Stat

The ISI's Journal for the Rapid
Dissemination of Statistics Research

(wileyonlinelibrary.com) DOI: 10.1002/sta4.121

Longitudinal functional additive model with continuous proportional outcomes for physical activity data

Haocheng Li[a]*, Sarah Kozey Keadle[b], Victor Kipnis[c] and Raymond J. Carroll[d,e]

Received 1 September 2016; Accepted 9 September 2016

Motivated by physical activity data obtained from the BodyMedia FIT device (http://www.bodymedia.com), we take a functional data approach for longitudinal studies with continuous proportional outcomes. The functional structure depends on three factors. In our three-factor model, the regression structures are specified as curves measured at various factor points with random effects that have a correlation structure. The random curve for the continuous factor is summarized using a few important principal components. The difficulties in handling the continuous proportion variables are solved by using a quasilikelihood-type approximation. We develop an efficient algorithm to fit the model, which involves the selection of the number of principal components. The method is evaluated empirically by a simulation study. This approach is applied to the BodyMedia data with 935 males and 84 consecutive days of observation, for a total of 78,540 observations. We show that sleep efficiency increases with increasing physical activity, while its variance decreases at the same time.

Keywords: BodyMedia FIT device; continuous proportions; functional data; mixed effects model; physical activity; sleep efficiency

1 Introduction

Motivated by physical activity data, we take a functional data approach for longitudinal studies on continuous proportional outcomes. The research is aimed at understanding the influence of physical activity on sleep efficiency. The response variable, sleep efficiency, is measured by the ratio of daily sleep time to lying down time for each participant (Lambiase et al., 2013), and thus, it is a continuous proportion. The major explanatory factor, physical activity level, is measured by the daily minutes of moderate to vigorous physical activity. The intensity of physical activity is evaluated in a unit called METs, with 1 MET being the energy required to sit quietly, a quantity that depends on one's body weight and other characteristics. Moderate to vigorous physical activity has 3–6 METs and is roughly when a person is moving fast enough or strenuously enough to burn off three to six times as much energy per minute as when she or he

[a]Departments of Oncology and Community Health Sciences, University of Calgary, Calgary, AB T2N 1N4, Canada
[b]Kinesiology Department, California Polytechnic State University, San Luis Obispo, CA 93407, USA
[c]Biometry Research Group, DCP, National Cancer Institute, Bethesda, MD 20892, USA
[d]Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143, USA
[e]School of Mathematical and Physical Sciences, University of Technology Sydney, Broadway, NSW 2007, Australia
*Email: haocheng.li@ucalgary.ca

is sitting quietly. Month and weekday effects may also influence sleep efficiency, and we take them into account as two additional factors. Therefore, our functional structure depends on physical activity, month and weekday effects factors.

It has been widely reported that greater physical activity leads to greater sleep efficiency, both marginally (Lambiase et al., 2013; Oudegeest-Sander et al., 2013) and, in longitudinal data, within-person (Ekstedt et al., 2013). However, these studies mainly use correlation-based or linear regression methods, which do not take advantage of newly developed instruments, such as the BodyMedia FIT device (bodymedia.com), the ActiGraph device (actigraphcorp.com) and the ActivPal device (paltechnologies.com). These devices can measure physical activity data continuously (e.g. minute-by-minute) for an extended period (e.g. weeks). In this study, we use data from the BodyMedia FIT device. The BodyMedia FIT device is a multi-sensor armband that measures skin temperature, heat flux, galvanic skin response, and motion through a three-axis accelerometer. With this device, there is the opportunity for new and more powerful statistical methods to understand physical activity and other outcomes, both within and between individuals.

The data we have are summaries of daily minutes of moderate to vigorous physical activity (MVPA) and daily sleep efficiency over 12-week periods with starting dates distributed throughout the calendar year. There are numerous questions that arise, including the following:

- What is the effect of daily MVPA minutes on daily sleep efficiency?
- If daily MVPA minutes lead to greater sleep efficiency, is the effect constant or is there some point where the effect of increasing MVPA plateaus or even decreases?
- Does increasing MVPA minutes have influence on the stability (variability) of sleep efficiency?

The purpose of this paper is to develop initial methods that can be used to answer the questions and conjectures described earlier. We do not claim that the methods are the last word in the analysis of physical activity and sleep efficiency, but we believe that our work is novel because we take a functional data analysis approach to the question, while simultaneously recognizing explicitly that sleep efficiency is a variable that necessarily is constrained to the unit interval $(0, 1)$, a constraint that, to the best of our knowledge, has not been considered in the functional data literature.

There are many studies for continuous proportional responses, but not for longitudinal functional data. One naive way is to ignore that the proportional outcomes should be in the unit interval $(0, 1)$, and then fit the responses by ordinary linear regression, but this potentially gives predictions outside the unit interval (Kieschnick & McCullough, 2003). The beta distribution (Ferrari & Cribari-Neto, 2004) can also be employed to analyze the data. Simas et al. (2010) and Zhao et al. (2012) used beta regression with functional forms for predictors, but they were limited by assuming that all observations are independent. Verkuilen & Smithson (2012) and Figueroa-Zúñiga et al. (2013) used mixed effects in the beta regression to model correlated data, but their fixed and random effects structures were not in functional formulations. On the other hand, continuous proportions can be analyzed by first taking the logit transformation of the outcomes and then using linear regression to fit the data. However, as we will show, this approach can lead to biased results.

There are of course many statistical papers that focus on functional data analysis, but as far as we are aware, there are only a few studies that focus on proportional responses. Hall et al. (2008) proposed a functional model for non-Gaussian data. However, the model estimation is partly based on a simplified first-order linear approximation formulation and is known to lead to biased results in some settings (Serban et al., 2013). Goldsmith et al. (2015) proposed a generalized multilevel function-on-scalar regression model for outcomes with exponential family. Gertheiss et al. (2015) discussed a marginal functional regression model for binary outcomes, and Scheipl et al. (2016) studied generalized functional additive mixed models for outcomes with exponential family distribution as well as others like beta distribution. However, these models cannot handle the effects from the three factors in our data.

In the case of correlated functional data, a large number of random effects are required to model the smooth random curves. Current computational methods for random effects mainly use Monte Carlo or Gauss–Hermite

(wileyonlinelibrary.com) DOI: 10.1002/sta4.121

quadrature approximations (Molenberghs & Verbeke, 2005), but in our context, these are computationally expensive. For example, Figueroa-Zúñiga et al. (2013) suggested that the beta regression could require more quadrature nodes than logistic regression.

We address this problem with the approach proposed by Cox (1996), where a quasilikelihood method (Wedderburn, 1974) is used to model continuous proportions. The quasilikelihood does not need to find a full distribution for outcomes but only requires the specification of the first and second moments. However, the existing methodology is limited to *independent* data, and we extend it to our longitudinal functional data scenario. In particular, the modelling of the correlation structure with respect to physical activity, month and weekday effects is necessary. To build a flexible model, we use functional random curves for the MVPA minutes. To avoid the dimension problem in random curves, we only use a few important functional principal components to summarize the random curves. This method is developed by Zhou et al. (2008) in the linear model, and we take this general approach to address a continuous proportional response.

A new efficient algorithm is proposed. The algorithm includes both features of penalized quasilikelihood (Breslow & Clayton, 1993) and the eigen-decomposition discussed in Yao et al. (2005). Because our problem involves quasi-likelihood modelling and random effects, a penalized quasilikelihood approach is convenient. On the other hand, the eigen-decomposition approach is efficient for simultaneous selection and estimation of the functional principal components. As a result, the new algorithm includes both procedures.

The paper is organized as follows. Section 2 describes the model, and Section 3 is for our algorithm in model fitting. Section 4 gives results from a simulation study. Section 5 analyzes the BodyMedia data set involving physical activity and suggests answers to the three questions posed earlier. Concluding remarks are given in Section 6.

2 Model

2.1 The mixed effects model for continuous proportional data

Let $Y_i(r,s,t)$ be a continuous proportional observation at MVPA minute r, month s and weekday t for subject $i = 1,\ldots,n$. Each subject has m_i observations. The possible values for s can be $1,2,\ldots,12$, which represents January to December, while t can be $1,2,\ldots,7$, indicating Sunday through Saturday. We use $Y_i(r_{ij},s_{ij},t_{ij})$ to denote the jth observation for subject i. Define $\mathbf{Y}_i = \{Y_i(r_{i1},s_{i1},t_{i1}),\ldots,Y_i(r_{im_i},s_{im_i},t_{im_i})\}^{\mathrm{T}}$.

According to the quasilikelihood method suggested by Wedderburn (1974) and McCullagh & Nelder (1989), we only specify the first and second moments for the outcomes. The mean and variance functions of $Y_i(r,s,t)$ given random effects are

$$\begin{aligned} E\{Y_i(r,s,t)|\mathbf{U}_i(r,s,t)\} &= H\{\mu(r,s,t)+\mathbf{U}_i(r,s,t)\}, \\ \mathrm{var}\{Y_i(r,s,t)|\mathbf{U}_i(r,s,t)\} &= \sigma^2[E\{Y_i(r,s,t)|\mathbf{U}_i(r,s,t)\}][1-E\{Y_i(r,s,t)|\mathbf{U}_i(r,s,t)\}], \end{aligned} \tag{1}$$

where $H(\cdot)$ denotes the logistic distribution function, $\mu(r,s,t)$ is a fixed curve, $\mathbf{U}_i(r,s,t)$ is a random effects curve and σ^2 is a dispersion parameter. We further assume that, given the random effects $\mathbf{U}_i(r,s,t)$, the variables in $\mathbf{Y}_i$ are independent.

Cox (1996) discussed other candidates for modelling $Y_i(r,s,t)$. For example, the variance structure can be specified as

$$\sigma^2[E\{Y_i(r,s,t)|\mathbf{U}_i(r,s,t)\}]^2[1-E\{Y_i(r,s,t)|\mathbf{U}_i(r,s,t)\}]^2.$$

However, unlike Cox's independent observation scenarios, our method involves the random effects curves $\mathbf{U}_i(r,s,t)$ to model the correlation structure. Therefore, we focus on model (1).

We further specify $\mu(r,s,t)$ and $\mathbf{U}_i(r,s,t)$ terms in (1) by additive models

$$\mu(r,s,t) = \mu_0(r) + \mu_1(s) + \mu_2(t); \tag{2}$$

$$\mathbf{U}_i(r,s,t) = \mathbf{U}_{i,0}(r) + \mathbf{U}_{i,1}(s) + \mathbf{U}_{i,2}(t), \tag{3}$$

where $\mu_0(r)$, $\mu_1(s)$ and $\mu_2(t)$ are fixed curves at r,s,t, and $\mathbf{U}_{i,0}(r)$, $\mathbf{U}_{i,1}(s)$, $\mathbf{U}_{i,2}(t)$ are random curves at r,s,t, respectively. For model identifiability, we set $\mu_1(1) = \mu_2(1) = 0$. We also assume that $\mathbf{U}_{i,0}(r)$, $\mathbf{U}_{i,1}(s)$ and $\mathbf{U}_{i,2}(t)$ are mutually independent for all r,s,t.

2.2 Basis functions

To model the fixed and random curves in (2) and (3), let $b_0(r) = \{b_{0,1}(r), \dots, b_{0,q_0}(r)\}^{\mathsf{T}}$, $b_1(s) = \{b_{1,1}(s), \dots, b_{1,q_1}(s)\}^{\mathsf{T}}$ and $b_2(t) = \{b_{2,1}(t), \dots, b_{2,q_2}(t)\}^{\mathsf{T}}$ be the vectors of orthogonal B-spline basis functions evaluated at physical activity minutes r, month s and weekday t, respectively. The orthogonal B-spline basis functions can be computed using an exact approach found in the R package "orthogonalsplinebasis" (Redd, 2011; R Core Team, 2016).

We model the fixed effects curves to be

$$\mu_0(r) = b_0^{\mathsf{T}}(r)\beta_0, \qquad \mu_1(s) = b_1^{\mathsf{T}}(s)\beta_1, \qquad \mu_2(t) = b_2^{\mathsf{T}}(t)\beta_2,$$

where $\beta_0, \beta_1, \beta_2$ are $q_0 \times 1$, $q_1 \times 1$, $q_2 \times 1$ regression coefficient vectors, and $s = 2, \dots, 12$ and $t = 2, \dots, 7$.

For the random effects curve $\mathbf{U}_{i,0}(r)$, we set

$$\mathbf{U}_{i,0}(r) = b_0^{\mathsf{T}}(r)u_{i,0},$$

where $u_{i,0}$ are $q_0 \times 1$ correlated random effects vectors. In practice, when q_0 is large, the estimation of the variance structure for $u_{i,0}$ could be difficult. Based on the principal component approach (Zhou et al., 2008), we summarize $u_{i,0}$ by using only a few principal components by setting $u_{i,0} \doteq \sum_{\ell=1}^{L} \theta_\ell \alpha_{i,\ell}$, where L is the number of principal components, θ_ℓ is the ℓth $q_0 \times 1$ orthogonal principal component vector and $\alpha_{i,\ell}$ is the ℓth principal component score. For identifiability, the principal components are sorted in decreasing order by the variance of $\alpha_{i,\ell}$ and the $\alpha_{i,\ell}$ is set to be independent across all $\ell = 1, \dots, L$. Denote $\Theta = (\theta_1, \dots, \theta_L)$ and $\alpha_i = (\alpha_{i,1}, \dots, \alpha_{i,L})^{\mathsf{T}}$. We assume $\alpha_i \sim \text{Normal}(0, \Delta)$, where $\Delta = \text{diag}(\Delta_1, \dots, \Delta_L)$, and thus $u_{i,0} \sim \text{Normal}(0, \Psi_0)$ with $\Psi_0 = \Theta\Delta\Theta^{\mathsf{T}}$. We further denote $f_\ell(r) = b_0^{\mathsf{T}}(r)\theta_\ell$ as the ℓth principal component curve.

Remark 1
Instead of using principal components, there are two commonly used models for the random effects curve $\mathbf{U}_{i,0}(r)$*, namely,*

$$\mathbf{U}_{i,0}(r) = u_{i,0}^*; \quad \textit{and} \quad \mathbf{U}_{i,0}(r) = u_{i,0}^* + u_{i,1}^* r,$$

where $u_{i,0}^*$ *and* $u_{i,1}^*$ *are scalar random effects. The first formulation only involves random intercepts, which implies homoscedasticity for* $\mathbf{U}_{i,0}(r)$ *across* r*. The second model has an additional random-slope term, so that the variance of* $\mathbf{U}_{i,0}(r)$ *is a quadratic function over* r*. However, as we show in the simulation study, both formulations can be limited when the random effects structure is complicated, and they can lead to biased estimates.*

For $\mathbf{U}_{i,1}(s)$ and $\mathbf{U}_{i,2}(t)$, we use dummy variables given as

$$\mathbf{U}_{i,1}(s) = \sum\nolimits_{k=1}^{12} I(s=k)u_{i,1,k} = I(s=1)u_{i,1,1} + I(s=2)u_{i,1,2} + \cdots + I(s=12)u_{i,1,12},$$
$$\mathbf{U}_{i,2}(t) = \sum\nolimits_{k=1}^{7} I(t=k)u_{i,2,k} = I(t=1)u_{i,2,1} + I(t=2)u_{i,2,2} + \cdots + I(t=7)u_{i,2,7},$$

(wileyonlinelibrary.com) DOI: 10.1002/sta4.121

where $I(\cdot)$ is an indicator function, and $u_{i,1} = (u_{i,1,1}, \ldots, u_{i,1,12})^{\mathrm{T}}$ and $u_{i,2} = (u_{i,2,1}, \ldots, u_{i,2,7})^{\mathrm{T}}$ are 12×1 and 7×1 random effects vectors. We assume $u_{i,1} \sim \mathrm{Normal}(0, \Psi_1)$ and $u_{i,2} \sim \mathrm{Normal}(0, \Psi_2)$ with $\Psi_1 = \mathrm{diag}(\Psi_{1,1}, \ldots, \Psi_{1,12})$ and $\Psi_2 = \mathrm{diag}(\Psi_{2,1}, \ldots, \Psi_{2,7})$.

Therefore, models (2) and (3) can be rewritten as

$$\mu(r,s,t) = b_0^{\mathrm{T}}(r)\beta_0 + b_1^{\mathrm{T}}(s)\beta_1 + b_2^{\mathrm{T}}(t)\beta_2; \tag{4}$$

$$\mathbf{U}_i(r,s,t) = b_0^{\mathrm{T}}(r)\Theta\alpha_i + \sum\nolimits_{k=1}^{12} I(s=k)u_{i,1,k} + \sum\nolimits_{k=1}^{7} I(t=k)u_{i,2,k}. \tag{5}$$

The modelling with B-splines involves six sets of parameters to be estimated: (i) the dispersion parameter: σ^2; (ii) the B-spline coefficients for the fixed effects: β_0, β_1 and β_2; (iii) the number of principal component: L; (iv) the B-spline coefficients for principal component functions: Θ; (v) the principal component scores' covariance matrix: Δ; and (vi) the covariance matrices for $u_{i,1}$ and $u_{i,2}$: Ψ_1 and Ψ_2.

3 Model fitting procedure

3.1 Second-order approximation for continuous proportions

Estimation of the parameters is complicated by the continuous proportional outcomes. We approximate the continuous proportions using a penalized quasilikelihood that includes a second-order approximation term. This method was introduced in Goldstein & Rasbash (1996), and it outperforms the methods proposed by Breslow & Clayton (1993). The method is as follows. Because $H(\cdot)$ is logistic distribution function, the first and second derivatives of $H(\cdot)$ are $H'(\cdot) = H(\cdot)\{1 - H(\cdot)\}$ and $H''(\cdot) = \{1 - 2H(\cdot)\}H'(\cdot)$. Let $g(\cdot) = 1/H'(\cdot)$.

Set $X(r,s,t) = \left\{b_0^{\mathrm{T}}(r), b_1^{\mathrm{T}}(s), b_2^{\mathrm{T}}(t)\right\}$. Let $I_1(s) = \{I(s=1), \ldots, I(s=12)\}^{\mathrm{T}}$, and $I_2(t) = \{I(t=1), \ldots, I(t=7)\}^{\mathrm{T}}$, and set $Z(r,s,t) = \left\{b_0^{\mathrm{T}}(r), I_1^{\mathrm{T}}(s), I_2^{\mathrm{T}}(t)\right\}$. Denote $\boldsymbol{\beta} = \left(\beta_0^{\mathrm{T}}, \beta_1^{\mathrm{T}}, \beta_2^{\mathrm{T}}\right)^{\mathrm{T}}$ and $\mathbf{u}_i = \left(u_{i,0}^{\mathrm{T}}, u_{i,1}^{\mathrm{T}}, u_{i,2}^{\mathrm{T}}\right)^{\mathrm{T}}$. Given known values of $(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{u}}_i)$, letting $\widehat{\eta}_i(r,s,t) = X(r,s,t)\widehat{\boldsymbol{\beta}} + Z(r,s,t)\widehat{\mathbf{u}}_i$, we use the approximate model

$$\begin{aligned} Y_i^*(r,s,t) &= g\{\widehat{\eta}_i(r,s,t)\}\left[Y_i(r,s,t) - H\{\widehat{\eta}_i(r,s,t)\}\right] + \widehat{\eta}_i(r,s,t) \\ &\quad - (1/2)g\{\widehat{\eta}_i(r,s,t)\}H''\{\widehat{\eta}_i(r,s,t)\}\left[Z(r,s,t)\widehat{\mathrm{var}}(\mathbf{u}_i - \widehat{\mathbf{u}}_i)Z^{\mathrm{T}}(r,s,t)\right] \\ &\approx X(r,s,t)\boldsymbol{\beta} + Z(r,s,t)\mathbf{u}_i + \epsilon_i(r,s,t), \end{aligned} \tag{6}$$

where $\epsilon_i(r,s,t) = \mathrm{Normal}\left[0, \sigma^2 g\{\widehat{\eta}_i(r,s,t)\}\right]$. The derivation of this approximation can be referred to Molenberghs & Verbeke (2005).

3.2 Estimation algorithm

According to the second-order approximation in (6), the transformed continuous proportional outcomes $Y_i^*(r,s,t)$ can be treated as continuous variables with normal distributions. We estimate the parameters using an Expectation/Conditional Maximization Either (ECME) algorithm (Schafer, 1998). The ECME algorithm updates fixed structure parameters by the Newton–Raphson approach and updates the random effects parameters by the Expectation Maximization (EM) method. We provide a brief sketch of the model estimation procedure here, and the details are in the Supporting Information for Appendix S1.

We set the initial numbers of principal components to be $L = q_0$, and thus, Ψ_0 is a full rank covariance matrix. We also give initial values for other parameters listed in Section 2.2. Then the iteration procedure is as follows:

1. Update $\boldsymbol{\beta}$ and σ^2 by a Newton–Raphson approach,

2. Update Ψ_0, Ψ_1 and Ψ_2 by the EM method, and
3. Update L, Θ and Δ with an eigen-decomposition of Ψ_0.

The entire procedure is iterated until convergence. For convergence properties of the ECME algorithm, see Liu & Rubin (1994).

3.3 Maximum penalized likelihood

The previous discussion focuses on the modelling of the response variables using basis functions. It is helpful, however, to introduce roughness penalties to regularize the fits of functions (Eilers & Marx, 1996). Denote θ_ℓ to be the ℓth column for Θ.

We penalize the loglikelihood and update the parameters in each iteration to maximize

$$\mathcal{L}_{pen} = \mathcal{L} - \tau_\beta \beta^{\mathsf{T}} D \beta - \tau_\theta \sum\nolimits_{\ell=1}^{L} \theta_\ell^{\mathsf{T}} D_0 \theta_\ell,$$

where $\mathcal{L}$ is defined in the Appendix S1 Equation (S.2), τ_β and τ_θ are penalty parameters, and the penalty matrices are $D_0 = \int b_0''(r) b_0''(r)^{\mathsf{T}}\,\mathrm{d}r$, $D_1 = \int b_1''(s) b_1''(s)^{\mathsf{T}}\,\mathrm{d}s$, $D_2 = \int b_2''(t) b_2''(t)^{\mathsf{T}}\,\mathrm{d}t$ and $D = \mathrm{diag}(D_0, D_1, D_2)$.

Using maximum penalized likelihood has only a minor effect on the estimation algorithm, although of course it has a major effect on the estimation results. We describe the details in the Supporting Information for Appendix S1.3.

In all of our work, we use fivefold crossvalidation to choose penalty parameters. We searched over a two-dimensional grid for τ_β and τ_θ. The tuning parameters are obtained by maximizing the cross-validated loglikelihood

$$-\sum\nolimits_{i=1}^{n} [\, m_i \log(2\pi) + \log\{|\widehat{\mathrm{cov}}(\mathbf{Y}_i)|\} + \{\mathbf{Y}_i - \widehat{E}(\mathbf{Y}_i)\}^{\mathsf{T}} \{\widehat{\mathrm{cov}}(\mathbf{Y}_i)\}^{-1} \{\mathbf{Y}_i - \widehat{E}(\mathbf{Y}_i)\}] \,,$$

where the estimates $\widehat{E}(\mathbf{Y}_i)$ and $\widehat{\mathrm{cov}}(\mathbf{Y}_i)$ are described in the Supporting Information for Appendix S1.4.

4 Simulation studies

We use a simulation of 500 runs to assess the performance of our longitudinal functional additive model. There are $n = 240$ subjects, and each subject has 84 visits observed in 12 weeks; this is similar to our BodyMedia data, but with a much smaller number of subjects. Each week has complete observations from Monday to Sunday. We set each month to have 4 weeks. All subjects are observed in 3 consecutive months. For example, subject 1 is observed from January to March and subject 2 is observed from February to April. Then $Y_i(r,s,t)$ is generated according to the beta distribution with density function conditional on $\mathbf{U}_i(r,s,t)$ as

$$\frac{\Gamma(\phi)}{\Gamma(\kappa\phi)\Gamma\{(1-\kappa)\phi\}} \{y_i(r,s,t)\}^{\kappa\phi-1}\{1-y_i(r,s,t)\}^{(1-\kappa)\phi-1},$$

where $\Gamma(\cdot)$ is the gamma function, $\kappa = E\{Y_i(r,s,t)|\mathbf{U}_i(r,s,t)\}$ and $\phi = 1/\sigma^2 - 1$. We set $E\{Y_i(r,s,t)|\mathbf{U}_i(r,s,t)\} = H\{\mu_0(r) + \mu_1(s) + \mu_2(t) + f_1(r)\alpha_{i,1} + f_2(r)\alpha_{i,2} + \mathbf{U}_{i,1}(s) + \mathbf{U}_{i,2}(t)\}$, where $\mu_0(r) = H(r/2 - 5.5)$, $\mu_1(s) = (s - 7)^2/36 - 1$ and $\mu_2(t) = -(t-3)^2/5 + 0.8$. The principal component curves are $f_1(r) = \sin\{2\pi r/22\}/\sqrt{11}$ and $f_2(r) = \cos\{2\pi r/22\}/\sqrt{11}$. We generate $\alpha_{i,1}$, $\alpha_{i,2}$, $\mathbf{U}_{i,1}(s)$ and $\mathbf{U}_{i,2}(t)$ as normally distributed with zero means and set $\Delta_1 = 12$, $\Delta_2 = 6$, $\Psi_{1,s} = 2$ for all s, and $\Psi_{2,t} = 1$ for all t. We also generate r as uniformly distributed in $[0, 22]$. For σ^2, we studied $\sigma^2 = 0.02$ by following the suggestion from Figueroa-Zúñiga et al. (2013) and $\sigma^2 = 1/30$, which is similar to the result of our data application in Section 5. Our method has good performances in both scenarios, and we only report the simulation results from $\sigma^2 = 1/30$ here.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.121

As a comparison with our method, three naive approaches are explored. The first approach (labelled as NAIVE1) follows our algorithm but uses a random intercepts model for $\mathbf{U}_{i,0}(r)$ as discussed in Remark 1. The second method (labelled as NAIVE2) is similar to NAIVE1 but uses a random slopes model. The third method (labelled as NAIVE3) uses an identical random effects structure as our method, but it first takes a logit transformation of the responses and then fits the outcomes by a linear functional data model (Zhou et al., 2008).

We use cubic B-spline basis function with 10 equispaced knots to fit $\mu_0(r)$ and use linear B-spline basis functions with five and four knots to fit $\mu_1(s)$ and $\mu_2(t)$, respectively. Convergence was achieved for all simulated data sets. The correct number of principal components was selected in all simulated data sets. Table S1 presents the mean estimates and the mean squared errors (MSE) of the parameters, which indicates good performance in the estimation of the model parameters for our approach.

Figure S1(a) and (b) show the true fixed curve $\mu_0(r)$ and the averaged estimates of the four methods. They indicate that NAIVE1, NAIVE2 and NAIVE3 approaches lead to obviously biased outcomes, while there is little bias for our method. Figure 1(c)–(f) represent the performance of our approach in fixed curves $\mu_1(s)$, $\mu_2(t)$ and the principal component curves $f_1(r)$, $f_2(r)$, respectively. Our approach captures all of the true curve patterns.

5 Application to physical activity data

In this section, we apply our methods to the BodyMedia data to help answer the question raised in Section 1. Our data involve 935 males, and each person has 84 observations consisting of daily METS and daily sleep efficiency. Referring back to the notation of Section 2, $Y_i(r,s,t)$ is the ratio of daily sleep time to lying down time, r is the average minutes of MVPA time on the current day and previous day, s is the month and t is the weekday. We use cubic B-spline basis function with 24 equispaced knots to fit $\mu_0(r)$, while other basis functions follow the settings in Section 4. The dispersion parameter σ^2 is estimated to be 0.034.

Figure S2 presents the estimation results for fixed effects curves $\mu_0(r)$, $\mu_1(s)$ and $\mu_2(t)$, and principal component curve $f_1(r)$. Figure S2(a) suggests that conditioning on the random effect $\mathbf{U}_i(r,s,t)$, the relation between MVPA minutes and sleep efficiency can be divided into three parts. From minutes 0 to 60, the increase of MVPA minutes leads to higher sleep efficiency. The effect of increasing MVPA minutes flattens out gradually between minutes 60 and 120, while the greater MVPA minutes have negative influence on sleep efficiency after minutes 120. Figure S2(b) suggests that sleep efficiency has a strong monthly trend, where January and February have higher sleep efficiency, while October has lower sleep efficiency. Figure S2(c) indicates that Sunday and Monday have lower sleep efficiency but the sleep efficiency on Wednesday and Thursday is higher. Thus, it suggests that sleep efficiency is greater in the middle of the week.

The number of principal components for MVPA minutes is selected to be 1. Figure S2(d) shows the principal component curve is decreasing with increasing MVPA time, which means that greater physical activity leads to less between-subject variability. The result implies subjects with more physical activity have more consistent sleep efficiency.

We also study the marginal mean and variance structure of $Y_i(r,s,t)$. The month is set to be January, and the weekday is Monday. Figure S3(a) presents the marginal mean of the outcomes evaluated at different MVPA minutes. It displays the increasing MVPA results in the improvement of sleep efficiency. However, the increase in sleep efficiency is up to about 120 MVPA minutes, and then it tails off. Figure S3(b) is the marginal variance of the responses. The variability of sleep efficiency is decreasing with increasing MVPA time. In particular, the variability at 200 MVPA minutes is about half of that at 0 MVPA minutes. This suggests that people with higher physical activity time will generally have more constant sleep efficiency.

We show the correlation structure of the outcomes with respect to MVPA time on Monday in January, corr$\{Y_i(j,1,2), Y_i(k,1,2)\}$, as a 3-D plot in Figure 4. The figure reaches its peak around $(j=0, k=0)$. This makes sense because lower MVPA time leads to higher variability in sleep efficiency, which causes greater correlation. On the other hand, the plot decreases as MVPA minutes increase. This is likely because, intuitively, sleep efficiency is relatively constant for people with longer MVPA minutes.

6 Discussion

We have proposed a three-factor joint modelling and estimation strategy for functional data with continuous proportions. The simulation results are encouraging, with little bias. The analysis of the BodyMedia data using the our method demonstrates its utility in real applications. Our conclusions are that daily sleep efficiency improves with increasing MVPA up to about 120 minutes and increasing MVPA results in a decrease in the variance of sleep efficiency throughout the range of MVPA minutes. The former conclusion makes sense in general; however, the plateau of mean daily sleep efficiency at about 120 MVPA minutes has not been reported previously, largely because fully linear modelling of this data is standard. We believe that the substantial decrease in the variability of sleep efficiency as MVPA minutes increase is also a new finding, with standard analyses focusing only on means.

Acknowledgements

Li was supported by discovery grants programme from the Natural Sciences and Engineering Research Council of Canada (NSERC, RGPIN-2015-04409). Carroll was supported by a grant from the National Cancer Institute (U01-CA057030). The authors thank BodyMedia, Inc. for making the data available to them.

References

Breslow, NE & Clayton, DG (1993), 'Approximate inference in generalized linear mixed models', *Journal of the American Statistical Association*, **88**, 9–25.

Cox, C (1996), 'Nonlinear quasi-likelihood models: applications to continuous proportions', *Computational Statistics and Data Analysis*, **21**, 449–461.

Eilers, PHC & Marx, BD (1996), 'Flexible smoothing with B-splines and penalties', *Statistical Science*, **11**, 89–121.

Ekstedt, M, Nyberg, G, Ingre, M, Örjan, E & Marcus, C (2013), 'Sleep, physical activity and BMI in six to ten-year-old children measured by accelerometry: a cross-sectional study', *International Journal of Behavioral Nutrition and Physical Activity*, **10**, 82–91.

Ferrari, S & Cribari-Neto, F (2004), 'Beta regression for modelling rates and proportions', *Journal of Applied Statistics*, **31**, 799–815.

Figueroa-Zúñiga, JI, Arellano-Valle, RB & Ferrari, SL (2013), 'Mixed beta regression: a Bayesian perspective', *Computational Statistics and Data Analysis*, **61**, 137–147.

Gertheiss, J, Maier, V, Hessel, EF & Staicu, AM (2015), 'Marginal functional regression models for analyzing the feeding behavior of pigs', *Journal of Agricultural, Biological, and Environmental Statistics*, **20**, 353–370.

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

Goldstein, H & Rasbash, J (1996), 'Improved approximations for multilevel models with binary responses', *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **159**, 505–513.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.121

The ISI's Journal for the Rapid
Dissemination of Statistics Research

Hall, P, Müller, HG & Yao, F (2008), 'Modelling sparse generalized longitudinal observations with latent gaussian processes', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**, 703–723.

Kieschnick, R & McCullough, BD (2003), 'Regression analysis of variates observed on (0, 1): percentages, proportions and fractions', *Statistical Modelling*, **3**, 193–213.

Lambiase, MJ, Gabriel, KP, Kuller, LH & Matthews, KA (2013), 'Temporal relationships between physical activity and sleep in older women', *Medicine and Science in Sports and Exercise*, **45**, 2362–2368.

Liu, C & Rubin, DB (1994), 'The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence', *Biometrika*, **81**, 633–648.

McCullagh, P & Nelder, JA (1989), *Generalized Linear Models*, *Chapman and Hall:* London, U.K.

Molenberghs, G & Verbeke, G (2005), *Models for Discrete Longitudinal Data*, *Springer*, New York, U.S.A.

Oudegeest-Sander, MH, Eijsvogels, TH, Verheggen, RJ, Poelkens, F, Hopman, MT, Jones, H & Thijssen, DH (2013), 'Impact of physical fitness and daily energy expenditure on sleep efficiency in young and older humans', *Gerontology*, **59**, 8–16.

R Core Team (2016), *R: a language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria.

Redd, A (2011), *orthogonalsplinebasis: orthogonal bspline basis functions*. R package version 0.1.5.

Schafer, JL (1998), *Some improved procedures for linear mixed models*, Technical Report, The Methodological Center, The Pennsylvania State University.

Scheipl, F, Gertheiss, J & Greven, S (2016), 'Generalized functional additive mixed models', *Electronic Journal of Statistics*, **10**, 1455–1492.

Serban, N, Staicu, AM & Carroll, RJ (2013), 'Multilevel cross-dependent binary longitudinal data', *Biometrics*, **69**, 903–913.

Simas, AB, Barreto-Souza, W & Rocha, AV (2010), 'Improved estimators for a general class of beta regression models', *Computational Statistics and Data Analysis*, **54**, 348–366.

Verkuilen, J & Smithson, M (2012), 'Mixed and mixture regression models for continuous bounded responses using the beta distribution', *Journal of Educational and Behavioral Statistics*, **37**, 82–113.

Wedderburn, RWM (1974), 'Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method', *Biometrika*, **61**, 439–447.

Yao, F, Müller, HG & Wang, JL (2005), 'Functional data analysis for sparse longitudinal data', *Journal of the American Statistical Association*, **100**, 577–590.

Zhao, W, Zhang, R, Huang, Z & Feng, J (2012), 'Partially linear single-index beta regression model and score test', *Journal of Multivariate Analysis*, **103**, 116–123.

Zhou, L, Huang, JZ & Carroll, RJ (2008), 'Joint modelling of paired sparse functional data using principal components', *Biometrika*, **95**, 601–619.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web-site.