Registration for exponential family functional data

Julia Wrobel[1*], Vadim Zipunnikov[2], Jennifer Schrack[3,4], and Jeff Goldsmith[1]

[1]Department of Biostatistics, Mailman School of Public Health, Columbia University
[2]Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University
[3]Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University
[4]Longitudinal Studies Section, Translational Gerontology Branch, National Institute on Aging, National Institutes of Health
[*]jw3134@cumc.columbia.edu

December 14, 2017

Abstract

We introduce a novel method for separating amplitude and phase variability in exponential family functional data. Our method alternates between two steps: the first uses generalized functional principal components analysis to calculate template functions, and the second estimates smooth warping functions that map observed curves to templates. Existing approaches to registration have primarily focused on continuous functional observations, and the few approaches for discrete functional data require a pre-smoothing step; these methods are frequently computationally intensive. In contrast, we focus on the likelihood of the observed data and avoid the need for preprocessing, and we implement both steps of our algorithm in a computationally efficient way. Our motivation comes from the Baltimore Longitudinal Study on Aging, in which accelerometer data provides valuable insights into the timing of sedentary behavior. We analyze binary functional data with observations each minute over 24 hours for 592 participants, where values represent activity and inactivity. Diurnal patterns of activity are obscured due to misalignment in the original data but are clear after curves are aligned. Simulations designed to mimic the application indicate that the proposed methods outperform competing approaches in terms of estimation accuracy and computational efficiency. Code for our method and simulations is publicly available.

Key Words: Accelerometers, warping, alignment, functional principal component analysis, generalized functional data, binary functional data

1   Introduction

In the most common setting for functional data analysis, the basic unit of observation is the real-valued curve $Y_i(t)$ for subjects $i \in 1, \ldots, N$. More recently, there has been interest in exponential family functional data, where $Y_i(t)$ comes from a non-Gaussian distribution; it is typically assumed that $Y_i(t)$ has a smooth and continuous latent mean, $\mu_i(t) = E[Y_i(t)]$. Our motivation is the study of activity and inactivity using data collected with accelerometers, a setting with binary functional data. Figure 1 shows binary curves $Y_i(t)$ for two participants taking the value 1 when the participant is active and 0 when the participant is inactive. A solid curve shows an estimate of the smooth latent mean $\mu_i(t)$, interpreted as the probability the subject will be active at each minute in the 24 hours of observation. Other recent examples of non-Gaussian functional data include agricultural studies on the feeding behavior of pigs, spectral backscatter from long range infrared light detection, and longitudinal studies of drug use (Gertheiss et al., 2015; Serban et al., 2013; Huang et al., 2014).

Functional data often include both phase displacement, the misalignment of major features shared across curves, and amplitude variability. The process underlying phase variation may itself be of interest; additionally, when the interest is primarily in the amplitude variation, phase variation can artificially distort analyses of amplitude and mask the shared data structure. Methods for curve *registration*, which transform functional data to align features, are focused on addressing the problem of phase variation. The goal of registration is to warp the functional domain, which we will refer to as *time*, so that phase variation is minimized and the major features of the curves are aligned. This process necessitates a distinction between *chronological time* (t_i^*), which is the originally observed time for each subject, and *internal time* (t), which is the unobserved time on which major features are aligned across subjects (chronological and internal time are often referred
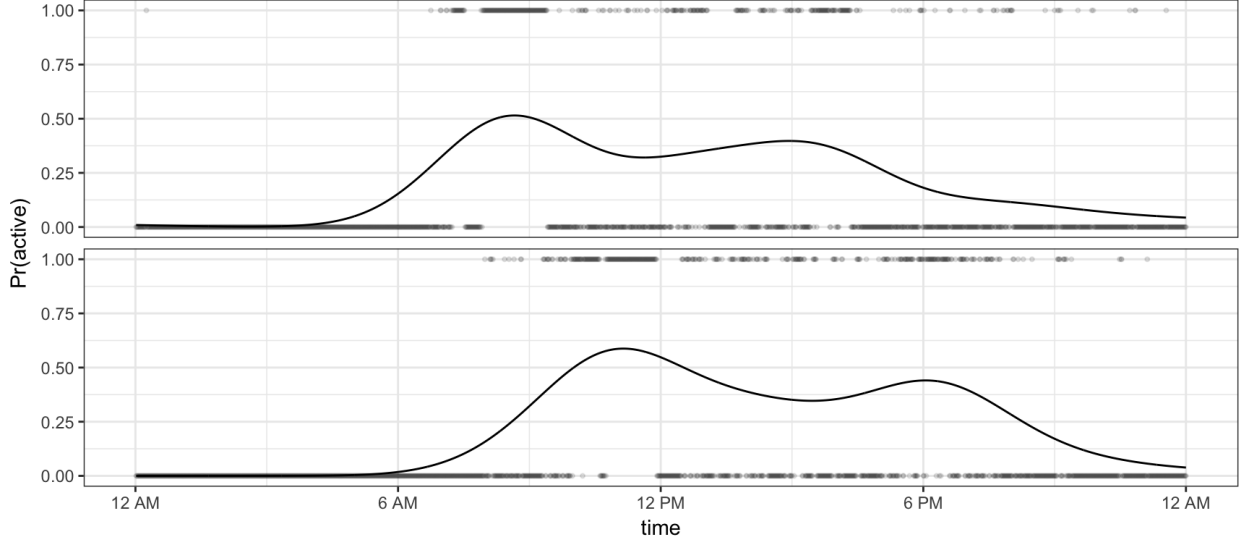
Figure 1: Points are binary curves for two subjects from the BLSA data before registration, where values of 1 and 0 represent activity and inactivity, respectively. The solid curves are estimates of the latent probability of activity, $\mu_i(t)$, and are fit for each subject using kernel smoothers.

to in the functional data literature as clock and system time, respectively). Stated differently, internal time is the true but unknown time over which aligned curves are generated and chronological time is the shifted time on which misaligned curves are observed. The registration problem amounts to recovering the subject-specific warping functions $h_i : t \mapsto t_i^*$ which map internal time to chronological time. Inverse warping functions $h_i^{-1}(t_i^*)$ can then be used to obtain aligned curves $Y_i(t)$ from observed data $Y_i(t_i^*)$. To emphasize the conceptual difference between chronological and internal times, we index t_i^* by subject but do not index t.

We are interested in registering actigraphy data that comes from the Baltimore Longitudinal Study of Aging. The BLSA is an observational study of healthy aging and included an accelerometer for monitoring activity (Schrack et al., 2014). Our dataset includes 592 people, for whom accelerometer observations are gathered over 24 hours in one-minute epochs giving chronological times on equally spaced grids of length 1440. We are especially interested in activity and inactivity, defined using a threshold of raw accelerometer observations, as both low activity levels and excessive sedentary behavior have been associated with poor health outcomes. Moreover, there is a growing research interest in understanding temporal/diurnal patterns of accumulation of seden-

tary time (Diaz et al., 2017; Martin et al., 2014; Yerrakalva et al., 2017). However, those analyses typically report diurnal averages that ignore the differences between subject specific wake time and mix together amplitude and phase.

The left panel of Figure 2 shows observed binary curves for all subjects against chronological time. In this lasagna plot (Swihart et al., 2010) subjects appear in rows, with active and inactive minutes shown in dark and light blue, respectively. This figure clearly shows the variability in the timing of inactivity across subjects, who may start or end the day at different times, and may accrue inactive minutes in sedentary bouts at different times. Such misalignment attenuates the diurnal patterns of activity that we believe to be present based on the naturally occurring circadian rhythm. The right panel of Figure 2 shows estimates of the mean $\mu_i(t)$ obtained using a Gaussian kernel smoother; these smooths illustrate the phase misalignment across subjects. The shift in timing of activity and inactivity is also seen in Figure 1. Specifically, the subject in the top row wakes up earlier, has a peak of activity, and then has a low activity level for the rest of the day, while subject in the bottom row has a similar but shifted pattern of behavior.
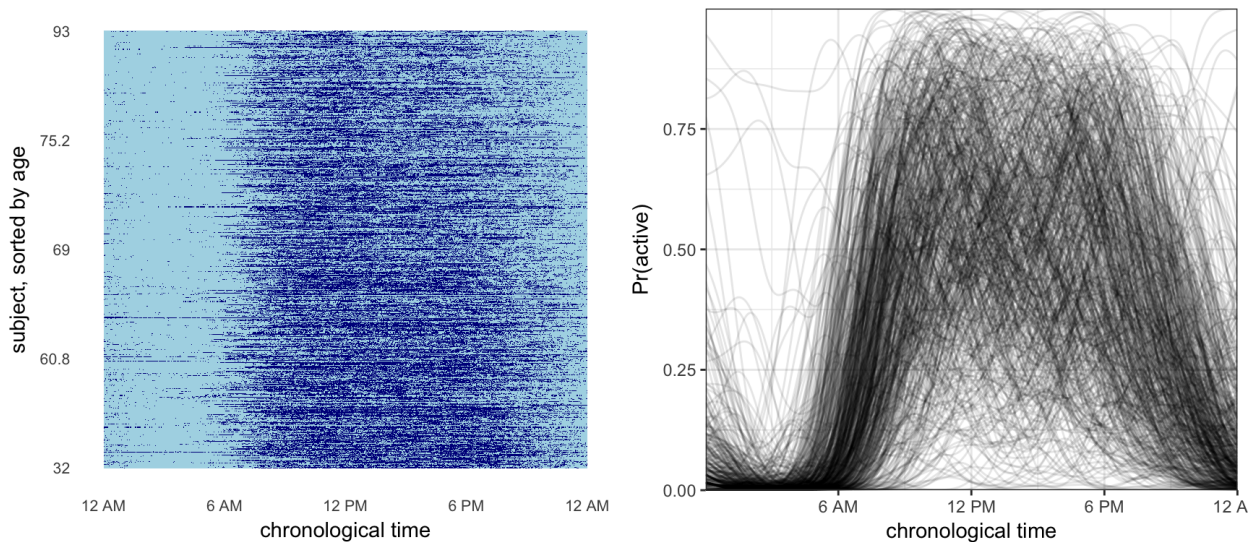

Figure 2: Plots of the unregistered data for 592 subjects at all 1440 minutes observed. At left is a lasagna plot, where row is the binary curve for a single subject and inactive and active observations are colored in light and dark blue, respectively. The rows are sorted by age, so that youngest subjects are at the bottom of the plot and oldest subjects are at the top. At right are smoothed curves for each subject, fit using kernel smoothers.

4

We propose novel methods for the registration of exponential family functional data, with particular emphasis on binary curves. Due to the large size of the data computational efficiency is critical, and we take this into consideration at each step of our method development. Section 2 provides a review of the relevant literature on registration and exponential-family functional principal components analysis; Section 3 details our methods; Section 4 shows simulation results, and Section 5 applies our method to the BLSA data. We conclude with a discussion in Section 6.

2   Literature Review

Our method draws on two distinct bodies of work in functional data analysis, which we review below. First, in 2.1, we review curve registration; this literature is primarily focused on Gaussian curves, with relatively little existing work for non-Gaussian curves. Then, in 2.2, we give an overview of exponential family FPCA, which is itself a relatively new area of interest in functional data analysis.

2.1   Registration

Several approaches for registering functional data have been proposed; we review these briefly, and suggest Marron et al. (2015) for a more detailed overview. *Dynamic time warping* uses dynamic programming to minimize distance between two functions (Sakoe and Chiba, 1978). Implementations of this class of methods are fast and provide a globally optimal solution under a pre-specified distance measure, but do not provide smooth, differentiable warping functions and are often not strictly monotonic. *Landmark registration* aligns key data structures called landmarks (e.g. maxima, minima, or other observable features) using piecewise linear functions (Gasser and Kneip, 1995). While landmark methods often provide excellent results at the position of the landmark, they can provide poor results between landmarks and are not easily automated. *Template registration* aligns each curve to a template curve by optimizing an objective function. This approach necessitates choosing the template, the objective function, and the optimization approach.

5

A common approach to template registration uses functional principal component analysis (FPCA) to select the template (Ramsay and Li, 1998; Kneip and Ramsay, 2008). First these methods estimate the template, and then estimate the warping functions for a given template; these steps are iterated until convergence. Warping functions are estimated using a sum of squared errors approach, often penalized to enforce smoothness. There is a large registration literature operating under and expanding this framework, including Sangalli et al. (2010), Vantini (2012), and Hadjipantelis et al. (2015). Intuitively, functional principal components describe the main directions of variation in a set of curves, making FPCA a natural tool for identifying the features to which data is registered.

Srivastava et al. (2011) introduce a metric for calculating warping functions based on the Fisher-Rao distance. They calculate a Karcher mean template and define a square root slope function transform (SRSF) of the observed curves. Minimizing the $\mathbb{L}^2$ norm between two SRSFs is equivalent to minimizing their Fisher-Rao distance. Since the SRSF uses the derivative of the observed curve, the data to be registered are required to be smooth. The Fisher-Rao metric has been the basis for several recent approaches to registration (Tucker et al., 2013, 2014), many of which are implemented in the `fdasrvf` package (Tucker, 2017).

Although most work in registration has focused on continuous data, there are two recent exceptions. Wu and Srivastava (2014) apply the SRSF approach to binary functional data by pre-smoothing data with a Gaussian kernel and registering the result. Panaretos and Zemel (2016) present a theoretical framework for separation of amplitude and phase variation of random point processes. The authors formalize a set of regularity conditions for warping functions that includes smoothness, proximity to the identity map, and unbiasedness, and establish a set of nonparametric estimators using the Fréchet mean and the L^2 Wasserstein distance. However, since these estimators register the unobserved probabilities of the point processes, the authors also begin by smoothing binary curves using kernel density estimation.

In contrast to the extensive literature on registration we develop an approach that can be applied to continuous and discrete data and does not require presmoothing. We also emphasize

computational efficiency, a matter of real concern given our high-dimensional data application.

2.2   FPCA for exponential family curves

Functional principal components analysis is popular for identifying modes of variation in functional data (Ramsay and Silverman, 2005). The most common approaches to FPCA decompose the variance-covariance matrix of de-meaned functional observations; see Ramsay and Silverman (2005), Yao et al. (2005), or Goldsmith et al. (2013) for details on this approach. Hall et al. (2008) adapted the methods in Yao et al. (2005) for binary functional data by positing a smooth latent Gaussian process and then estimating and decomposing the covariance of this process. Serban et al. (2013) refined and extended this approach by improving approximations in the estimation procedure, increasing accuracy for rare events, and allowing spatial structures. However, as demonstrated in Gertheiss et al. (2017), the adaptation of Yao et al. (2005) to exponential family data has an inherent bias due to reliance on a marginal rather than conditional mean estimate.

Probabilistic FPCA is an appealing alternative to the covariance smoothing approach. This framework conceptualizes PCA as a likelihood-based model, can be approached from a Bayesian perspective, and easily accounts for sparse or irregular data. Tipping and Bishop (1999) introduce probabilistic PCA, and a related approach is used by James et al. (2000) for functional data. van der Linde (2008) extends probabilistic FPCA to binary and count data through use of a Taylor-approximated likelihood function, while Goldsmith et al. (2015) uses a fully Bayesian parameter specification for generalized FPCA and function-on-scalar regression. These approaches often relate the expected value of observed data to a smooth latent process through a link function, and for this reason will be referred to as methods for generalized FPCA or GFPCA. Because all parameters are estimated simultaneously rather than sequentially, the probabilistic framework avoids the bias inherent in the covariance decomposition approach.

Our contributions to this literature focus on improving accuracy and efficiency for binary FPCA by estimating parameters in a probabilistic framework using a novel variational EM algorithm. To do this, we adapt the approach developed by Jaakkola and Jordan (1997) for logistic regression,

which has since been extended to binary (non-functional) PCA (Tipping, 1999; Schein et al., 2003). These methods rely on a variational approximation to the Bernoulli likelihood that is a true lower bound and allows for closed form updates of parameters. In contrast to van der Linde (2008), which uses a second-order Taylor expansion of the log likelihood to approximate a lower bound to the true distribution, our variational approximation is a true lower bound. While our method is optimized for binary data, similar derivations are possible for functional data from other exponential family distributions.

3   Methods

We first introduce the conceptual framework for our approach. Our goal is to estimate inverse warping functions h_i^{-1} which map unregistered *chronological time* t_i^* to registered *internal time* t such that $h_i^{-1}(t_i^*) = t$. Then for subject i, the unregistered and registered response curves are $Y_i(t_i^*)$ and $Y_i(t) = Y_i\left(h_i^{-1}(t_i^*)\right)$, respectively. Without loss of generality, we assume both t^* and t are on $[0, 1]$. We require that functions h_i^{-1} are monotonically increasing and satisfy the endpoint constraints $h_i^{-1}(0) = 0$ and $h_i^{-1}(1) = 1$. Notationally, we combine warping functions with exponential family GFPCA through the following:

$$E\left[Y_i(h_i^{-1}(t_i^*))|c_i, h_i^{-1}\right] = \mu_i(t)$$

$$g[\mu_i(t)] = \alpha(t) + \sum_{k=1}^{K} c_{ik}\psi_k(t). \tag{1}$$

The aligned response curves $Y_i\left(h_i^{-1}(t_i^*)\right)$ for each $t_i^* \in [0, 1]$ arise from the canonical exponential family of distributions with density

$$P\left(Y_i(h_i^{-1}(t_i^*))|\mu_i(t)\right) = \exp\left\{\left(Y_i(h_i^{-1}(t_i^*))g[\mu_i(t)] - b\left(g[\mu_i(t)]\right)\right)/\varphi + c(Y_i(h_i^{-1}(t_i^*)), \varphi)\right\} \tag{2}$$

where $E\left[Y_i(h_i^{-1}(t_i^*))|\mu_i(t)\right] = \mu_i(t) = b'\left(g[\mu_i(t)]\right)$, $Var\left[Y_i(h_i^{-1}(t_i^*))|\mu_i(t)\right] = b''\left(g[\mu_i(t)]\right)\varphi$, and φ is the dispersion parameter. The subject-specific means $\mu_i(t)$ implicitly condition on parameters in

model (1) and are used as templates in our warping step. Through a known link function g, the $\mu_i(t)$ are related to a linear predictor containing the population level mean $\alpha(t)$ and a linear combination of population level basis functions $\psi(t)$ and subject-specific score vectors $\boldsymbol{c}_i \sim N(0, \boldsymbol{I}_{K \times K})$. This formulation assumes that registered curves can be decomposed using GFPCA and, in doing so, places both registration and GFPCA in a single model.

Our estimation method is based on model (1) and alternates between the following steps:

1. **Subject-specific means $\mu_i(t)$ are estimated via probabilistic GFPCA, conditional on the current estimate of inverse warping functions $h_i^{-1}(t)$.**

2. **Inverse warping functions h_i^{-1} are estimated by maximizing the log likelihood of the exponential family distribution under monotonicity and endpoint constraints on h_i, conditional on the current estimate of μ_i.**

We iterate between steps (1) and (2) until curves are aligned.

Similar registration approaches for continuous-valued response curves have used the squared error loss for optimizing warping functions which, in a Gaussian setting, is equivalent to maximizing the likelihood function. However, our likelihood-based approach, which registers non-Gaussian data by extending the exponential-family framework, is novel. In contrast to registration methods for discrete functional data, we register observed binary curves using smooth templates rather than aligning pre-smoothed functional data. Because our application has 592 subjects measured at 1440 time points each, computational efficiency is critical. To this end we develop a novel fast approach to binary FPCA in Step 1, which we describe in Section 3.1, and optimize speed in estimating warping functions in Step 2, which is described in Section 3.2.

3.1 Binary FPCA

We first detail our novel EM approach to binary FPCA. Model (1) provides a conceptual framework, assuming that each curve $Y_i(t)$ is evaluated over internal time $t \in [0, 1]$. In practice, data for subject i is observed on the discrete grid, $\boldsymbol{t}_i = \{t_{i1}, \ldots, t_{iD_i}\}$, which may be irregular across subjects, and

therefore (in contrast to t) is indexed by subject. Functions indexed by the vector $\boldsymbol{t}_i$ are $D_i \times 1$ vectors of those functions evaluated on the observed time points (e.g. $Y_i(\boldsymbol{t}_i) = [Y_i(t_{i1}), \ldots, Y_i(t_{iD_i})]^T$ and $\psi_k(\boldsymbol{t}_i) = [\psi_k(t_{i1}), \ldots, \psi_k(t_{iD_i})]^T)$. The population level mean $\alpha(t)$ and principal components $\psi_k(t)$, $1 \leq k \leq K$, are expanded using a fixed B-spline basis, $\boldsymbol{\Theta}_\phi(t)$, of K_ϕ basis functions $\theta_1(t), \ldots, \theta_{K_\phi}(t)$. Let $\boldsymbol{\Theta}_\phi(\boldsymbol{t}_i)$ be the $D_i \times K_\phi$ B-spline matrix evaluated at $\boldsymbol{t}_i$ and a $1 \times K_\phi$ vector when evaluated at a single point t_{ij}; then $\alpha(\boldsymbol{t}_i) = \boldsymbol{\Theta}_\phi(\boldsymbol{t}_i)\boldsymbol{\alpha}_\Theta$ and $\Psi(\boldsymbol{t}_i) = [\psi_1(\boldsymbol{t}_i), \ldots, \psi_K(\boldsymbol{t}_i)] = \boldsymbol{\Theta}_\phi(\boldsymbol{t}_i)\boldsymbol{\Psi}_\Theta$ where the vector $\boldsymbol{\alpha}_\Theta$ and matrix $\boldsymbol{\Psi}_\Theta$ of size $K_\phi \times K$ contain the spline coefficients for the mean and principal components, respectively. Observed on the discrete grid $\boldsymbol{t}_i$, the linear predictor in (1) becomes

$$g[\mu_i(\boldsymbol{t}_i)] = \boldsymbol{\Theta}_\phi(\boldsymbol{t}_i)\left(\boldsymbol{\alpha}_\Theta + \boldsymbol{\Psi}_\Theta \boldsymbol{c}_i\right). \tag{3}$$

We estimate parameters in model (3) using an EM algorithm that incorporates a variational approximation. We assume $\boldsymbol{c}_i \sim MVN(0, I)$. For the binary case that is our main interest, $g(\cdot)$ is the logit function, for each point on the grid for the ith subject, $Y_i(t_{ij}) \sim Bernoulli(\mu_i(t_{ij}))$ where $\mu_i(t_{ij}) = P(Y_i(t_{ij}) = 1|\boldsymbol{c}_i)$. It is convenient to rewrite the probability density function as

$$P(Y_i(t_{ij})|\boldsymbol{c}_i) = g^{-1}\left\{[2Y_i(t_{ij}) - 1][\boldsymbol{\Theta}_\phi(t_{ij})\left(\boldsymbol{\alpha}_\Theta + \boldsymbol{\Psi}_\Theta \boldsymbol{c}_i\right)]\right\}, \tag{4}$$

so that the full unobserved joint likelihood for the observations and score vectors is

$$L(\boldsymbol{Y}, \boldsymbol{c}) \propto \prod_{i=1}^{I}\prod_{j=1}^{D_i} g^{-1}\left\{[2Y_i(t_{ij}) - 1][\boldsymbol{\Theta}_\phi(t_{ij})\left(\boldsymbol{\alpha}_\Theta + \boldsymbol{\Psi}_\Theta \boldsymbol{c}_i\right)]\right\} \times \prod_{i=1}^{I}\exp\left\{-\frac{\boldsymbol{c}_i^T\boldsymbol{c}_i}{2}\right\}. \tag{5}$$

Let scalar $A_i(t_{ij}) = \boldsymbol{\Theta}_\phi(t_{ij})\left(\boldsymbol{\alpha}_\Theta + \boldsymbol{\Psi}_\Theta \boldsymbol{c}_i\right)$ and $\lambda(z) = \frac{0.5 - g^{-1}(z)}{2z}$. A variational approximation to (4), based on the approximation in Jaakkola and Jordan (1997), is

$$\tilde{P}\left(Y_i(t_{ij})|\boldsymbol{c}_i, \xi_i(t_{ij})\right) = g^{-1}(\xi_i(t_{ij}))\exp\left\{\frac{(2Y_i(t_{ij}) - 1)A_i(t_{ij}) - \xi_i(t_{ij})}{2} + \lambda\left(\xi_i(t_{ij})\right)\left(A_i(t_{ij})^2 - \xi_i(t_{ij})^2\right)\right\}. \tag{6}$$

The resulting variational joint likelihood is

$$\tilde{L}(\boldsymbol{Y}, \boldsymbol{c}) \propto \prod_{i=1}^{I} \prod_{j=1}^{D_i} \tilde{P}\left(Y_i(t_{ij}) | \boldsymbol{c}_i, \xi_i(t_{ij})\right) \times \prod_{i=1}^{I} \exp\left\{ -\frac{\boldsymbol{c}_i^T \boldsymbol{c}_i}{2} \right\}. \tag{7}$$

We use an EM algorithm to obtain parameter estimates from (7) by (i) finding the posterior distribution of the scores; (ii) maximizing $\tilde{L}(\boldsymbol{Y}, \boldsymbol{c})$ with respect to $\boldsymbol{\xi}$; and (iii) maximizing the variational likelihood with respect to $\boldsymbol{\alpha}_\Theta$ and $\boldsymbol{\Psi}_\Theta$. These three steps are described in Sections 3.1.1, 3.1.2, and 3.1.3, with more details given in Appendix A.1. Convergence is attained when the squared difference between parameter estimates and their previous solution become arbitrarily small.

3.1.1 Calculating posterior scores

The posterior scores for each subject, derived via Bayes' rule, follow a multivariate normal distribution $\boldsymbol{c}_i | Y_i(\boldsymbol{t}_i), \xi_i(\boldsymbol{t}_i) \sim MVN(\boldsymbol{m}_i, \boldsymbol{C}_i)$ with:

$$\boldsymbol{C}_i = \left(\boldsymbol{I}_{K \times K} - 2\boldsymbol{\Psi}_\Theta^T \boldsymbol{\Theta}_\phi(\boldsymbol{t}_i)^T [diag(\lambda(\xi_i(\boldsymbol{t}_i)))] \boldsymbol{\Theta}_\phi(\boldsymbol{t}_i) \boldsymbol{\Psi}_\Theta \right)^{-1}$$

and

$$\boldsymbol{m}_i = \boldsymbol{C}_i \left(\boldsymbol{\Psi}_\Theta^T \boldsymbol{\Theta}_\phi(\boldsymbol{t}_i)^T (Y_i(\boldsymbol{t}_i) - \frac{1}{2}) + 2\boldsymbol{\Psi}_\Theta^T \boldsymbol{\Theta}_\phi(\boldsymbol{t}_i)^T [diag(\lambda(\xi_i(\boldsymbol{t}_i)))] \boldsymbol{\Theta}_\phi(\boldsymbol{t}_i) \boldsymbol{\alpha}_\Theta \right)$$

where $\xi_i(\boldsymbol{t}_i)$ is a vector of length D_i and $diag(\lambda(\xi_i(\boldsymbol{t}_i)))$ is a $D_i \times D_i$ diagonal matrix.

3.1.2 Maximizing $\tilde{L}(\boldsymbol{Y}, \boldsymbol{c})$ with respect to $\boldsymbol{\xi}$

We maximize the variational likelihood with respect to ξ_i^2, obtaining

$$\hat{\xi}_i(t_{ij})^2 = E_{\tilde{P}_{post}}(A_i(t_{ij})^2)$$

$$= \boldsymbol{\alpha}_\Theta^T \boldsymbol{\Theta}_\phi(t_{ij})^T \boldsymbol{\Theta}_\phi(t_{ij}) \boldsymbol{\alpha}_\Theta + 2\boldsymbol{\alpha}_\Theta^T \boldsymbol{\Theta}_\phi(t_{ij})^T \boldsymbol{\Theta}_\phi(t_{ij}) \boldsymbol{\Psi}_\Theta \boldsymbol{m}_i$$

$$+ \text{tr}\left[\boldsymbol{\Psi}_\Theta^T \boldsymbol{\Theta}_\phi(t_{ij})^T \boldsymbol{\Theta}_\phi(t_{ij}) \boldsymbol{\Psi}_\Theta \boldsymbol{C}_i \right] + \boldsymbol{m}_i^T \boldsymbol{\Psi}_\Theta^T \boldsymbol{\Theta}_\phi(t_{ij})^T \boldsymbol{\Theta}_\phi(t_{ij}) \boldsymbol{\Psi}_\Theta \boldsymbol{m}_i$$

where the expectation is taken with respect to the posterior distribution $\tilde{P}(\boldsymbol{c}_i | Y_i(\boldsymbol{t}_i), \xi_i(\boldsymbol{t}_i))$, using

estimates of $\boldsymbol{\alpha}_\Theta$ and $\boldsymbol{\Psi}_\Theta$ from the previous iteration.

3.1.3 Maximizing $\tilde{L}(\boldsymbol{Y}, \boldsymbol{c})$ with respect to $\boldsymbol{\alpha}_\Theta$ and $\boldsymbol{\Psi}_\Theta$

In this step we jointly estimate vectors of spline coefficients, which distinguishes our approach from previous binary PCA techniques and which entails additional complexity in the derivation of updates. The introduction of the spline basis and associated coefficients lowers the dimensionality of the estimation problem and enforces smoothness of the resulting $\hat{\mu}_i(t)$.

In order to obtain updates for our population-level basis coefficients, we introduce a new representation of the model which is mathematically equivalent to the parameterization in model (4) and easier to maximize. Let $\boldsymbol{s}_i = (\boldsymbol{c}_i^T, 1)^T$ of dimension $(K+1) \times 1$ and $\boldsymbol{\Phi} = (\boldsymbol{\Psi}_\Theta^T, \boldsymbol{\alpha}_\Theta)^T$ of dimension $(K+1) \times K_\phi$, and $vec(\boldsymbol{\Phi})$ be a vectorized version of $\boldsymbol{\Phi}$ with dimension $K_\phi(K+1) \times 1$. We can rewrite $A_i(\boldsymbol{t}_i)$ as

$$A_i(\boldsymbol{t}_i) = \boldsymbol{\Theta}_\phi(\boldsymbol{t}_i)(\boldsymbol{\alpha}_\Theta + \boldsymbol{\Psi}_\Theta \boldsymbol{c}_i) = (\boldsymbol{\Theta}_\phi(\boldsymbol{t}_i) \otimes \boldsymbol{s}_i^T)vec(\boldsymbol{\Phi})$$

where $\otimes$ is the Kronecker product. Maximizing the variational log-likelihood in this reparameterized form gives updates

$$vec(\hat{\boldsymbol{\Phi}}) = -\left[\sum_i 2\boldsymbol{\Theta}_\phi(\boldsymbol{t}_i)^T[diag(\lambda(\xi_i(\boldsymbol{t}_i)))]\boldsymbol{\Theta}_\phi(\boldsymbol{t}_i) \otimes \widehat{\boldsymbol{s}_i\boldsymbol{s}_i^T})\right]^{-1}\left[\sum_i \left(Y_i(\boldsymbol{t}_i) - \frac{1}{2}\right)^T \left(\boldsymbol{\Theta}_\phi(\boldsymbol{t}_i) \otimes \hat{\boldsymbol{s}}_i^T\right)\right]$$

where $\hat{\boldsymbol{s}}_i = (\boldsymbol{m}_i^T, 1)^T$ and

$$\widehat{\boldsymbol{s}_i\boldsymbol{s}_i^T} = \begin{pmatrix} \boldsymbol{C}_i + \boldsymbol{m}_i\boldsymbol{m}_i^T & \boldsymbol{m}_i \\ \boldsymbol{m}_i^T & 1 \end{pmatrix}$$

The first K rows of $vec\hat{\boldsymbol{\Phi}}$ are the K columns of $\hat{\boldsymbol{\Psi}}_\Theta$, and the last K_ϕ rows are $\hat{\boldsymbol{\mu}}_\Theta$.

3.2 Binary Registration

We now turn to the second step in our iterative algorithm, in which warping functions are estimated for each subject conditionally on the target function $\mu_i(t)$. Conceptually, our approach is to

maximize the exponential family likelihood function given by integrating the density in equation (2) over time. We maximize with respect to the inverse warping function $h_i^{-1}(t_i^*)$, subject to the constraint that $h_i^{-1}(t_i^*)$ is monotonic with endpoints fixed at the minimum and maximum of our time domain. For binary data we maximize the Bernoulli log-likelihood

$$l(h_i^{-1}; Y_i, \mu_i) = \int \left(Y_i(t_i^*) \log \mu_i[h_i^{-1}(t_i^*)] + (1 - Y_i(t_i^*)) \log(1 - \mu_i[h_i^{-1}(t_i^*)]) \right). \tag{8}$$

Again, functions are observed on a discrete grid in practice, and we differentiate between subject-specific finite grids for chronological time $\boldsymbol{t}_i^* = \{t_{i1}^*, ..., t_{iD_i}^*\}$ and internal time $\boldsymbol{t}_i = \{t_{i1}, ..., t_{iD_i}\}$. Using notation similar to Section 3.1, we let $Y_i(\boldsymbol{t}_i^*)$, $Y_i(\boldsymbol{t}_i)$, and $h_i^{-1}(\boldsymbol{t}_i^*)$ be $D_i \times 1$ vectors corresponding to observed responses, registered responses, and inverse warping functions, respectively. We expand $h_i^{-1}(\boldsymbol{t}_i^*)$ using a B-spline basis, $\boldsymbol{\Theta}_h(\boldsymbol{t}_i^*)$, of dimension $D_i \times K_h$. Then warping functions take the form $h_i^{-1}(\boldsymbol{t}_i^*) = \boldsymbol{\Theta}_h(\boldsymbol{t}_i^*)\boldsymbol{\beta}_i = \boldsymbol{t}_i$. The $K_h \times 1$ vector of spline coefficients $\boldsymbol{\beta}_i$ allows us to express $h_i^{-1}(\boldsymbol{t}_i^*)$, and is the target of our estimation problem. We estimate $\boldsymbol{\beta}_i$ separately for each subject using constrained optimization and iterate over subjects.

We modify the conceptual likelihood in equation (8) to incorporate the spline basis expansion of h^{-1} and to express data over the observed finite grid, which yields

$$l\left(\boldsymbol{\beta}_i; Y_i(\boldsymbol{t}_i^*), \mu_i(\cdot)\right) \propto \sum_{j=1}^{D_i} \left(Y_i(t_{ij}^*) \log \mu_i[\boldsymbol{\Theta}_h(t_{ij}^*)\boldsymbol{\beta}_i] + (1 - Y_i(t_{ij}^*)) \log(1 - \mu_i[\boldsymbol{\Theta}_h(t_{ij}^*)\boldsymbol{\beta}_i]) \right). \tag{9}$$

Recall that $\mu_i(\cdot)$ from (3) is the subject-specific mean found in the FPCA step. Estimates are constrained to be monotonic with fixed endpoints. The constraints ensure that our resulting estimates for t are monotonic and span the desired domain. We implement these constraints using linear constraint matrices, which we provide in the Appendix A.2.

The constrained optimization can be made more efficient with an analytic form of the gradient. For the general exponential family case, this gradient is

$$\frac{dl(Y_i(\boldsymbol{t}_i^*), \boldsymbol{\beta}_i)}{d\boldsymbol{\beta}_i} = \frac{1}{\varphi} \sum_{j=1}^{D_i} \left[\left(Y_i(t_{ij}^*) - b'\left(g((\mu_i(t_{ij})) \right) \right) \times \Theta_h(t_{ij}^*)^T \Theta'_\phi(\boldsymbol{\Theta}_h(t_{ij}^*)\boldsymbol{\beta}_i) \left(\boldsymbol{\alpha}_\Theta + \boldsymbol{\Psi}_\Theta \boldsymbol{c}_i \right) \right], \quad (10)$$

where $\boldsymbol{\Theta}'_\phi(\boldsymbol{t}_i)$ is a $D_i \times K_h$ matrix of first derivatives of the B-spline basis functions used to reconstruct $\boldsymbol{t}_i$, and $b'\left(g((\mu_i(t_{ij})) \right) = \mu_i(t_{ij}) = g^{-1}(\Theta_\phi(\boldsymbol{\Theta}_h(t_{ij}^*)\boldsymbol{\beta}_i)(\boldsymbol{\alpha}_\Theta + \boldsymbol{\psi}_\Theta \boldsymbol{c}_i))$. For the Bernoulli loss function $\varphi = 1$ and $g^{-1}(z) = \frac{1}{1+e^{-z}}$, so the gradient becomes

$$\frac{dl(Y_i(\boldsymbol{t}_i^*), \boldsymbol{\beta}_i)}{d\boldsymbol{\beta}_i} = \sum_{j=1}^{D_i} \left[\left(Y_i(t_{ij}^*) - \frac{1}{1 + e^{-\Theta_\phi(\boldsymbol{\Theta}_h(t_{ij}^*)\boldsymbol{\beta}_i)(\boldsymbol{\alpha}_\Theta + \boldsymbol{\psi}_\Theta \boldsymbol{c}_i)}} \right) \times \Theta_h(t_{ij}^*)^T \Theta'_\phi(\boldsymbol{\Theta}_h(t_{ij}^*)\boldsymbol{\beta}_i) \left(\boldsymbol{\alpha}_\Theta + \boldsymbol{\Psi}_\Theta \boldsymbol{c}_i \right) \right].$$
$$(11)$$

The addition of this analytic derivative drastically improves the performance of the warping function estimation process.

3.3 Implementation

Our methods are implemented in `R` and are publicly available on GitHub as part of the `registr` package. For Step 1, binary FPCA is custom-written with a `C++` backend for estimation. For Step 2, we implement linearly constrained optimization with the `constrOptim()` function, which uses an adaptive barrier method. The default for `constrOptim()` in `R` uses a numeric method to calculate the gradient of the loss function; in place of this we use the analytic gradient in equation (11) to improve accuracy and computational efficiency. Though our simulated and real data examples are observed on a dense regular grid, the `registr` package handles both sparse and irregular functional data. For visualizing results, `registr` is compatible with the developer version of `refund.shiny`, an `R` package that produces interactive graphics for functional data analyses (Wrobel et al., 2016).

4 Simulations

We assess the accuracy and computational efficiency of our method using data simulated to mimic those in our motivating study, and compare to a competing approach described below.

4.1 Simulation Design

Binary functions in simulated datasets are designed to exhibit a circadian rhythm, so that simulated participants are more likely to be inactive at the beginning and end of the observation domain ("day") and more likely to be active in the middle of the day. Overall activity levels vary across simulated participants, as do the timing of the active period.

We then generate a grid of chronological times $\boldsymbol{t}_i^*$, which is equally spaced and shared across subjects. We generate inverse warping functions $h_i^{-1}(\boldsymbol{t}_i^*)$ using a B-spline basis with 3 degrees of freedom; coefficients are chosen from a uniform(0,1) distribution and placed increasing order to ensure monotonically increasing warping functions. The internal times $\boldsymbol{t}_i$ for each subject are obtained by evaluating the inverse warping functions at $\boldsymbol{t}_i^*$. We simulate latent probability curves over internal time using

$$E\Big[Y_i(\boldsymbol{t}_i)|c_i\Big] = \mu_i(\boldsymbol{t}_i)$$
$$g[\mu_i(\boldsymbol{t}_i)] = \alpha(\boldsymbol{t}_i) + c_i \times \psi(\boldsymbol{t}_i)$$

(12)

where $\alpha(\boldsymbol{t}_i) = 1.5 * (0 - \sin(2\pi\boldsymbol{t}_i) - \cos(2\pi\boldsymbol{t}_i)$, $\psi(\boldsymbol{t}_i) \propto (0 - \sin(2\pi\boldsymbol{t}_i) - \cos(2\pi\boldsymbol{t}_i))$, and $c_i \sim N(0,15)$. Binary observations $Y_i(\boldsymbol{t}_i)$ are sampled from a Bernoulli distribution using the probabilities $\mu_i(\boldsymbol{t}_i)$. Unregistered data $Y_i(\boldsymbol{t}^*)$, observed over the grid $\boldsymbol{t}_i^*$, are defined by the warping functions $h_i(\boldsymbol{t}_i)$. Figure 3 shows an example of a single simulated dataset, including latent probability curves on both $\boldsymbol{t}_i^*$ and $\boldsymbol{t}$ (first row, first and second columns) and observed binary data (second row, first and second columns).

We evaluate the performance of our algorithm as a function of sample size and grid length. We simulate 50 datasets for each combination of sample sizes (50, 100, and 200) and grid lengths (taking values 100, 200, 400). For each simulated dataset we apply the methods in Section 3, denoted *registr* in text and figures below, setting $K_\phi = 5$, $K_h = 3$, and using 1 FPC. To provide a frame of reference we compare to an approach that registers pre-smoothed data. We first smooth the binary data using a generalized additive model, implemented using the `gam()` function from the `mgcv` package (Wood,

15

2017), then apply the inverse link function to obtain estimated probability curves, and register these using the default settings for the `fdasrvf::time_warping()` function (Tucker, 2017). This is analogous to the approach in Wu and Srivastava (2014), and is denoted *fdasvrf* in text and figures below. Methods are compared in terms of estimation accuracy and computation time, with accuracy quantified using mean integrated squared error (MISE). For each subject, integrated squared error calculations are made comparing the estimated inverse warping functions for each method, $\hat{h}_i^{-1}(\boldsymbol{t}_i^*)$, to the true inverse warping functions $h_i^{-1}(\boldsymbol{t}_i^*)$ such that $ISE = \int_0^1 \left(h_i^{-1}(\boldsymbol{t}_i^*) - \hat{h}_i^{-1}(\boldsymbol{t}_i^*) \right)^2 dt_i$. MISE is then the average of ISEs across subjects. A sensitivity analysis of our method's performance across values of K_ϕ and K_h is given in Appendix B.1. Code for our simulations is publicly available.

4.2   Simulation Results

Figure 3 shows a simulated dataset with 100 subjects observed over a grid with 200 time points. From left to right, columns show observed (unregistered) data, data aligned using the true internal time t, data aligned using the estimated internal time $\hat{t}_{registr}$ obtained from the *registr* method, and data aligned using the estimated internal time $\hat{t}_{fdasvrf}$ obtained from the *fdasvrf* method. The top row shows the latent mean curves, the middle row shows lasagna plots of observed binary data, and the bottom row shows inverse warping functions using t, $\hat{t}_{registr}$, and $\hat{t}_{fdasvrf}$. The latent probability curves, on clock and internal time, illustrate the simulation design in terms of the relative magnitudes of phase and amplitude variability. The binary curves illustrate the observed data, which includes a period of higher activity for each subject. The results for *registr* are encouraging, both for the latent curves and for the binary activity data in that phase variation is largely removed. Some amount of misalignment remains, which is attributable to the inherent sampling variability introduced when binary points are generated from the latent probabilities. The *fdasrvf* method also works reasonably well, although visual inspection of the probability curves and binary data suggests somewhat poorer alignment.

Figure 4 summarizes results across simulated datasets at different sample sizes and grid lengths; for reference, the data in Figure 3 has a median MISE for the *registr* method relative to other
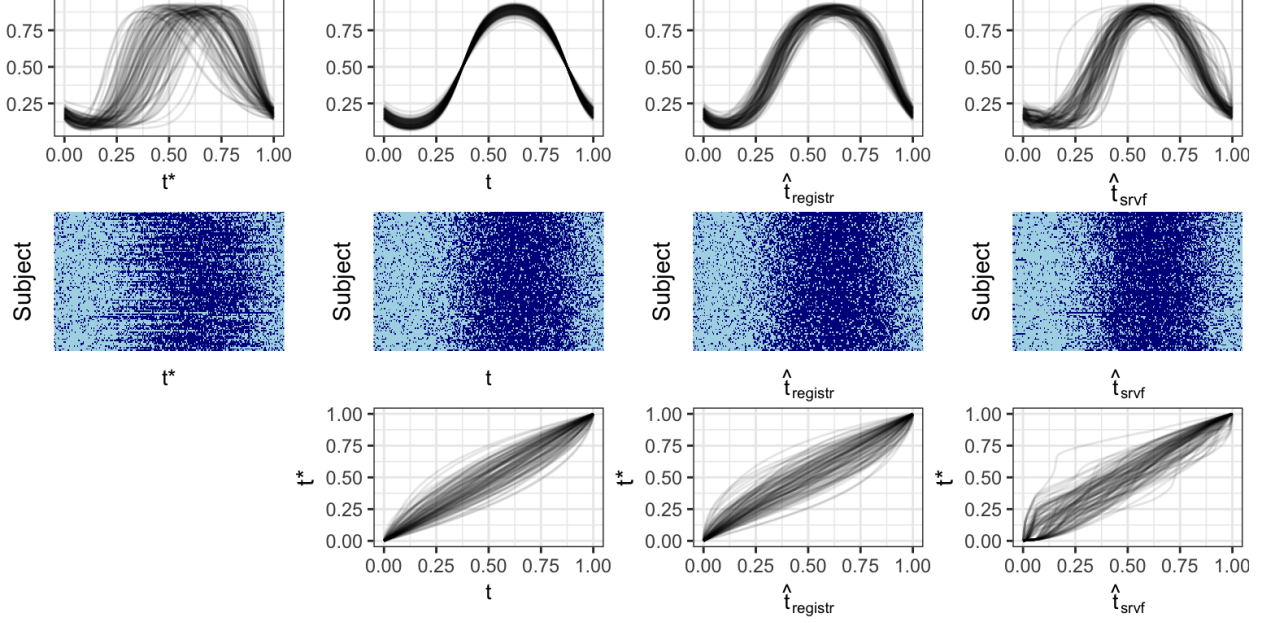
Figure 3: For top and center rows, from left to right we have: unregistered curves, curves registered using true inverse warping functions, curves registered using *registr* method, curves registered using *fdasrvf* method. The top row shows the true latent probability curves which are used to generate the binary curves but not used to estimate warping since they are unknown in a real data application. The middle row shows the binary curves as a lasagna plot, as in Figure 2. The bottom row shows the true, *registr* method, and *fdasrvf* method inverse warping functions.

datasets generated with 100 subjects and 200 time points. The columns of Figure 4, from left to right, show results for datasets with 50, 100, and 200 subjects, respectively, and grid lengths of 100, 200, and 400 are shown within each panel. The top row shows box plots of MISE and the bottom row shows median computation times. Across all settings, *registr* outperforms *fdasrvf* in terms of the MISE; this is consistent with observations in Figure 3. With respect to computation time, although the methods are similar for small sample sizes and grid lengths, *registr* scales well in both of these, while the burden grows dramatically for *fdasvrf*. Indeed, *fdasrvf* may be prohibitively expensive for our real data.

5 Analysis

We now apply our method described in Section 3 to the BLSA data. These data contain 592 subjects with activity counts every minute over 24 hours, for a total of 1440 measurements per
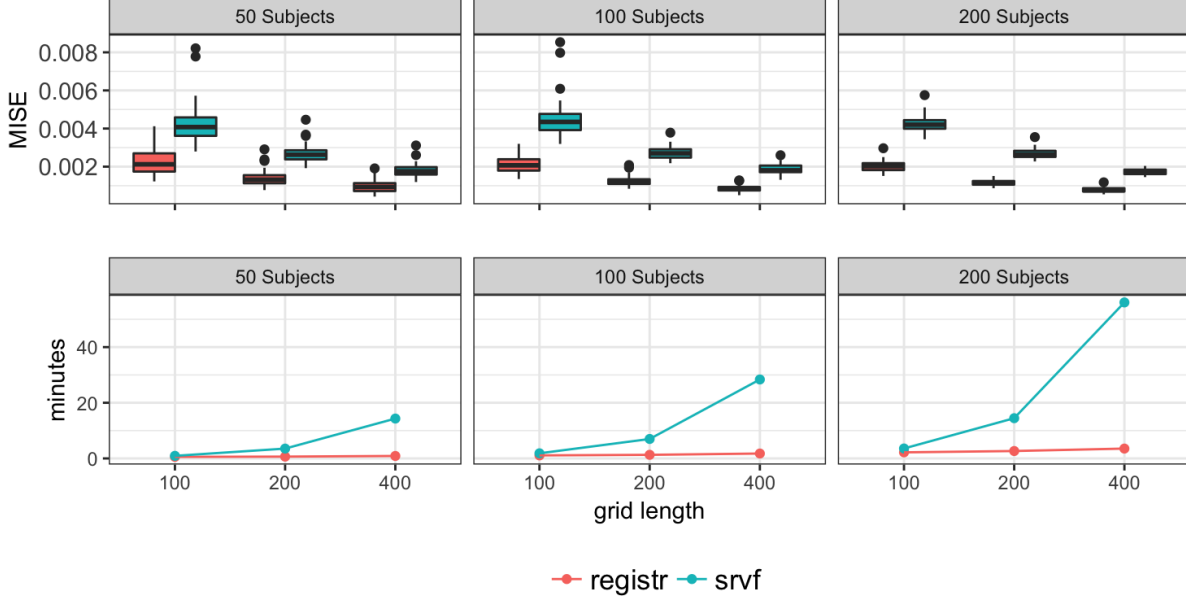
Figure 4: This figure shows mean integrated squared errors (top row) and median computation times (bottom row) for *registr* (in red) and *fdasrvf* (in green) methods across varying sample sizes and grid lengths. The columns, from left to right, show sample sizes 50, 100, and 200, respectively. Within each panel we compare grid lengths of 100, 200, and 400.

subject. BLSA participants wore the accelerometer for 5 days; we average across these days to establish a typical diurnal pattern for each participant, and then threshold the result at values of 20 counts per minute to obtain the binary activity curve to be registered. We fix the dimensions of the B-spline basis function to $K_\phi = 8$ and $K_h = 4$ and number of functional principal components to $K = 2$. Total computation time for this analysis was 17 minutes.

Figure 5 shows the registered curves from the BLSA dataset, which can be compared with the observed data in Figure 2. After registration, there are two clear activity peaks: people tend to be active for an extended period of time after they wake up; this period is followed by a mid-day dip in activity, and a second, smaller, period of activity in the afternoon and evening. Figure 6 emphasizes this point, and the effect of registration, by plotting the subjects from Figure 1 after registration. The data for these two subjects are more closely aligned, as are the latent probabilities curves estimated from the aligned data. The left panel of Figure 5 shows the inverse warping functions which transform the BLSA data from the unregistered to the registered space.

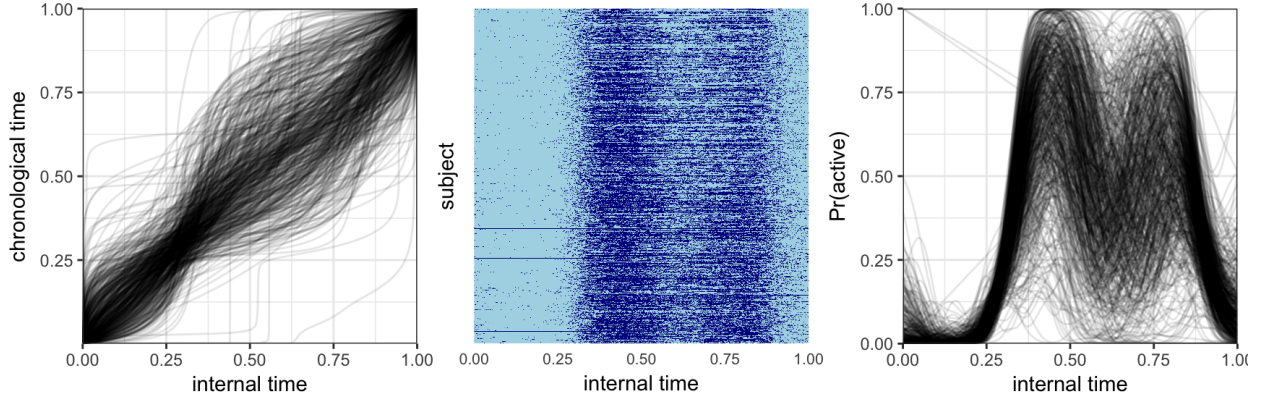The results of the applying the registration method to these data are consistent with expecta-

Figure 5: Plots of the registered BLSA data. Left panel shows inverse warping functions from alignment of the data; center panel shows a lasagna plot of the aligned binary data; and right panel shows smooths of the aligned data. See Figure 2 for the unregistered data.

tions, in that the diurnal activity pattern observed across subjects after registration contains both morning and afternoon active periods and a period of relative inactivity around lunchtime. These results also emphasize the importance of assessing and removing phase variability in studies of daily activity patterns. Figure (7) shows the effects of the estimated principal component basis functions for the BLSA data after the registration process. The first principal component is a vertical shift around the population mean, $\alpha(t)$, indicating a higher or lower probability of being active. More interesting is the second principal component, which shows that some subjects have higher probability of activity earlier in the day, while others have higher probability of activity later in the day. The existence and number of "chronotypes", or subjects who intrinsically prefer certain hours of the day (like the colloquial night owls or early birds), is the subject of intense debate in the circadian rhythm literature (Adan et al., 2012; Putilov et al., 2015). Aligning observed activity data as a processing step may help inform this debate, and our results are consistent with the existence of distinct chronotypes in this population.

Appendix B.2 contains additional analysis results for the registration of data from each day of the week separately. These additional results and their interpretation are similar to those presented
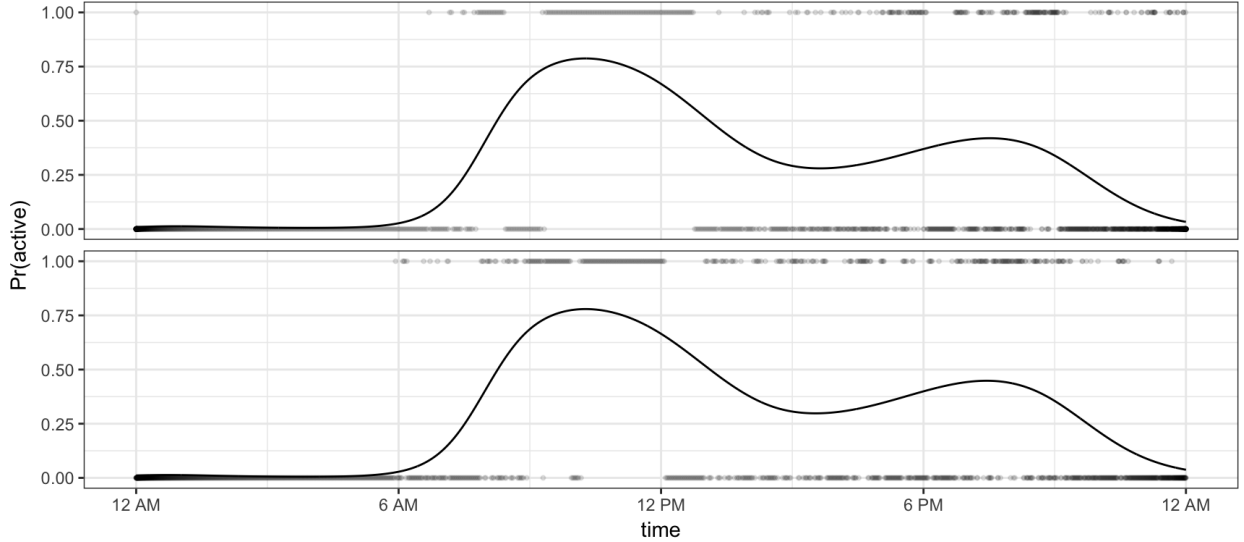
Figure 6: These are binary curves for the same two subjects from the BLSA data as in Figure 1 but now the curves are registered. Here the lines represent estimates of the latent probability that come from our binary FPCA algorithm.

in this Section.

6   Discussion

We present a novel approach to curve registration for functional data from exponential family distributions which avoids the need for pre-smoothing, and our attention to computational efficiency is necessitated by our data application. Simulations suggest our approach compares favorably to competing methods in the settings we examined. Our scientific results are plausible and meaningful in the context of activity measurement. Finally, our code for registration and binary probabilistic FPCA is publicly available as part of the `registr` package.

Because of the nature of our application, we optimize performance for registering binary curves. While our method can be applied to functional data from any exponential family, one will not reap the computational benefits we highlight here without at least some additional work optimizing the FPCA algorithm for additional distributions. In particular, we expect that computationally efficient implementations for the Poisson distribution will be relevant for studies of activity intensity using accelerometer data. For our application, we chose to threshold activity count data and
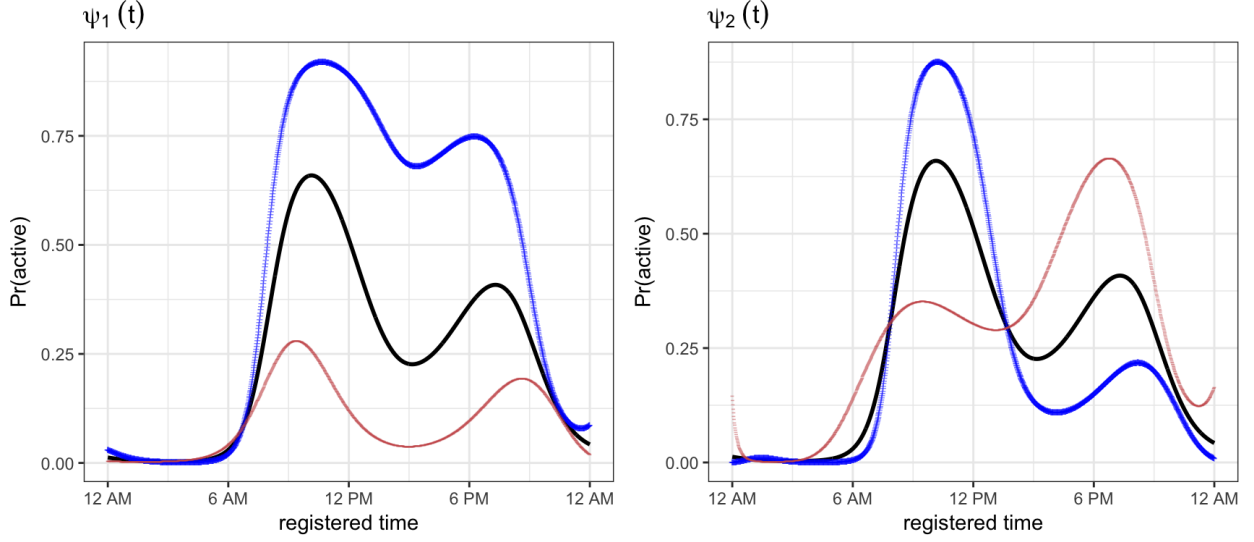
Figure 7: Estimated binary FPCA basis functions after registration process, illustrated by plotting $g^{-1}\left[\alpha(t) \pm \psi_k(t)\right]$ for basis functions $k \in \{1, 2\}$.

register the resulting binary curves, though we could have chosen to register the raw counts using a Poisson distribution. Exploratory analyses suggested that aligning raw activity counts may be overly influenced by extreme values (during intense activity, for example) rather than aligning the general patterns of activity and inactivity which more accurately reflect circadian rhythms. It is worth noting, however, that registering activity count data is made possible by the methodological developments described in this paper.

Though we focus on amplitude alignment for this paper, the inverse warping functions contain information on phase variation and are potential analysis objects of interest for future scientific work. Subsequent analyses will examine whether aligned data are more clearly affected by covariates like age and sex, and how the phase alignment relates to these covariates. Understanding patterns of phase variation across days within the same subject may be of interest as well, although doing so will necessitate incorporating a multilevel structure into the preceding methods. Finally, we note that our emphasis has been on the temporal structure of inactivity, and additional work to connect these results with the accrual of sedentary minutes in bouts is needed.

References

Adan, A., Archer, S. N., Hidalgo, M. P., Di Milia, L., Natale, V., and Randler, C. "Circadian typology: a comprehensive review." Chronobiology international, 29(9):1153–1175 (2012).

Diaz, K., Howard, V., Hutto, B., and et al. "Patterns of sedentary behavior and mortality in U.S. middle-aged and older adults: A national cohort study." Annals of Internal Medicine (2017).

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

Gertheiss, J., Goldsmith, J., and Staicu, A.-M. "A note on modeling sparse exponential-family functional response curves." Computational Statistics and Data Analysis, 105:46–52 (2017).

Gertheiss, J., Maier, V., Hessel, E., and Staicu, A.-M. "Marginal Functional Regression Models for Analyzing the Feeding Behavior of Pigs." Journal of Agricultural, Biological, and Environmental Statistics,, To Appear (2015).

Goldsmith, J., Greven, S., and Crainiceanu, C. M. "Corrected Confidence Bands for Functional Data using Principal Components." Biometrics, 69:41–51 (2013).

Goldsmith, J., Zipunnikov, V., and Schrack, J. "Generalized multilevel function-on-scalar regression and principal component analysis." Biometrics, 71:344–353 (2015).

Hadjipantelis, P. Z., Aston, J. A. D., Müller, H.-G., and Evans, J. P. "Unifying Amplitude and Phase Analysis: a Compositional Data Approach to Functional Multivariate Mixed-Effects Modeling of Mandarin Chinese." Journal of the American Statistical Association, 110:545–559 (2015).

Hall, P., Müller, H.-G., and Yao, F. "Modelling sparse generalized longitudinal observations with latent Gaussian processes." Journal of the Royal Statistical Society: Series B, 70:703–723 (2008).

Huang, H., Yehua, L., and Guan, Y. "Joint Modeling and Clustering Paired Generalized Longitudinal Trajectories with Application to Cocaine Abuse Treatment Data." Journal of the American Statistical Association, 109.508:1412–1424 (2014).

Jaakkola, T. S. and Jordan, M. I. "A variational approach to Bayesian logistic regression models and their extensions." In Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics (1997).

James, G. M., Hastie, T. J., and Sugar, C. A. "Principal component models for sparse functional data." Biometrika, 87:587–602 (2000).

Kneip, A. and Ramsay, J. O. "Combining Registration and Fitting for Functional Models." Journal of the American Statistical Association, 103:1155–1165 (2008).

Marron, J. S., Ramsay, J. O., Sangalli, L. M., and Srivastave, A. "Functional Data Analysis of Amplitude and Phase Variation." Statistical Science, 30(4):468–484 (2015).

Martin, K. R., Koster, A., Murphy, R. A., Van Domelen, D. R., Hung, M. Y., Brychta, R. J., Chen, K. Y., and Harris, T. B. "Changes in Daily Activity Patterns with Age in U.S. Men and Women: National Health and Nutrition Examination Survey 2003?04 and 2005?06." Journal of the American Geriatrics Society, 62(7):1263–1271 (2014).
URL http://dx.doi.org/10.1111/jgs.12893

Panaretos, V. M. and Zemel, Y. Z. "Amplitude and phase variation of point processes." The Annals of Statistics, 44:771–812 (2016).

Putilov, A. A., Donskaya, O. G., and Verevkin, E. G. "How many diurnal types are there? A search for two further ?bird species?" Personality and Individual Differences, 72:12–17 (2015).

Ramsay, J. O. and Li, X. "Curve registration." Journal of the Royal Statistical Society: Series B (Statistical Methodology), 60:351–363 (1998).

Ramsay, J. O. and Silverman, B. W. Functional Data Analysis. New York: Springer (2005).

Sakoe, H. and Chiba, S. "Dynamic programming algorithm optimization for spoken word recognition." IEEE transactions on acoustics, speech, and signal processing, 26(1):43–49 (1978).

Sangalli, L. M., Secchi, P., Vantini, S., and Vitelli, V. "k-Mean Alignment for Curve Clustering." Computational Statistics & Data Analysis, 54:1219–1233 (2010).

Schein, A. I., Saul, L. H., and Ungar, A. "A generalised linear model for principal component analysis of binary data." In Proceedings of the 9th International Workshop on Artificial Intelligence and Statistics (2003).

Schrack, J. A., Zipunnikov, V., Goldsmith, J., Bai, J., Simonshick, E. M., Crainiceanu, C. M., and Ferrucci, L. "Assessing the "Physical Cliff": Detailed Quantification of Aging and Physical Activity." Journal of Gerontology: Medical Sciences (2014).

Serban, N., Staicu, A.-M., and Carrol, R. J. "Multilevel Cross-Dependent Binary Longitudinal Data." Biometrics, 69:903–913 (2013).

Srivastava, A., Wu, W., Kurtek, S., Klassen, E., and Marron, J. S. "Registration of Functional Data Using Fiser-Rao Metric." arXiv preprint arXiv, 1103.3817 (2011).

Swihart, B., Caffo, B., James, B. D., Schwartz, B. S., and Punjabi, N. M. "Lasagna plots: A saucy alternative to spaghetti plots." Epidemiology, 21(5):621–625 (2010).

Tipping, M. E. "Probabilistic Visualisation of High-dimensional binary data." Advances in neural information processing systems, 592–598 (1999).

Tipping, M. E. and Bishop, C. "Probabilistic Principal Component Analysis." Journal of the Royal Statistical Society: Series B, 61:611–622 (1999).

Tucker, J., Wu, W., and Srivastava, A. "Analysis of proteomics data: Phase amplitude separation using an extended Fisher-Rao metric." Electronic Journal of Statistics, 8:1724–1733 (2014).

Tucker, J. D. fdasrvf: Elastic Functional Data Analysis (2017). R package version 1.8.1.
  URL http://CRAN.R-project.org/package=fdasrvf

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

van der Linde, A. "Variational Bayesian Functional PCA." Computational Statistics and Data Analysis, 53:517–533 (2008).

Vantini, S. "On the definition of phase and amplitude variability in functional data analysis." TEST, 21(4):676–696 (2012).

Wood, S. mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation (2017). R package version 1.8-22.
  URL http://CRAN.R-project.org/package=mgcv

Wrobel, J., Young Park, S., Maria Staicu, A., and Goldsmith, J. "Interactive graphics for functional data analyses." Stat, 5:108–118 (2016).

Wu, W. and Srivastava, A. "Analysis of spike train data: Alignment and comparisons using the extended Fisher-Rao metric." Electronic Journal of Statistics, 8:1776–1785 (2014).

Yao, F., Müller, H., and Wang, J. "Functional data analysis for sparse longitudinal data." Journal of the American Statistical Association, 100(470):577–590 (2005).

Yerrakalva, D., Cooper, A. J., Westgate, K., Khaw, K. T., Wareham, N. J., Brage, S., Griffin, S. J., and Wijndaele, K. "The descriptive epidemiology of the diurnal profile of bouts and breaks in sedentary time in older English adults." International Journal of Epidemiology, dyx123 (2017).
  URL +http://dx.doi.org/10.1093/ije/dyx123

A   Appendix: methods

Here we provide extra details of our methods.

A.1   Updating $\boldsymbol{\alpha}_\Theta$ and $\boldsymbol{\Psi}_\Theta$ for binary FPCA

Recall that we obtain parameter updates by maximizing the variational likelihood given in equation (7). In section 3.1.3 we obtain updates for $\boldsymbol{\alpha}_\Theta$ and $\boldsymbol{\Psi}_\Theta$ by reparameterizing equation (7) such that $\boldsymbol{\Phi} = (\boldsymbol{\Psi}_\Theta^T, \boldsymbol{\alpha}_\Theta)^T$. This reparameterization leads to the variational log-likelihood below:

$$
\begin{aligned}
\tilde{l}(\boldsymbol{Y}, \boldsymbol{c}) \;\propto\; & \sum_{j=1}^{D_i}\sum_{i=1}^{I} \log \tilde{P}\!\left(Y_i(t_{ij})|\boldsymbol{c}_i, \xi_i(t_{ij})\right) - \sum_i \boldsymbol{c}_i^T \boldsymbol{c}_i \\
\propto\; & \sum_i \left(\left(Y_i(\boldsymbol{t}_i) - \frac{1}{2}\right)^T A_i(\boldsymbol{t}_i) - \frac{1}{2}\xi_i(\boldsymbol{t}_i)\mathbb{1}_{D_i\times 1} + A_i^T(\boldsymbol{t}_i)[diag\big(\lambda\left(\xi_i(\boldsymbol{t}_i)\right)\big)]A_i(\boldsymbol{t}_i)\right) \\
\propto\; & \sum_i \left(Y_i(\boldsymbol{t}_i) - \frac{1}{2}\right)^T (\boldsymbol{\Theta}_\phi(\boldsymbol{t}_i) \otimes \boldsymbol{s}_i^T)vec(\boldsymbol{\Phi}) \\
+\; & \sum_i vec(\boldsymbol{\Phi})^T (\boldsymbol{\Theta}_\phi(\boldsymbol{t}_i)^T \otimes \boldsymbol{s}_i)[diag\big(\lambda\left(\xi_i(\boldsymbol{t}_i)\right)\big)](\boldsymbol{\Theta}_\phi(\boldsymbol{t}_i) \otimes \boldsymbol{s}_i^T)vec(\boldsymbol{\Phi}).
\end{aligned}
$$

Maximizing with respect to $\boldsymbol{\Phi}$ gives estimates $\hat{\boldsymbol{\Phi}}$.

A.2   Optimization constraints for the warping step

Section 3.3 refers to optimization constraints for the R function `constrOptim()` implemented in the warping step of our algorithm. We constrain inverse warping function to be monotonic with fixed endpoints, and these constraints are enforced through $\boldsymbol{\beta}_i$, the warping function B-spline coefficients for each subject. To ensure that estimated inverse warping functions $\hat{h}_i^{-1}$ span the same domain as chronological time t_i^*, we fix the outer coefficients $\beta_{i,1}$ and β_{i,K_h}. Thus in practice we estimate the $K_h - 2$ inner spline coefficients $\boldsymbol{\beta}_{i,inner} = (\beta_{i,2}, ..., \beta_{i,K_h-1})^T$.

To enforce monotonicity of the warping functions we must ensure $\beta_2 < \beta_2 < ... < \beta_{K_h-1}$. Using the notation from the `constrOptim()` function, we define a matrix ui and a vector ci such that

$$ui \times \boldsymbol{\beta}_{i,inner} - ci \geq 0. \tag{A.1}$$

This leads to a ui matrix of dimension $(K_h - 1 \times K_h - 2)$ and a size $K_h - 1$ vector, ci that take the forms:

$$ui = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & & & -1 & 1 \\ 0 & & & 0 & -1 \end{pmatrix}$$

and

$$ci = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ -1 \end{pmatrix}$$

such that

$$\begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & & & -1 & 1 \\ 0 & & & 0 & -1 \end{pmatrix} \begin{pmatrix} \beta_{i,2} \\ \beta_{i,3} \\ \vdots \\ \beta_{i,K_h-1} \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ -1 \end{pmatrix} > \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

B Appendix: simulations and analysis

Here we provide extra results from simulations and analysis of BLSA data.

B.1 Optimizing parameters

As a sensitivity analysis we evaluate our method as a function of parameters K_ϕ and K_h. We evaluated all combinations of $K_\phi \in \{5, 10, 15\}$, $K_h \in \{3, 4, 5, 6\}$ and grid length $D \in \{50, 100, 200\}$ using the same simulation setup and performance metrics as in Section 4. Mean integrated squared errors are given in Figure (B.1) and computation times across these simulation scenarios are given in Figure (B.2).
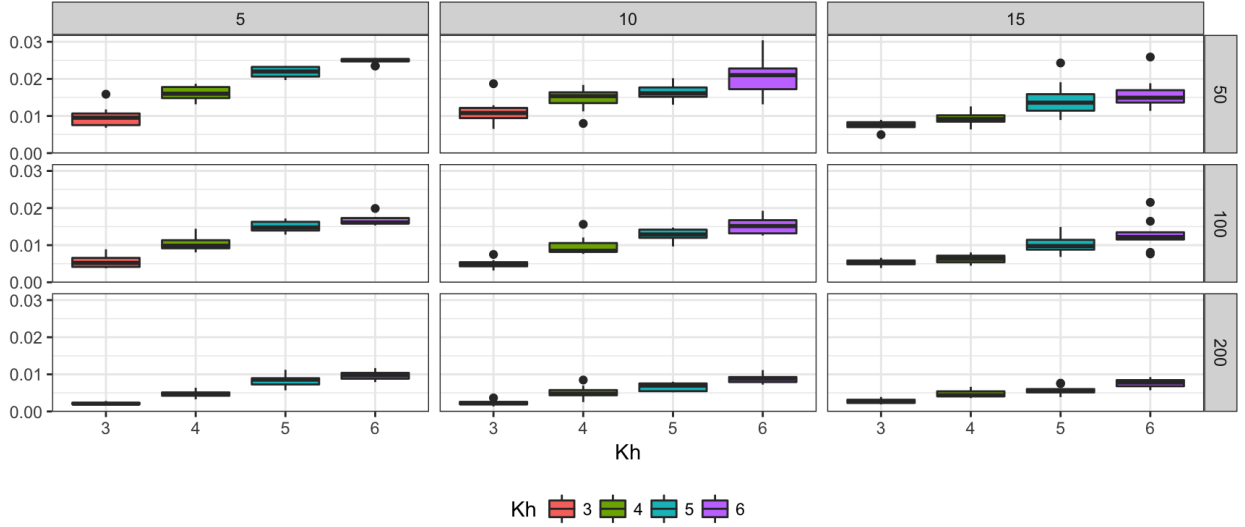

Figure B.1: Parameter sensitivity across values of K_ϕ and K_h for *registr* method. Shown are mean integrated squared error (MISE) summaries across 10 datasets for each parameter scenario. Columns represent distinct values of K_ϕ and rows distinct grid lengths D.

Both MISE and computation time increase linearly with K_h. Mean integrated squared errors decrease slightly with increasing K_ϕ, and computation time slightly increases with increasing K_ϕ.

B.2 Analysis of weekdays for BLSA

In our primary analysis we averaged across visits for subjects. Here we separate visits by day of the week and look at day-specific effects. Figure B.3 shows unregistered and registered binary and smooth curves for each day of the week. Our algorithm consistently identifies similar patterns across days of the week. Alignment may be slightly better on week days than weekends, which suggests an area for future exploration.
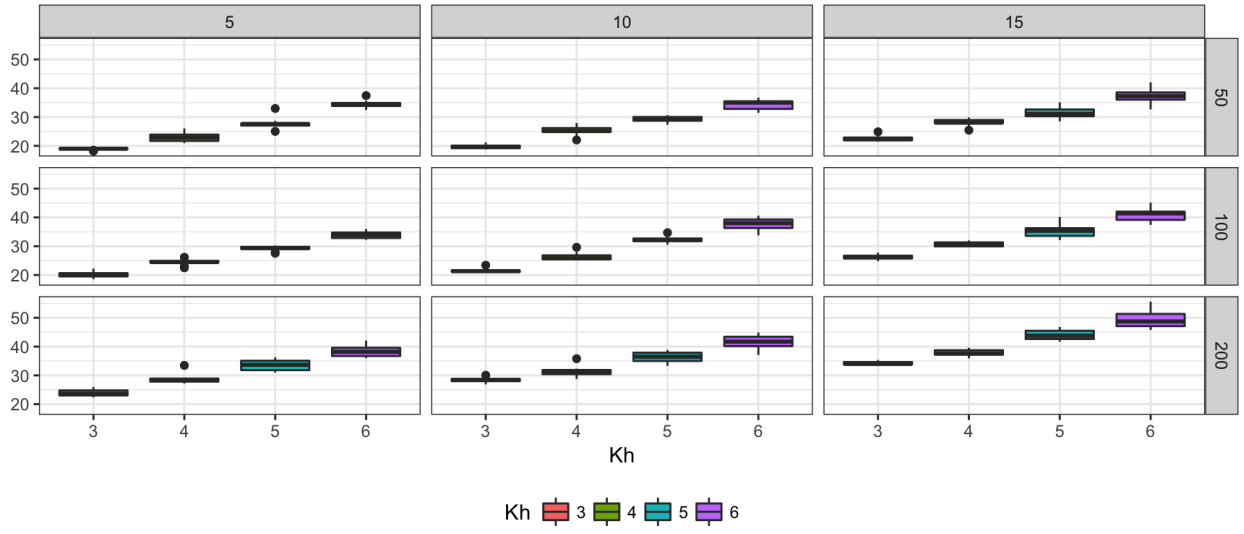
Figure B.2: Parameter sensitivity across values of K_ϕ and K_h for *registr* method. Shown are boxplots of computation time (in seconds) across 10 datasets for each parameter scenario. Columns represent distinct values of K_ϕ and rows distinct grid lengths D.
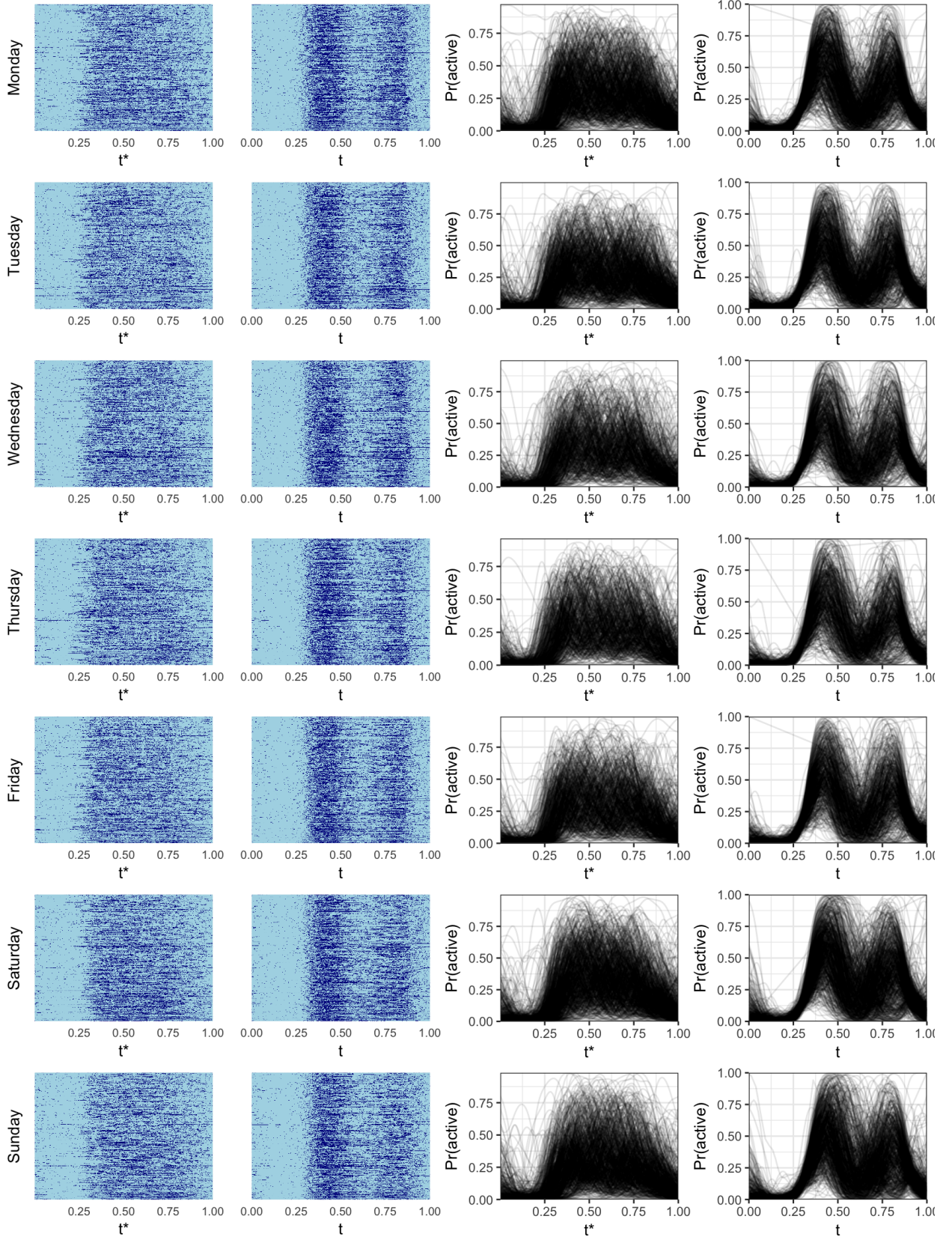
Figure B.3: Analysis results for each day of the week.