frederic Ferraty's (Mathematics; University of Toulouse, France) website (http://www.math.univ-toulouse.fr/~ferraty/SOFTWARES/NPFDA/npfda-datasets.html). These data are a subset of the original data with monthly sea surface temperature records from June of 1950 to May of 2004. For this analysis, the bi-monthly observations are added to the data to prevent significant changes to the shape of a given function due to interpolation error. Also, light smoothing is applied to all functions.

The goal of our study is to predict how high temperatures will stay in the remaining part of the year in conjunction with how long temperatures will drop before they rise again based on the first seven months of temperature recordings from the lowest temperature recording in the previous year.

For this purpose, the data are restructured to define a "year" as the period of time between the lowest temperatures in consecutive calendar years. For example, the first year in our data set ranged from September 1950 to September 1951. Note, these "years" will not all be 12 months in length, and our final data had "years" that ranged from 11 to 14 months. For our analysis, we will concentrate on a subset of this group of restructured functions where the previous year's lowest temperature for each selected temperature profile is between 19.5 °C and 21 °C. Twenty-nine of the original temperature profiles fit this criterion. We will use the first 28 functions to predict the remaining portion of the 29th function based on the first 7 months of sea surface temperature observed in that year.

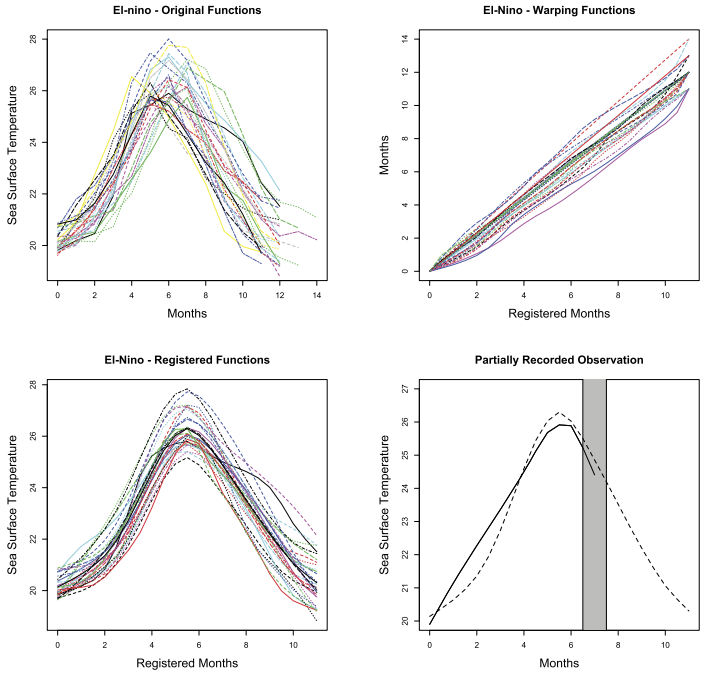

Figure 6: El-niño Data. **Top Left** Original 28 profiles of sea surface temperature. **Top Right** Estimated warping functions. As can be seen here, the time period of the original data ranged from 11 to 14 months. **Lower Left** Estimated registered temperature profiles. **Lower Right** The solid line is observation 29 recorded for 7 months. The dashed line is the estimated target function. The grey shaded area spans the 5 time points that are considered for the final time of the partial registration.

For the purpose of registration, all functions need to be recorded over the same interval of time. As mentioned above, in this particular case our data is recorded over a time periods that range from 11 to 14 months. An easy remedy to this situation is to perform a simple initial warping to each function that rescales every observation to an 11 month time frame. In our final analysis, this initial warping is accounted for when determining the final base functions used for the prediction algorithm.

The original unregistered functions and the functions registered using the GP model described in Section 2 are plotted in Figure 6. For this data set, to register significant features in the sample while retaining function variation beyond a scaling and vertical shift of the target function, individual warping parameters, γ_{w_i}, $i = 1, \ldots, 28$ were utilized instead of γ_w in (11). Significant differences in the amplitude variation in the original functions that is unassociated with temporal variation prevented the use of a global parameter. However, only 3 unique warping parameters in total were necessary.

Using the empirical mean of the 28 original registered functions as the target function, the first 7 months of sea surface temperature records from observation 29 are

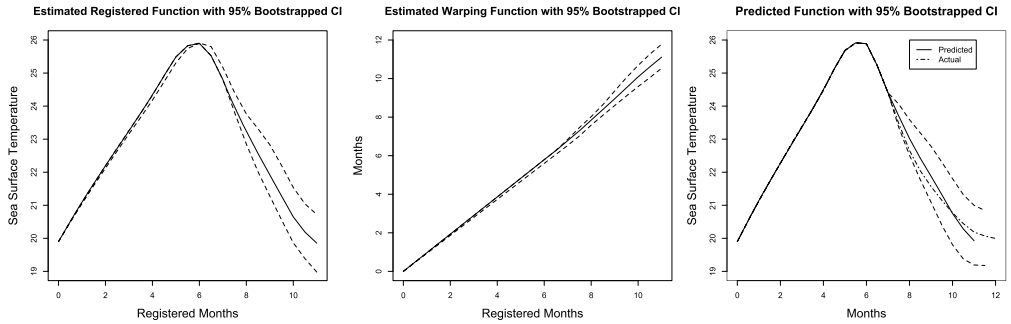

Figure 7: Estimates and Bootstrapped Confidence Intervals. **Left** Estimated registered function with 95% bootstrapped confidence interval. Note: In parts of the domain, the estimated confidence interval and estimated registered function overlap. This is largely due to a bimodal distribution of the last registered time. **Center** Estimated warping function with 95% bootstrapped confidence interval. **Right** Estimated unregistered function with 95% bootstrapped confidence interval. The dashed and dotted line is the true unregistered function.

registered to a piece of the target function where the final registered time is allowed to vary from 6.5 to 7.5 months. Between these months, a finer time interval corresponding to weekly records is used to allow for additional flexibility in determining the final registered time. The partially recorded function is plotted with the target function in the lower right panel of Figure 6. The grey shaded area includes the time points considered for the final time of the partial registration. After the optimal registration of the partially recorded observation is determined, estimates of the entire registered function, warping function, and unregistered function are determined using the model outlined in Section 5.1.

One-hundred bootstrapped samples were used to estimate the variability in the predictions of all three estimated functions. Figure 7 plots the initial estimates with the 95% bootstrapped confidence intervals. In addition, the plot of the estimated unregistered function also includes the true value of this function.

The primary advantage of registering the partially recorded observation before estimating future values is that we can capture variation in amplitude and timing separately. In Figure 7, the first plot captures the variability in the future level of sea surface temperature (amplitude variation), and answers the question,"How high can we expect sea surface temperatures to stay?". The second plot captures the variability in the timing of future observations (temporal variation) which addresses the question of, "When can we expect sea level temperatures to begin rising again?". The confidence interval for the unregistered function seen in the last frame of Figure 7, combines both amplitude and temporal variation to estimate the future trajectory of sea surface temperature for this year. In this illustration it can be seen that the main difference in the estimated and actual temperature profiles lies in the timing of the lowest observation. However, for this observation, the sea surface temperature at 12 months is not much different

than the sea surface temperature at 11.5 months. The predicted timing of the lowest temperature was 11.1 months.

One of the most notable features of this analysis is that there is little uncertainty in the registration of the first 7 months of sea surface temperatures. The most prominent feature in the data is the peak temperature that occurs anywhere from 4 to 8 months in the original data. In our partially recorded observation, as seen in Figure 6, the peak of the target function and the partially recorded observation are already closely aligned. Additionally, this observation happens to be similar in shape to the target function. The combination of these features resulted in only a minimal amount of variation in the estimated registered and warping functions in the first 7 months. However, we note here, this phenomenon is an artifact of these particular data, and in other analyses more variation in the registered timing of the partially recorded observation would be expected.

The El-niño data set provides a challenging registration problem. The registered functions vary significantly in directions beyond the target function. Choosing curve specific registration parameters enabled features common to all functions to be registered while retaining prominent features in each individual curve. This is just one example of the difficulties that can arise in registering functional data and in turn how these challenges can be addressed to analyze data that does not fit the "ideal" registration problem.

6 Discussion

In this paper we have developed a methodology for Bayesian registration that accounts for uncertainty in the registered functions, the warping functions, and the parameters associated with them. The hierarchical structure of this model allows multiple inferential procedures to be included in one analysis. We give an example where functional regularization and registration are performed in one model. However, these models will accommodate any combination of inferential procedures for which an appropriate prior can be defined. For instance, the models proposed here can easily be extended to a functional linear regression model where the registered functions or registered latent functions are considered as covariates.

While our registration algorithm provides high quality estimates in a highly flexible model, the associated computational costs due to running a MCMC sampling scheme for a high-dimensional model are considerable. To address these costs, we have proposed the Adapted Variational Bayes algorithm. This algorithm has been shown to converge in a similar way to traditional variational Bayes, but may not converge to a global maximum. However, an initial warping can be performed to move a function closer to its optimal registered value if the algorithm converges to a local maximum.

The bijective relationship between the base and warping functions in this model makes it possible to perform functional prediction in the context of registration. Using the estimated values of the base and registered functions from an initial registration of N functions, we use approximate distributions for these functions to predict future

values of the registered, warping, and unregistered functions of a new function that is only partially observed. Furthermore, the AVB algorithm makes it possible to re-sample from these distributions to bootstrap confidence intervals for these predicted values.

While the AVB algorithm provides estimates that are similar to their MCMC counterparts, for some models MCMC sampling remains the optimal inferential procedure. For example, in the model that combines smoothing and registration AVB estimates are approximate and preferably are only used to initialize an MCMC sampler. While we have used Metropolis within Gibbs samplers, these are likely not optimal. Determining the most efficient method of sampling from these joint posterior distributions is left for future work.

For simplicity, we have used inverse-Gamma or Gamma priors for the variance components in these models. The best choice of priors for these components is a uniform prior on the square root of the variance components as suggested by Gelman (2006). This is particularly a problem when the variance component is small or there is very little data to estimate these components. For the models presented here, Gamma priors for the smoothing parameters are sufficient as small changes in these parameters do not significantly affect the model. However, in the analysis of the noisy boys' growth velocity data in Appendix C (Earls and Hooker, 2016), we saw some evidence that the "uninformative" inverse-Gamma prior resulted in a slightly upward biased estimate of the noise variance.

Modeling functions as Gaussian processes in a Bayesian hierarchical model offers a unified approach to performing multiple inference procedures for functional data within one model. In Earls and Hooker (2014), we established the properties of functional data estimates determined by approximating functional distributions over a finite subset of observed time points and estimating functions at unobserved time points with linear interpolation. Furthermore, we demonstrated that providing smoothing information in scale or covariance matrices results in regularized functional estimates. Here we extend this work by using informative covariance matrices to register functions where the registered functions are modeled as Gaussian processes. Future work includes adapting these models for other areas of inference for functional data and continuing to improve on the models we have proposed here.

Supplementary Material

Appendices for Variational Bayes for Functional Data Registration, Smoothing, and Prediction (DOI: 10.1214/16-BA1013SUPP; .pdf).

References

Beran, R. (1990). "Calibrating prediction regions." *Journal of the American Statistical Association*, 85(411): 715–723. MR1138352. 575

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer, New York. MR2247587. doi: http://dx.doi.org/10.1007/978-0-387-45528-0. 563, 570

Brumback, C. and Lindstrom, J. (2004). "Self-modeling with flexible, random time transformations." *Biometrics*, 60: 461–470. MR2066281. doi: http://dx.doi.org/10.1111/j.0006-341X.2004.00191.x. 558

Earls, C. and Hooker, G. (2014). "Bayesian covariance estimation and inference in latent Gaussian process models." *Statistical Methodology*, 18: 79–100. MR3151865. doi: http://dx.doi.org/10.1016/j.stamet.2013.10.001. 559, 561, 579

Earls, C. and Hooker, G. (2016). "Appendices for Variational Bayes for Functional Data Registration, Smoothing, and Prediction." *Bayesian Analysis*. doi: http://dx.doi.org/10.1214/16-BA1013SUPP. 558, 559, 560, 562, 565, 566, 570, 571, 579

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer, New York. MR2229687. 572

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper." *Bayesian Analysis*, 1(3): 515–534. MR2221284. 579

Gervini, D. and Gasser, T. (2004). "Self-modeling warping functions." *Journal of the Royal Statistical Society, Series B*, 66: 959–971. MR2102475. doi: http://dx.doi.org/10.1111/j.1467-9868.2004.B5582.x. 558, 567

Gervini, D. and Gasser, T. (2005). "Nonparametric maximum likelihood estimation of the structure of a sample of curves." *Biometrika*, 92: 801–820. MR2234187. doi: http://dx.doi.org/10.1093/biomet/92.4.801. 558

Goldsmith, J., Wand, M., and Crainiceanu, C. (2011). "Functional regression via variational Bayes." *Electronic Journal of Statistics*, 5(572). MR2813555. doi: http://dx.doi.org/10.1214/11-EJS619. 563, 564

James, G. (2007). "Curve alignment by moments." *The Annals of Applied Statistics*, 1(2): 480–501. MR2415744. doi: http://dx.doi.org/10.1214/07-AOAS127. 558, 567

Kneip, A. and Gasser, T. (1992). "Statistical tools to analyze data representing a sample of curves." *The Annals of Statistics*, 1(2): 480–501. MR1186250. doi: http://dx.doi.org/10.1214/aos/1176348769. 558

Kneip, A. and Gasser, T. (1995). "Searching for structure in curve samples." *Journal of the American Statistical Association*, 90: 1179–1188. 558

Kneip, A. and Ramsay, J. O. (2008). "Combining registration and fitting for functional models." *Journal of the American Statistical Association*, 103(483): 1155–1165. MR2528838. doi: http://dx.doi.org/10.1198/016214508000000517. 558

Lange (1989). "Robust statistical modeling using the t distribution." *Journal of the American Statistical Association*, 84(408): 881–896. MR1134486. 575

Liu, X. and Muller, H. (2004). "Functional convex averaging and synchronization for time-warped random curves." *Journal of the American Statistical Association*, 99: 687–699. MR2090903. doi: http://dx.doi.org/10.1198/016214504000000999. 567

Liu, X. and Yang, M. (2009). "Simultaneous curve registration and clustering for functional data." *Computational Statistics and Data Analysis*, 53: 1361–1376. MR2657097. doi: http://dx.doi.org/10.1016/j.csda.2008.11.019. 558

Omerod, J. and Wand, M. (2010). "Explaining variational approximations." *The American Statistician*, 64: 140–153. MR2757005. doi: http://dx.doi.org/10.1198/tast.2010.09058. 563

Raket, L., Sommer, S., and Markussen, B. (2014). "A nonlinear mixed-effects model for simultaneous smoothing and registration of functional data." *Pattern Recognition Letters*, 38: 1–7. 558

Ramsay, J. O. and Li, X. (1998). "Curve registration." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 60(2): 351–363. MR1616045. doi: http://dx.doi.org/10.1111/1467-9868.00129. 558, 559, 562, 566, 567

Ramsay, J. O. and Silverman, B. (2005). *Functional Data Analysis*. Springer, New York. MR2168993. 558

Rao, C. (1945). "Information and accuracy attainable in the estimation of statistical parameters." *Bulletin of Calcutta Mathematical Society*, 37: 81–91. MR0015748. 567

Ronn, B. (2001). "Nonparametric maximum likelihood estimation of shifted curves." *Journal of the Royal Statistical Society, B*, 63: 243–259. MR1841413. doi: http://dx.doi.org/10.1111/1467-9868.00283. 558

Sakoe, H. and Chiba, S. (1978). "Dynamic programming algorithm optimization for spoken word recognition." *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(1): 43–49. 558

Sangalli, L., Secchi, P., Vantini, S., and Vitelli, V. (2010). "k-mean alignment for curve clustering." *Computational Statistics and Data Analysis*, 54: 1219–1233. MR2600827. doi: http://dx.doi.org/10.1016/j.csda.2009.12.008. 558

Silverman, B. (1995). "Incorporating parametric effects into functional principal components analysis." *Journal of the Royal Statistical Society*, 57: 673–689. MR1354074. 558

Srivastava, A., Wu, W., Kurtek, S., Klassen, E., and Marron, J. (2011). "Registration of functional data using Fisher–Rao metric." *arXiv:1103.3817*. 558, 559, 562, 566, 567, 568

Tang, R. and Muller, H. (2008). "Pairwise curve synchronization for functional data." *Biometrika*, 95(4): 875–889. MR2461217. doi: http://dx.doi.org/10.1093/biomet/asn047. 558, 567

Telesca, D. and Inoue, L. (2007). "Bayesian hierarchical curve registration." *Journal of the American Statistical Association*, 103(481): 328–339. MR2420237. doi: http://dx.doi.org/10.1198/016214507000001139. 558

Tuddenham, R. and Snyder, M. (1954). "Physical growth of California boys and girls from birth to eighteen years." *University of California Publications in Child Development I*, 183–364. 569

Tzikas, D. G., Likas, A. C., and Galatsanos, N. P. (2008). "The variational approximation for Bayesian inference." *Signal Processing Magazine, IEEE*, 25(6): 131–146. 564

Wang, B. and Titterington, D. (2005). "Inadequacy of interval estimates corresponding to variational Bayesian approximations." *Proc. 10th Int. Wrkshp Artificial Intelligence and Statistics*, 373–380. 570

Wang, K. and Gasser, T. (1997). "Alignment of curves by dynamic time warping." *The Annals of Statistics*, 25(3): 1251–1276. MR1447750. doi: http://dx.doi.org/10.1214/aos/1069362747. 558

Zhang, Y. and Telesca, D. (2014). "Joint clustering and registration of functional data." *arXiv:1403.7134*. 558