Published in final edited form as:

J Am Stat Assoc. 2018 ; 113(523): 992–1002. doi:10.1080/01621459.2017.1379402.

Latent Variable Poisson Models for Assessing the Regularity of Circadian Patterns over Time

Sungduk Kim and Paul S. Albert

Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD

Abstract

Many researchers in biology and medicine have focused on trying to understand biological rhythms and their potential impact on disease. A common biological rhythm is circadian, where the cycle repeats itself every 24 hours. However, a disturbance of the circadian pattern may be indicative of future disease. In this article, we develop new statistical methodology for assessing the degree of disturbance or irregularity in a circadian pattern for count sequences that are observed over time in a population of individuals. We develop a latent variable Poisson modeling approach with both circadian and stochastic short-term trend (autoregressive latent process) components that allow for individual variation in the degree of each component. A parameterization is proposed for modeling covariate dependence on the proportion of these two model components across individuals. In addition, we incorporate covariate dependence in the overall mean, the magnitude of the trend, and the phase-shift of the circadian pattern. Innovative Markov chain Monte Carlo sampling is used to carry out Bayesian posterior computation. Several variations of the proposed models are considered and compared using the deviance information criterion. We illustrate this methodology with longitudinal physical activity count data measured in a longitudinal cohort of adolescents.

Keywords

Autoregressive process; Circadian pattern; Longitudinal count data; Physical activity; Random effect; Serial correlation; Shrinkage prior; Stick-breaking prior

1. Introduction

Many researchers in biology and medicine have focused on understanding biological rhythms and their potential impact on disease. A common biological rhythm is circadian, where the cycle repeats itself every 24 hours. There has been substantial statistical methodological development that characterizes circadian patterns in biomarkers. Greenhouse, Kass, and Tsay (1987), Brown and Czeisler (1992), Wang and Brown (1996), Wang, Ke, and Brown (2003), and Albert and Hunsberger (2005) proposed statistical

CONTACT Sungduk Kim kims2@mail.nih.gov Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD 20892.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

methodology for modeling circadian mean structure in continuous longitudinal data. Ogbagaber et al. (2012) extended the shape-invariant model proposed by Wang, Ke, and Brown (2003) to longitudinal physical activity count data. Hu et al. (2007) and Indic et al. (2012) propose the use of nonlinear dynamic approaches for studying circadian patterns in physical activity data over time. All this work, however, focuses only on characterizing the circadian patterns across a population of individuals. Specifically, the focus is on characterizing the underlying circadian shape pattern, amplitude, and phase-shift. However, abnormal biology may reflect itself in disturbed circadian patterns, and scientific interest may be on examining what factors relate to having regular or disturbed (irregular) circadian patterns. In addition to modeling the mean structure in circadian rhythms, this article focuses on understanding what factors influence the degree of circadian as compared with stochastic short-term trend across individuals. In our proposed approach, we characterize both circadian and stochastic variation, and allow for each individual to have an additive and varying component of each.

This methodology was motivated by a recent study that examined the circadian pattern in longitudinal physical activity counts among adolescent girls (Ogbagaber et al. 2012). In this study, 96 adolescent girls completing the 9th grade were asked to wear a physical activity watch (actiwatch2) for a period of 7 days (Kushida et al. 2001). When the actiwatch2 detected nonwearing of the watch at any time during the 7 days, then the participants were asked to redo the whole 7-day followup. Physical activity was characterized by counting physical motions over 15-minute intervals. Ogbagaber et al. (2012) developed a shape-invariant modeling approach for introducing covariates dependence on the mean structure. With this approach, they showed that the circadian pattern was different between obese and nonobese girls. This approach characterized the mean circadian pattern, but did not examine how covariates may affect the relative proportion of circadian versus stochastic short-term trend across participants.

A close examination of the physical activity pattern across individuals shows markedly different types of physical activity patterns in these subjects (Figure 1). Figure 1(a) shows an individual adolescent with a pronounced circadian pattern across the seven days of follow-up. Figure 1(b) shows an adolescent with a slightly disturbed circadian pattern. Furthermore, Figure 1(c) shows an individual with a severely disturbed circadian pattern. This analysis suggested that some individuals have a regular circadian pattern, while others show a much more irregular pattern in physical activity. Interestingly, the degree of a circadian versus stochastic short-term trend may be reflective of future outcomes or important demographic variables. For example, irregular physical activity may result in higher rates of obesity, or perhaps, may be higher in teenagers who are already obese. Developing a modeling strategy that allows us to examine this important public health question is of interest. Figure 2 shows the overall mean and autocorrelation function (ACF) for these longitudinal count data. The figure shows a clear overall circadian pattern across the seven days of follow-up. However, the ACF shows a dampening of the daily correlation across time, suggesting that there may be additional sources of variation in addition to the circadian pattern.

Generalized linear mixed models (GLMMs) have been used to estimate the mean structure and variation for longitudinal discrete data (Karim and Zeger 1992; McCulloch et al. 2008).

Various extensions of GLMMs have been proposed to incorporate serial correlation using a latent process (Harvey 1989; Smith 1979; Zeger 1988; Albert et al. 2002). Chen and Ibrahim (2000) proposed a Bayesian approach for a time series of counts by constructing a likelihood-based generalization of a model considered by Zeger (1988). Hay and Pettitt (2001) considered the Bayesian analysis of a time series of counts with a random effect having a time series correlation structure. Recently, Kim et al. (2013) proposed a hierarchical Poisson regression model to incorporate overdispersion, heterogeneity, and serial correlation as well as a semiparametric mean structure. These approaches focused on either making inference about the overall mean structure or characterizing the between- to within- subject variation in longitudinal count data. Our focus is on characterizing (through covariates and random effects) the degree of circadian versus stochastic variation in longitudinal physical activity count data. There are different approaches to the statistical analysis of periodic data. The classical time series approach is based on Fourier expansions (Huettel, Song, and McCarthy 2004). A nonparametric method is proposed by Ramsay and Silverman (2005) in which the original series is split into individual cycles and each cycle is treated as an independent curve. A more efficient method is semiparametric shape-invariant modeling (Lawton and Sylvestre 1971; Lawton, Sylvestre, and Maggio 1972; Lindstrom 1995; Wang and Brown 1996; Hurtgen and Gervini 2009). Zeger and Diggle (1994) and Zhang et al. (1998) proposed a semiparametric stochastic mixed model, which assumes parametric covariate effects and nonparametric time effect, and accounts for the within-subject correlation using random effects and a stochastic process. Wang, Ke, and Brown (2003) proposed the shape-invariant mixed effects models for circadian rhythms with smoothing Spline ANOVA decompositions to study potential covariate effects on the common shape function. Albert and Hunsberger (2005) proposed a regression-based approach for analyzing longitudinal continuous circadian rhythm data where individual variations were incorporated through random effects added to the mean, amplitude, and phase-shift. For counts data with periodicity, Ogbagaber et al. (2012) recently proposed to use a shape-invariant Poisson models for circadian rhythm count data, which is an adaptation of Albert and Hunsberger (2005) to the longitudinal physical activity example. All these approaches focus on characterizing the circadian pattern in the mean structure, while in this article, the focus is on characterizing individual differences in the degree of circadian variation across individuals.

In Section 2, we formulate a Poisson model that allows for subject-specific variation in the relative contribution of a circadian pattern and a stochastic latent process, the latter of which governs short-term trend that characterizes irregular circadian variation. The circadian pattern is modeled with a harmonic representation that allows for individual variation in the phase-shift across subjects. Section 3 discusses Bayesian estimation including the proposed prior distributions and Bayesian computations. We demonstrate the performance of the proposed methodology through a simulation study in Section 4. Section 5 presents an analysis of the longitudinal physical activity data that motivated the modeling strategy. A discussion follows in Section 6.

2. The Latent Variable Poisson models

2.1. Model Framework

Suppose that i denotes individual ($i=1, \dots, I$) and j denotes the follow-up interval ($j=1, \dots, J$). We assume that there are I individuals in the study, each contributing J time intervals. Let m denote the day of continuous activity monitoring ($m=1, \dots, M$). Further, k denotes the time interval on day m ($k=1, \dots, K$) where k , beginning at 12:00 am, repeats itself on each new day. Let Y_{ij} be a sequence of counts for the i th individual at the j th time interval. Also, let $\mathbf{z}_i = (z_{i1}, z_{i2}, \dots, z_{i,q_1})'$, $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{i,q_2})'$, $\mathbf{w}_i = (w_{i1}, w_{i2}, \dots, w_{i,q_3})'$, and

$\mathbf{u}_i = (u_{i1}, u_{i2}, \dots, u_{i,q_4})'$; denote q_1, q_2, q_3 , and q_4 -dimensional vectors of covariates for

individual i . Note that $\mathbf{x}_i$, $\mathbf{w}_i$, and $\mathbf{u}_i$ can share common covariates with $\mathbf{z}_i$. Also,

$\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_{q_1})'$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_{q_2})'$, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_{q_3})'$, and $\boldsymbol{\phi} = (\phi_1, \dots, \phi_{q_4})'$ are the

corresponding vectors of regression coefficients, respectively. We define $h(k)$ and b_{ij} as the circadian pattern and stochastic short-term variation components. $h(k)$ follows a harmonic representation, while b_{ij} is an autoregressive (AR) process. We allow for participants to have differing degrees of these two components by introducing a mixing parameter that depends

on both covariates and random effects. Denote $p_i = \frac{\exp(\mathbf{x}_i' \boldsymbol{\beta} + \sigma_{\xi} \xi_i)}{1 + \exp(\mathbf{x}_i' \boldsymbol{\beta} + \sigma_{\xi} \xi_i)}$ as a mixing parameter,

where ξ_i is random effect with mean 0 and variance 1, and σ_{ξ} is a standard deviation (SD) of ξ_i . Further, denote $\sigma_i = \exp(\mathbf{w}_i' \boldsymbol{\theta})$ as a scale parameter. The observed count process, y_{ij} is assumed to follow an independent Poisson distribution with mean λ_{ij}

$$\log \lambda_{ij} = \mathbf{z}_i' \boldsymbol{\alpha} + \sigma_{\tau} \tau_i + \sigma_{\gamma} \gamma_{ij} + \sigma_i (p_i h(k) + (1 - p_i) b_{ij}), \quad (1)$$

where τ_i is a individual-level random effect that induces exchangeable correlation between individuals, and γ_{ij} is a time-level random effect that accounts for overdispersion. The term $\sigma_i (p_i h(k) + (1 - p_i) b_{ij})$ reflects the circadian/stochastic pattern where σ_i governs the amplitude of the combined process. The random effects are assumed to be independent of each other with $\tau_i \sim N(0, 1)$, $\gamma_{ij} \sim N(0, 1)$, and $\xi_i \sim N(0, 1)$, where $N(a, b)$ denotes a normal distribution with mean a and variance b . Furthermore, σ_{τ} and σ_{γ} are the SDs of τ_i and γ_{ij} , respectively.

We model the stochastic component, b_{ij} as an AR(1) process,

$$\begin{aligned} b_{i1} &= \eta_{i1}, \eta_{i1} \sim N(0, 1), \\ b_{ij} &= \rho b_{i,j-1} + \eta_{ij}, \quad j \geq 2, \end{aligned} \quad (2)$$

where $\eta_{ij} \sim N(0, 1 - \rho^2)$ for $j \geq 2$. Here ρ , where $0 < \rho < 1$, is an autocorrelation parameter that determines how rapidly the serial correlation decreases between $b_{i,j-1}$ and b_{ij} . We note

that $\text{Var}(b_{ij}) = 1$ for $j = 1, \dots, J$ and $I = 1, \dots, I$. We model the circadian component, $h_i^*(k)$, as a harmonic function with a Fourier wavelet basis that can be parameterized as

$$h_i^*(k) = \sum_{l=1}^L \varphi_l^* \cos(2\pi l(k/K + \zeta_{il})), \quad (3)$$

where k/K is in fractions of a day ranging from 0 to 1 starting at midnight, L is the number of harmonic terms where a large L results in a flexible underlying circadian rhythm, and $0 < \zeta_{il} < 1$. Furthermore, L is the number of harmonic terms with $L = 1$ reducing to a simple sinusoidal circadian pattern and a large value of L allowing for a very flexible pattern. The parameters φ_l^* assess the contribution of each of the L terms to the harmonic function, while the ζ_{il} is a parameter that characterizes each participant phase-shift as a percentage of a 24-hour period. Variation in the phase-shift for the l th harmonic function and i th participant, ζ_{il} ,

is characterized by $\zeta_{il} = \frac{\exp(\delta_l + u_l' \phi + \sigma_\psi \psi_i)}{1 + \exp(\delta_l + u_l' \phi + \sigma_\psi \psi_i)}$ where δ_l is an intercept for the phase-shift of

the l th harmonic function, ψ_i is an individual-level random effect with mean 0 and variance 1, and σ_ψ is a SD of ψ_i . Furthermore, the harmonic function in (3) can be rewritten as

$$h_i^*(k) = \sum_{l=1}^L [a_{il}^* \cos(2\pi l k/K) + b_{il}^* \sin(2\pi l k/K)], \quad (4)$$

where $a_{il}^* = \varphi_l^* \cos(2\pi l \zeta_{il})$, $b_{il}^* = \varphi_l^* \sin(2\pi l \zeta_{il})$, and $-\infty < a_{il}^*$ and $b_{il}^* < \infty$. In addition, $a_{il}^2 + b_{il}^2 = \varphi_l^{*2}$ for $l = 1, \dots, L$. Note that the harmonic function in (4) is the alternative form of (3) which has been used in the literature. The random effect ψ_i is assumed to be independent of τ_i , γ_{ij} , b_{ij} , and ξ_i with $\psi_i \sim N(0, 1)$.

In order to fairly compare the relative contribution of the circadian versus the stochastic component (i.e., for p_i to be a meaningful quantity), we need to standardize the harmonic function $h_i^*(k)$ so that it has comparable variation across time as the unit variance stochastic process b_{ij} . Thus, we define $h_i(k)$ as the standardized version,

$$h_i(k) = \frac{h_i^*(k)}{\int_0^1 h_i^*(t)^2 dt} = \sum_{l=1}^L \varphi_l^2 \sqrt{2} \cos(2\pi l(k/K + \zeta_{il})), \quad (5)$$

where $t = k/K$, $\varphi_l = \frac{\varphi_l^{*2}}{\sum_{l'=1}^L \varphi_{l'}^{*2}}$, $0 < \varphi_l < 1$, and $\sum_{l=1}^L \varphi_l = 1$. Further, from (4), the harmonic function can be also rewritten as

$$h_i(k) = \sum_{l=1}^L [a_{il} \cos(2\pi lk/K) + b_{il} \sin(2\pi lk/K)], \quad (6)$$

where $a_{il} = \varphi_l^{\frac{1}{2}} \sqrt{2} \cos(2\pi l \zeta_{il})$ and $b_{il} = -\varphi_l^{\frac{1}{2}} \sqrt{2} \sin(2\pi l \zeta_{il})$. Note that $\sum_{l=1}^L (a_{il}^2 + b_{il}^2) = 2$ and $-\sqrt{2} < a_{il}$ and $b_{il} < \sqrt{2}$. Our formulation extends the shape-invariant circadian model (Wang, Ke, and Brown 2003; Albert and Hunsberger 2005) to incorporate the stochastic short-term variation into the longitudinal process.

2.2. Hierarchical Centering Reparametrization

Due to the complexity of the model in (1), we propose to use the hierarchical centering reparameterization technique of Ibrahim, Chen, and Ryan (2000), which is particularly suitable for the implementation of the Markov chain Monte Carlo (MCMC) sampling for our problem. Ibrahim, Chen, and Ryan (2000) considered the hierarchical centering reparameterization technique of Gelfand, Sahu, and Carlin (1996) for a Poisson regression model with a latent AR(1) process for a time series of counts. They observed that the mixing and convergence of Gibbs sampler were improved. In addition, Kim et al. (2013) also proposed a reparameterization scheme that proves to enhance the convergence of Gibbs sampler for the hierarchical Poisson regression model incorporating overdispersion, heterogeneity, and serial correlation as well as a semiparametric mean structure. Based on the proposed model described by (1), (2), (5) and random effects such as τ_i , γ_{ij} , b_{ij} , ξ_i and ψ_i , we first tried a Gibbs sampler without the hierarchical centering reparameterization technique, and all the parameters converged very slowly. To improve convergence, we considered the reparameterization:

$$b_{ij}^* = z_i' \alpha + \sigma_\tau \tau_i + \sigma_\gamma \gamma_{ij} + \sigma_i p_i h_i(k) + \sigma_i (1 - p_i) b_{ij}, \quad (7)$$

where $E(b_{ij}^*) = z_i' \alpha + \sigma_\tau \tau_i + \sigma_\gamma \gamma_{ij} + \sigma_i p_i h_i(k)$ and $Var(b_{ij}^*) = \sigma_i^2 (1 - p_i)^2$. Using the hierarchical centering in (7), the proposed model in (1) can be rewritten as

$$\begin{aligned} y_{ij} &\sim \text{Poisson}(\exp(b_{ij}^*)) \quad \text{and} \\ b_{i1}^* &= \mu_{i1} + \epsilon_{i1}, \epsilon_{i1} \sim N(0, \sigma_i^2 (1 - p_i)^2), \\ b_{ij}^* - \mu_{ij} &= \rho(b_{i,j-1}^* - \mu_{i,j-1}) + \epsilon_{ij}, \quad j \geq 2, \end{aligned} \quad (8)$$

where $\mu_{ij} = z_i' \alpha + \sigma_\tau \tau_i + \sigma_\gamma \gamma_{ij} + \sigma_i p_i h_i(k)$, $\epsilon_{ij} \sim N(0, (1 - \rho^2) \sigma_i^2 (1 - p_i)^2)$ for $j = 2, j = (m-1)K + k$, $k = 1, \dots, K$, $m = 1, \dots, M$. The convergence for the data analysis and (Section 4) was much improved using this reparameterization. Let $\mathbf{y} = (y_{11}, y_{12}, \dots, y_{LMK})'$, $\mathbf{Z} = (z_1, z_2, \dots, z_M)'$, $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M)'$, $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_M)'$, and $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_M)'$. Also let $\boldsymbol{\delta} = (\delta_1, \delta_2, \dots, \delta_L)'$, $\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \dots, \varphi_L)'$, $\boldsymbol{\tau} = (\tau_1, \tau_2, \dots, \tau_J)'$, $\boldsymbol{\xi} = (\xi_1, \xi_2, \dots, \xi_J)'$, $\boldsymbol{\psi} = (\psi_1, \psi_2, \dots,$

$\psi_D, \mathbf{b}^* = (b_{11}^*, b_{12}^*, \dots, b_{I,MK}^*)'$, and $\boldsymbol{\gamma} = (\gamma_{11}, \gamma_{12}, \dots, \gamma_{I,MK})'$. Then, we let $D_{\text{obs}}(\mathbf{y}, \mathbf{Z}, \mathbf{X}, \mathbf{W}, \mathbf{U})$ and $D = (\mathbf{y}, \mathbf{Z}, \mathbf{X}, \mathbf{W}, \boldsymbol{\tau}, \boldsymbol{\xi}, \boldsymbol{\psi}, \mathbf{b}^*, \boldsymbol{\gamma})$ denote the observed and complete data, respectively. The complete data likelihood function can be written as

$$\begin{aligned} & L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\delta}, \boldsymbol{\varphi}, \rho, \sigma_\tau, \sigma_\xi, \sigma_\psi, \sigma_\gamma | D) \\ &= \prod_{i=1}^I \prod_{m=1}^M \prod_{k=1}^K \exp[y_{ij} b_{ij}^* - \exp(b_{ij}^*) - \log(y_{ij}!)] \\ &\times \prod_{i=1}^I \prod_{m=1}^M \prod_{k=1}^K N(b_{ij}^* | \mu_{ij} + \rho(b_{i,j-1}^* - \mu_{i,j-1}), (1 - \rho^2)\sigma_i^2(1 - p_i)^2) \\ &\times \prod_{i=1}^I \prod_{m=1}^M \prod_{k=1}^K N(r_{ij} | 0, 1) \\ &\times \prod_{i=1}^I N(\tau_i | 0, 1) \times N(\xi_i | 0, 1) \times N(\psi_i | 0, 1), \end{aligned} \quad (9)$$

where μ_{ij} is given in (8) and $j = (m - 1)K + k$. Further, we obtain the observed likelihood function after integrating out all random effects such as $\boldsymbol{\tau}, \boldsymbol{\xi}, \boldsymbol{\psi}, \boldsymbol{\gamma}$, and $\mathbf{b}^*$ from the complete-data likelihood function in (9).

3. Posterior Inference

3.1. Prior and Posterior Distributions

We consider a joint prior distribution for $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\delta}, \boldsymbol{\varphi}, \rho, \sigma_\tau, \sigma_\xi, \sigma_\psi, \sigma_\gamma)$. We assume that the joint prior for $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\delta}, \boldsymbol{\varphi}, \rho, \sigma_\tau, \sigma_\xi, \sigma_\psi, \sigma_\gamma)$ is of the form $\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\delta}, \boldsymbol{\varphi}, \rho, \sigma_\tau, \sigma_\xi, \sigma_\psi, \sigma_\gamma) = \pi(\boldsymbol{\alpha})\pi(\boldsymbol{\beta})\pi(\boldsymbol{\theta})\pi(\boldsymbol{\phi})\pi(\boldsymbol{\delta})\pi(\boldsymbol{\varphi})\pi(\rho)\pi(\sigma_\tau)\pi(\sigma_\xi)\pi(\sigma_\psi)\pi(\sigma_\gamma)$. We first specify the prior distribution for the amplitude $\boldsymbol{\varphi}_l$ using a stick-breaking representation:

$\varphi_l = v_l \prod_{h=1}^{l-1} (1 - v_h)$, where $v_l \sim \text{Beta}(a_0, b_0)$ for $l < L$ and $v_L = 1$, which ensures that $\sum_{l=1}^L \varphi_l = 1$. Note that stick-breaking prior is commonly used in Bayesian nonparametric modeling such as the Dirichlet process, and $\text{Beta}(a, b)$ denotes the beta distribution with shape a and scale b . We also specify a shrinkage prior for the phase-shift, $\delta_l: \delta_l \sim N(0, \varphi_l \sigma_\delta^2)$ and $\sigma_\delta^2 \sim \text{Inv-Ga}(a_1, b_1)$. Furthermore, we assume that

$$\begin{aligned} & \boldsymbol{\alpha} \sim N_{q_1}(0, c_1 \mathbf{I}_{q_1}), \boldsymbol{\beta} \sim N_{q_2}(0, c_2 \mathbf{I}_{q_2}), \boldsymbol{\theta} \sim N_{q_3}(0, c_3 \mathbf{I}_{q_3}), \\ & \boldsymbol{\phi} \sim N_{q_4}(0, c_4 \mathbf{I}_{q_4}), \sigma_\tau \sim \text{Ga}(a_2, b_2), \sigma_\xi \sim \text{Ga}(a_3, b_3), \\ & \sigma_\psi \sim \text{Ga}(a_4, b_4), \sigma_\gamma \sim \text{Ga}(a_5, b_5), \text{ and } \rho \sim \text{Unif}(0, 1), \end{aligned} \quad (10)$$

where $l = 1, \dots, L$, $\text{Unif}(a, b)$ denotes the uniform distribution with ranged (a, b) , $\text{Ga}(a, b)$ denotes the gamma distribution with mean a/b , $\text{Inv-Ga}(a, b)$ denotes the inverse gamma

distribution with shape a and scale b , and $c_1, c_2, c_3, c_4, a_0, b_0, a_1, b_1, a_2, b_2, a_3, b_3, a_4$, and b_5 are the prespecified hyperparameters. In this article, we used $c_1 = 100, c_2 = 100, c_3 = 100, c_4 = 100, a_0 = 1, b_0 = 1, a_1 = 1, b_1 = 1, a_2 = 1, b_2 = 0.1, a_3 = 1, b_3 = 0.1, a_4 = 1, b_4 = 0.1, a_5 = 0.1$, and $b_5 = 0.1$. These choices ensure that the prior distribution for $(\alpha, \beta, \theta, \phi, \delta, \nu, \rho, \sigma_\tau, \sigma_\xi, \sigma_\psi, \sigma_\gamma, \sigma_\delta^2)$ is nearly non-informative. Let $\mathbf{v} = (v_1, v_2, \dots, v_L)'$. Based on the prior distributions in (10), the joint posterior distribution of $\alpha, \beta, \theta, \phi, \delta, \nu, \rho, \sigma_\tau, \sigma_\xi, \sigma_\psi, \sigma_\gamma$, and σ_δ^2 based on the complete data D is given by

$$\begin{aligned} & \pi(\alpha, \beta, \theta, \phi, \delta, \nu, \rho, \sigma_\tau, \sigma_\xi, \sigma_\psi, \sigma_\gamma, \sigma_\delta^2 | D) \\ & \propto L(\alpha, \beta, \theta, \phi, \delta, \nu, \rho, \sigma_\tau, \sigma_\xi, \sigma_\psi, \sigma_\gamma | D) \\ & \quad \times \pi(\alpha)\pi(\beta)\pi(\theta)\pi(\phi)\pi(\delta|\phi, \sigma_\delta^2)\pi(\nu) \\ & \quad \times \pi(\rho)\pi(\sigma_\tau)\pi(\sigma_\xi)\pi(\sigma_\psi)\pi(\sigma_\gamma)\pi(\sigma_\delta^2), \end{aligned} \quad (11)$$

where $L(\alpha, \beta, \theta, \phi, \delta, \nu, \rho, \sigma_\tau, \sigma_\xi, \sigma_\psi, \sigma_\gamma | D)$ is defined in (9). We note that by using the stick-breaking prior, we sample v_l from its conditional posterior distribution without constraints, thereby obtaining the posterior estimates of ϕ_l . Furthermore, we incorporate a shrinkage prior for δ_l that shrinks the variance of the prior distribution toward zero when the corresponding value of ϕ_l is near zero, creating stable estimation in situations where there is very little information about δ_l . A description of the MCMC algorithm is given in the Section 3.3.

3.2. Model Comparison

The model formulation in (1) includes a random effect γ_{ij} that induces overdispersion relative to the model without this effect. We use the deviance information criterion (DIC) developed by Spiegelhalter et al. (2002) to determine whether this extra source of variation is necessary in our application. Due to the nature of the latent variables and random effects, it is not easy to integrate out $\tau_i, \gamma_{ij}, \xi_i, \psi_i$, and b_{ij} analytically from model (1). Although the analytically intractable integrals may be evaluating through numerical integration or Monte Carlo methods, these methods are computationally expensive due to the high dimension integrations. Instead of integrating out all random effects to compute the DIC, we let $\zeta_{ij} = \sigma_\tau \tau_i + \sigma_\gamma \gamma_{ij} + \sigma_i(p_i h_i(k) (1 - p_i) b_{ij})$, which can not be observed. Similar to an extension of a version of the DIC discussed in Huang, Chen, and Ibrahim (2005) for generalized linear models with missing covariates, we here treated the ζ_{ij} as a parameter. Let $\Omega = (\alpha, \zeta)$, where $\zeta = (\zeta_{11}, \zeta_{12}, \dots, \zeta_{L,MK})'$. We define the deviance function as

$$\text{Dev}(\Omega) = -2\log L^*(\alpha, \zeta | D_{\text{obs}}), \quad (12)$$

where

$$L^*(\alpha, \varsigma | D_{\text{obs}}) = \prod_{i=1}^I \prod_{m=1}^M \prod_{k=1}^K \exp \left[y_{ij} (z'_i \alpha + \varsigma_{ij}) - \exp(z'_i \alpha + \varsigma_{ij}) - \log(y_{ij}!) \right].$$

The DIC is given by

$$\text{DIC} = \text{Dev}(\bar{\Omega}) + 2p_D, \quad (13)$$

Where $p_D = \overline{\text{Dev}}(\Omega) - \text{Dev}(\bar{\Omega})$, $\overline{\text{Dev}}(\Omega)$ is the posterior mean of $\text{Dev}(\Omega)$, and $\bar{\Omega}$ is the posterior mean of Ω . In (13), the first term of the DIC measures the goodness-of-fit, and p_D is the effective number of model parameters. Using the extension to the DIC as proposed by Huang et al. (2005) in the presence of missing covariates, we compute $\bar{\alpha} = E(\alpha | D_{\text{obs}})$, $\bar{\varsigma}_{ij} = E(\varsigma_{ij} | D_{\text{obs}})$, and $\overline{\text{Dev}}(\Omega) = E[\text{Dev}(\Omega) | D_{\text{obs}}]$. This way of computing the DIC is possible because we have values of ς_{ij} at each MCMC iteration and that given ς , no other parameters except α are needed. The DIC in (13) is a Bayesian measure of fit or adequacy with $2p_D$ being the respective dimension penalty term. The smaller the DIC value, the better the model fits the data. Other properties of the DIC can be found in Spiegelhalter et al. (2002) and Huang et al. (2005).

3.3. Computational Development

We present the detailed development of the MCMC sampling algorithm only for full model in (8) as the model without random effects for overdispersion have similar computational developments. Although the analytic evaluation of the joint posterior distribution of $(\alpha, \beta, \theta, \phi, \delta, v, \rho, \sigma_\tau, \sigma_\xi, \sigma_\psi, \sigma_\gamma, \sigma_\delta^2, \tau, \xi, \psi, b^*, \gamma)$ given in (11) is not possible, the proposed model allows us to develop an efficient MCMC sampling algorithm to sample from (11). We here briefly discuss how to sample from each of the conditional posterior distributions. (I) Observe that

$$\tau_i | \alpha, \beta, \theta, \phi, \delta, v, \rho, \sigma_\tau, \sigma_\xi, \sigma_\psi, \sigma_\gamma, \sigma_\delta^2, \xi, \psi, b^*, \gamma, D_{\text{obs}} \sim N \left(A_{\tau_i}^{-1} B_{\tau_i}, A_{\tau_i}^{-1} \right), \quad (14)$$

where $A_{\tau_i} = 1 + \sum_m \sum_k \frac{(1-\rho)\sigma_\tau^2}{(1+\rho)\sigma_i^2(1-p_i)^2}$ and $B_{\tau_i} = \sum_m \sum_k \frac{\sigma_\tau^2 b_{ij}^a}{(1+\rho)\sigma_i^2(1-p_i)^2}$ with

$b_{ij}^a = b_{ij}^* - \rho b_{i,j-1}^* - \sigma_\gamma (\gamma_{ij} - \rho \gamma_{i,j-1}) - (1-\rho)(z'_i \alpha + \sigma_i p_i h_i(k))$. (II) Observe that

$$\gamma_{ij} | \alpha, \beta, \theta, \phi, \delta, v, \rho, \sigma_\tau, \sigma_\xi, \sigma_\psi, \sigma_\gamma, \sigma_\delta^2, \tau, \xi, \psi, b^*, D_{\text{obs}} \sim N \left(A_{\gamma_{ij}}^{-1} B_{\gamma_{ij}}, A_{\gamma_{ij}}^{-1} \right), \quad (15)$$

where $A_{\gamma_{ij}} = \frac{(1+\rho^2)\sigma_\gamma^2}{(1-\rho^2)\sigma_i^2(1-p_i)^2} + 1$ and $B_{\gamma_{ij}} = \frac{\sigma_\gamma(\mu_{\gamma_{ij}} + \sigma_\gamma\rho\gamma_{i,j-1} - \rho(\mu_{\gamma_{i,j}+1} - \sigma_\gamma\gamma_{i,j+1}))}{(1-\rho^2)\sigma_i^2(1-p_i)^2}$

with $\mu_{\gamma_{ij}} = b_{ij}^* - \rho b_{i,j-1}^* - (1-\rho)(z_i'\alpha + \sigma_\tau\tau_i + \sigma_i p_i h_i(k))$. (III) The conditional posterior density for $[b^*|\alpha, \beta, \theta, \phi, \delta, v, \rho, \sigma_\tau, \sigma_\xi, \sigma_\psi, \sigma_\gamma, \sigma_\delta^2, \tau, \xi, \psi, \gamma, D_{\text{obs}}]$ has the form

$$\begin{aligned} & [b_{ij}^*|\alpha, \beta, \theta, \phi, \delta, v, \rho, \sigma_\tau, \sigma_\xi, \sigma_\psi, \sigma_\gamma, \sigma_\delta^2, \tau, \xi, \psi, \gamma, D_{\text{obs}}] \\ & \propto \exp[y_{ij}b_{ij}^* - \exp(b_{ij}^*)] \\ & \times \exp\left[-\frac{(b_{ij}^* - \mu_{ij} - \rho(b_{i,j-1}^* - \mu_{i,j-1}))^2}{2(1-\rho^2)\sigma_i^2(1-p_i)^2}\right] \\ & \times \exp\left[-\frac{(b_{i,j+1}^* - \mu_{i,j+1} - \rho(b_{i,j}^* - \mu_{i,j}))^2}{2(1-\rho^2)\sigma_i^2(1-p_i)^2}\right], \end{aligned} \quad (16)$$

where μ_{ij} is given in (8). The expression in (16) is log-concave and we use the adaptive rejection algorithm of Gilks and Wild (1992) to sample b_{ij}^* from

$[b_{ij}^*|\alpha, \beta, \theta, \phi, \delta, v, \rho, \sigma_\tau, \sigma_\xi, \sigma_\psi, \sigma_\gamma, \sigma_\delta^2, \tau, \xi, \psi, \gamma, D_{\text{obs}}]$. (IV) The conditional posterior density for $[\sigma_\gamma|\alpha, \beta, \theta, \phi, \delta, v, \rho, \sigma_\tau, \sigma_\xi, \sigma_\psi, \sigma_\delta^2, \tau, \xi, \psi, b^*, \gamma, D_{\text{obs}}]$ has the form

$$\begin{aligned} & [\sigma_\gamma|\alpha, \beta, \theta, \phi, \delta, v, \rho, \sigma_\tau, \sigma_\xi, \sigma_\psi, \sigma_\delta^2, \tau, \xi, \psi, b^*, \gamma, D_{\text{obs}}] \\ & \propto \prod_{i=1}^I \prod_{m=1}^M \prod_{k=1}^K \exp\left[-\frac{(b_{ij}^* - \mu_{ij} - \rho(b_{i,j-1}^* - \mu_{i,j-1}))^2}{2(1-\rho^2)\sigma_i^2(1-p_i)^2}\right] \\ & \times (\sigma_\gamma)^{a_5-1} \exp(-b_5\sigma_\gamma), \end{aligned} \quad (17)$$

where μ_{ij} is given in (8). The expression in (17) is log-concave and we use the adaptive rejection algorithm of Gilks and Wild (1992) to sample σ_γ from

$[\sigma_\gamma|\alpha, \beta, \theta, \phi, \delta, v, \rho, \sigma_\tau, \sigma_\xi, \sigma_\psi, \sigma_\delta^2, \tau, \xi, \psi, b^*, \gamma, D_{\text{obs}}]$. (V) The conditional posterior density for $[\sigma_\tau|\alpha, \beta, \theta, \phi, \delta, v, \rho, \sigma_\xi, \sigma_\psi, \sigma_\gamma, \sigma_\delta^2, \tau, \xi, \psi, b^*, \gamma, D_{\text{obs}}]$ has the form

$$\begin{aligned} & [\sigma_\tau|\alpha, \beta, \theta, \phi, \delta, v, \rho, \sigma_\xi, \sigma_\psi, \sigma_\gamma, \sigma_\delta^2, \tau, \xi, \psi, b^*, \gamma, D_{\text{obs}}] \\ & \propto \prod_{i=1}^I \prod_{m=1}^M \prod_{k=1}^K \exp\left[-\frac{(b_{ij}^* - \mu_{ij} - \rho(b_{i,j-1}^* - \mu_{i,j-1}))^2}{2(1-\rho^2)\sigma_i^2(1-p_i)^2}\right] \\ & \times (\sigma_\tau)^{a_2-1} \exp(-b_2\sigma_\tau), \end{aligned} \quad (18)$$

where μ_{ij} is given in (8). The expression in (18) is log-concave and we use the adaptive rejection algorithm of Gilks and Wild (1992) to sample σ_τ from

$[\sigma_\tau | \alpha, \beta, \theta, \phi, \delta, v, \rho, \sigma_\xi, \sigma_\psi, \sigma_\gamma, \sigma_\delta^2, \tau, \xi, \psi, b^*, \gamma, D_{\text{obs}}]$. (VI) Observe that

$$\alpha | \beta, \theta, \phi, \delta, v, \rho, \sigma_\tau, \sigma_\xi, \sigma_\psi, \sigma_\gamma, \sigma_\delta^2, \tau, \xi, \psi, b^*, \gamma, D_{\text{obs}} \sim N(A_\alpha^{-1} b_\alpha, A_\alpha^{-1}), \quad (19)$$

where $A_\alpha = c_1 I_{q_1} + \sum_i \sum_m \sum_k \frac{(1-\rho)z_i z'_i}{(1+\rho)\sigma_i^2(1-p_i)^2}$ and $b_\alpha = \sum_i \sum_m \sum_k \frac{z_i b_{ij}^b}{(1+\rho)\sigma_i^2(1-p_i)^2}$ with

$b_{ij}^b = b_{ij}^* - \rho b_{i,j-1}^* - \sigma_\gamma(\gamma_{ij} - \rho\gamma_{i,j-1}) - (1-\rho)(\sigma_\tau\tau_i + \sigma_i p_i h_i(k))$. (VII) We apply the collapsed Gibbs technique of Liu (1994) via the following identity:

$$\begin{aligned} & [\delta, \sigma_\delta^2 | \alpha, \beta, \theta, \phi, v, \rho, \sigma_\tau, \sigma_\xi, \sigma_\psi, \sigma_\gamma, \tau, \xi, \psi, b^*, \gamma, D_{\text{obs}}] \\ &= [\sigma_\delta^2 | \delta, \alpha, \beta, \theta, \phi, v, \rho, \sigma_\tau, \sigma_\xi, \sigma_\psi, \sigma_\gamma, \tau, \xi, \psi, b^*, \gamma, D_{\text{obs}}] \quad (20) \\ &\times [\delta | \alpha, \beta, \theta, \phi, v, \rho, \sigma_\tau, \sigma_\xi, \sigma_\psi, \sigma_\gamma, \tau, \xi, \psi, b^*, \gamma, D_{\text{obs}}]. \end{aligned}$$

That is, we sample δ after collapsing on σ_δ^2 . Observe that

$$\sigma_\delta^2 | \delta, \alpha, \beta, \theta, \phi, v, \rho, \sigma_\tau, \sigma_\xi, \sigma_\psi, \sigma_\gamma, \tau, \xi, \psi, b^*, \gamma, D_{\text{obs}} \sim \text{Inv-Ga}\left(a_1 + \frac{L}{2}, b_1 + \frac{1}{2} \sum_{l=1}^L \frac{\delta_l^2}{\varphi_l}\right), \quad (21)$$

where $\varphi_l = v_l \prod_{h=1}^{l-1} (1 - v_h)$. Finally, (VIII) the conditional posterior densities for $\beta, \theta, \phi, \delta, v, \rho, \sigma_\xi, \sigma_\psi, \xi$, and ψ do not have closed form. We use the Metropolis–Hastings algorithm (Hastings 1970) to sample $\beta, \theta, \phi, \delta, v, \rho, \sigma_\xi, \sigma_\psi, \xi$, and ψ from their conditional posterior distributions.

4. Simulation Study

We perform a simulation study to illustrate the proposed model. The simulation is constructed to closely match the data analysis in Section 5. Specifically, we simulate data with the same number of subjects ($I = 89$) and follow-up times as for the application, and we fix the covariates at their observed values. Furthermore, we incorporate covariates that were significant in the application, and evaluate the performance of the model with $L = 7$, which corresponds to the model fit to the application data where the amplitude values for φ_l are close to zero for $l > 7$. We generate random effects as $\tau_i \sim N(0, 1)$, $\gamma_{ij} \sim N(0, 1)$, $\xi_i \sim N(0, 1)$, and $\psi_i \sim N(0, 1)$. In addition, we generate the individual stochastic variation, b_{ij} , as in (2). True parameter values were chosen close to the ones estimated in the analysis and 500

simulated datasets were generated. The prior distributions used were those specified in Section 3.1. We used 20,000 Gibbs samples to compute all the posterior estimates, including the mean and 95% highest posterior density (HPD) intervals after a burn-in of 10,000 iterations.

Table 1 presents mean of the posterior means (Mean), standard deviation of the posterior means (SD), and the coverage probability (CP) of the 95% HPD intervals for the simulation study. The estimates for all parameters in Table 1 are close to the true values, and are therefore nearly unbiased. In addition, the coverage rates for 95% HPD intervals are close to 95%.

5. Physical Activity in Teenage Girls

We used the proposed latent variable Poisson model (1) to analyze the longitudinal physical activity count data discussed in Section 1. The analytic goal was to investigate the circadian pattern as well as the proportion of circadian versus stochastic short-term trend (degree of irregularity in the pattern across time) across participants in these data.

The response variable Y_{ij} is a physical activity count over a 15-minute period for the i th individual at the j th time interval. We consider four covariates for investigation: body mass index (BMI), race, family affluence score (FAS), and parental education (PEd). BMI is dichotomized (1 if BMI ≥ 30 kg/m² (obese); 0 otherwise (no obese)), Race is also dichotomized ((Race1, Race2) = (0, 0) for Hispanic, (1, 0) for African American, and (0, 1) for Others). The covariate PEd was reported using a seven point scale (1 if less than a high school diploma; 2 if high school diploma; 3 if GED; 4 if some college or technical school; 5 if associates degree; 6 if bachelors degree; or 7 if graduate degree) from the highest level of education completed by the first guardian in the household. The adolescent provided an estimate of family socioeconomic status using the FAS, which is a validated measure of socioeconomic status. The possible scores of FAS are from 0 to 9. The covariates PEd and FAS were treated as continuous.

Two subjects were removed from the analysis that had long periods of zero or nearly zero counts which were not detected by the watch's nonwear algorithm (See Figure S in Section S1 of the supplementary materials). In total, we have $I = 89$ subjects with complete covariate information, $M = 7$ days, $K = 96$ observations for each day, and $J = 672$ observations for each subject. Table 2 presents some descriptive statistics for the 89 participants included in the analysis: number of observations and percent for categorical variables, mean and standard deviation (SD) for continuous variables. Since all participating girls were measured in the summer of the 9th grade, it was anticipated that most individuals would have a nonconstrained schedule since they were unlikely to have summer jobs (although these data were not explicitly collected). Also, Figure 2 (a) shows little variation across the days (Monday to Sunday) suggesting that a day of the week effect is not necessary.

To help the numerical stability and to improve convergence for the MCMC sampling algorithm, we standardized all covariates. The means and standard deviations are (0.169, 0.377) for obesity, (0.258, 0.440) for race1, (0.371, 0.486) for race2, (5.045, 1.566) for PAS,

and (3.708, 1.779) for PEd, respectively. In all the Bayesian computations, we used 20,000 Gibbs samples, which were taken from every fifth iteration, after a burn-in of 4000 iterations for each model, to compute all the posterior estimates, including means (Estimates), SDs, 95% HPD intervals, as well as the DICs. The computer programs were written in FORTRAN 95 using IMSL subroutines with double precision accuracy. The convergence of the MCMC sampling algorithm for all the parameters was checked based on the recommendations of Cowles and Carlin (1996). All trace and autocorrelation plots showed good convergence and mixing of the MCMC sampling algorithm (See trace plots in Section S3 of the supplementary materials).

We fit the model with different numbers of harmonic functions ($L = 5, 10, 15$, and 20) to characterize the circadian pattern in physical activity and to examine the sensitivity of inferences to the size of L . The posterior estimates, including the means, SDs, and 95% HPD intervals of the parameters for the model with $L = 10$ are reported in Tables 3 and 4. The posterior estimates for the model with $L = 5, 15$, and 20 were similar and the results are given in Tables in Section S2 of the supplementary materials. Furthermore, the posterior estimates of ϕ_l for model with $L = 10, 15$, and 20 are close to zero for $l > 7$, suggesting that the model with $L = 7$ is enough to fully characterize the circadian pattern in physical activity. The model allows us to examine the differing effects of covariates on the overall mean structure ($z_i^T \alpha$), the shape-invariant scale function (σ_i), the phase-shift (ζ_{ij}), and the mixing parameter (p_i).

From Table 3, mean physical activity was only influenced by obesity, where heavier teenage girls had lower overall levels. The scale parameter (σ_i), which reflects the amplitude or inherent variation of the circadian/stochastic process, significantly increased with obesity and family wealth (FAS), and decreased for African Americans and with parental education (PEd). The phase-shift (ζ_{ij}) was shifted later for obese participants compared with their nonobese counterparts, and earlier with increasing parental education. With regard to the relative contribution of circadian to stochastic short-term trend, 95% HPD intervals include zero for all covariates, which demonstrates little or no effect of the covariates. However, the variation across individuals, although not explained by these covariates, was large (i.e., the 5th and 95th percentiles of p_i are 0.31 to 0.68, respectively). This confirms the observed individual variation in the regularity of circadian patterns shown in Figure 1. The posterior estimate of ρ (a serial correlation for b_{ij}) is 0.7895, demonstrating a strong dependence structure in the stochastic variation characterizing the stochastic short-term trend. Furthermore, from Table 4, the posterior estimates of ϕ for the model with $L = 10$ are close to zero for $l > 7$ (the same results for $L = 15$ and 20), and 95% HPD intervals for δ include zero for $l = 4$ and $l > 7$, demonstrating that including additional harmonic terms beyond 7 is not necessary for this application.

As a sensitivity analysis, we also fit a model without overdispersion. We did this to (i) examine whether overdispersion is important in the modeling framework; (ii) provide a sensitivity analysis for examining regression effects under a potentially misspecified model; (iii) examine the effect of excluding this variation term on inferences about the mixing parameters p_i since the overdispersion adds additional independent variation that is not

included in the stochastic short-term trend b_{ij} . Based on the DIC, the model with overdispersion showed a substantially better fit than the reduced model without overdispersion, indicating the importance of including this parameter in our formulation. DIC for the model with $L = 10$ are reported in Table 5. Specifically, for the model that did not account for overdispersion, we found that the value of the DIC ($L = 10$) is 629088.97, which is higher than the DIC for the proposed model. Furthermore, the model with $L = 5, 15$, and 20 had similar results (data not shown). In addition, we found similar results for p_i with both models. Specifically, we found that no covariates influenced p_i , and that the 5th and 95th percentiles of p_i were 0.32 and 0.58, respectively, for the model without overdispersion which is close to the values reported previously for the model with overdispersion. Further, parameter estimates for the other regression coefficients were similar between the two models (data not shown).

Figure 3 shows the posterior estimates of the circadian pattern, $\sigma_i h_i(k)$, in physical activity over a 24-hour period and presents both the population-averages of the posterior estimates as well as individual plots of five students chosen at random for the model with overdispersion. The plots demonstrate a circadian rhythm with relatively constant physical activity during the day and decreased physical activity during the sleeping hours. Additionally, the plot demonstrates sizable between-subject variation in the circadian pattern across students, with the variation manifesting itself in amplitude as well as phase-shift variation. Figure 4(a) is a plot of residuals against the fitted values, defined as $\exp(\hat{\lambda}_{ij})$, where $\hat{\lambda}_{ij}$ is the posterior mean of λ_{ij} defined in (1). There are no discernible patterns in this residual plot, which suggests that the model fits the data well. In addition, Figure 4 (b) presents an ACF plot of the residuals and shows that the residuals are nearly independent across time, demonstrating that the proposed model describes the dependence structure well. Figure 5 shows a plot of the estimated overall mean over time, $E(\exp(\lambda_{ij}))$ in (1) obtained from the posterior samples of the parameters. This plot adjusts for obesity, race, FAS, and PED in the mean and variance structure, and takes full advantage of the specification of our flexible model structure. The figure demonstrates a pronounced circadian pattern over the 7-day followup that is close to the one described in Figure 2. Figure 6 shows the posterior estimates of the mixing parameter p_i . The figure shows that the participants have a large range in p_i with values ranging from 0.2283 to 0.669 among the 89 participants. Interestingly, since a majority of the estimated p_i 's are below 0.5, these results demonstrate that, for a majority of participants, the short-term stochastic trend dominates the circadian pattern.

6. Discussion

The focus of this article was on characterizing the heterogeneity in the degree of circadian versus stochastic variation (irregular short-term trend) in longitudinal physical activity counts across time among teenage girls. This new formulation allows investigators to examine the effect of covariates on multiple features of these complex patterns including: (i) overall mean structure; (ii) scale or amplitude parameter of the circadian pattern; (iii) phase-shift of the circadian pattern; and (iv) the degree of circadian versus short-term trend manifested by a stochastic process. Others (Wang, Ke, and Brown 2003; Albert and Hunsberger 2005) have considered shape-invariant formulations for continuous longitudinal

data that have addressed (i)–(iii). The assessment of the degree of circadian versus short-term trend was motivated by the analysis of physical activity data where we saw some individuals that showed pronounced circadian patterns while others showed clear departures from these patterns. Incorporating this variation presented unique methodological challenges that, for computational efficiency, were addressed from a Bayesian perspective.

The application of Bayesian methodology to this application has some novel aspects. First, we use a stick-breaking representation for estimating the amplitudes, φ_j . This approach, which has been primarily used in nonparametric Bayes methodology such as a Dirichlet process, allows us to avoid inherent constraints in sampling from the conditional posterior distribution. Second, in order to avoid estimation instability in estimating the phase-shifts δ_j associated with amplitudes that are small, we proposed a shrinkage prior with variance $\varphi_j \sigma_\delta^2$, whereby the prior distribution is heavily centered at zero in situations where the amplitudes are small and there is very little information about the phase-shift. Third, a hierarchical centering reparametrization was employed in order to improve convergence of Gibbs sampler.

The analysis demonstrated some interesting findings including that obesity was associated with a decreased overall mean structure, an increased shape or amplitude parameter, and a delay in the phase of the circadian pattern among teenage girls. Interestingly, although we saw marked variation among individuals in the proportion of stochastic variation that can be explained by the circadian pattern (circadian pattern versus stochastic short-term trend), this variation was not explained by any of the known demographic covariates. It is possible that this variation may be explained by other biologically based markers that were not available to us. It is also possible that large or small values in this proportion may be indicative of subsequent teenage or early adult behavior. This is an area of future research.

The model was developed for longitudinal physical activity count data, and to directly answer analytic questions for this physical activity study. However, the model and estimation procedures can be adapted to other generalized linear model outcomes quite simply. We assume Gaussian random effects in this formulation which describe these data well. However, for other applications more flexible random effect distributions can be used to more flexibly model individual and within subject variation across time. Although the shape-invariant parameterization is appealing in its model parsimony, it may not be appropriate in situations where individuals may have very different underlying patterns from each other. Alternative parameterizations may be needed in this case.

The model distinguishes between circadian and stochastic variation, which requires relatively long sequences of data on each individual. In fact, we could not distinguish between these two sources of variation without observing more than one daily cycle of longitudinal count data. Fortunately we can distinguish heterogeneity in the sources of variation with the seven days of follow-up observed in this study.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding

This research was supported by the intramural research program of the National Cancer Institute.

References

- Albert PS, Follmann DA, Wang SA, and Suh EB (2002), “A Latent Autoregressive Model for Longitudinal Binary Data Subject to Informative Missingness,” *Biometrics*, 58, 631–642. [1993] [PubMed: 12229998]
- Albert PS, and Hunsberger S (2005), “On Analyzing Circadian Rhythms Data Using Nonlinear Mixed Models with Harmonic Terms,” *Biometrics*, 61, 1115–1120. [1992,993,995,1000] [PubMed: 16401286]
- Brown EN, and Czeisler CA (1992), “The Statistical Analysis of Circadian Phase and Amplitude in Constant-routine Core-temperature Data,” *Journal of Biological Rhythms*, 7, 177–202. [1992] [PubMed: 1421473]
- Chen M-H, and Ibrahim JG (2000), “Bayesian Predictive Inference for Time Series Count Data,” *Biometrics*, 56, 678–685. [1993] [PubMed: 10985202]
- Cowles C, and Carlin BP (1996), “Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review,” *Journal of the American Statistical Association*, 91, 883–904. [1998]
- Gelfand AE, Sahu SK, and Carlin BP (1996), “Efficient Parametrisations for Generalized Linear Mixed Models” (with discussion), in *Bayesian Statistics 5*, eds. by Bernardo JM, Berger JO, Dawid AP and Smith AFM, Oxford, Oxford University Press [1995]
- Gilks WR, and Wild P (1992), “Adaptive Rejection Sampling for Gibbs Sampling,” *Applied Statistics*, 41, 337–348. [1996,997]
- Greenhouse J, Kass R, and Tsay R (1987), “Fitting Nonlinear Models with ARMA Errors to Biological Rhythm Data,” *Statistics in Medicine*, 55, 239–244. [1992]
- Harvey AC (1989), *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge: Cambridge University Press [1993]
- Hastings WK (1970), “Monte Carlo Sampling Methods Using Markov Chains and Their Applications,” *Biometrika*, 57, 97–109. [1997]
- Hay JL, and Pettitt A (2001), “Bayesian Analysis of Time Series of Counts,” *Biostatistics*, 2, 433–444. [1993] [PubMed: 12933634]
- Hu K, Scheer FA, Ivanov PC, Buijs RM, and Shea SA (2007), “The Suprachiasmatic Nucleus Functions Beyond Circadian Rhythm Generation,” *Neuroscience*, 149, 508–517. [1992] [PubMed: 17920204]
- Huang L, Chen M-H, and Ibrahim JG (2005), “Bayesian Analysis for Generalized Linear Models with Nonignorable Missing Covariates,” *Biometrics*, 61, 767–780. [1996] [PubMed: 16135028]
- Huettel SA, Song AW, and McCarthy G (2004), *Functional Magnetic Resonance Imaging*, Sunderland, UK: Sinauer Associates [1993]
- Hurtgen H, and Gervini D (2009), “Semiparametric Shape-invariant Models for Periodic Data,” *Journal of Applied Statistics*, 36, 1055–1065. [1993]
- Ibrahim JG, Chen M-H, and Ryan LM (2000), “Bayesian Variable Selection for Time Series Count Data,” *Statistica Sinica*, 10, 971–987. [1995]
- Indic P, Murray G, Maggini C, Amore M, Meschi T, Borghi L, Baldessarini RJ, and Salvatore P (2012), “Multi-scale Motility Amplitude Associated with Suicidal Thoughts in Major Depression,” *PLoS One*, 7, e38761 [1992] [PubMed: 22701706]
- Karim MR, and Zeger SL (1992), “Generalized Linear Models with Random Effects—salamander Mating Revisited,” *Biometrics*, 48, 631–644. [1993] [PubMed: 1637985]
- Kim S, Chen Z, Zhang Z, Simons-Morton BG, and Albert PS (2013), “Bayesian Hierarchical Poisson Regression Models, An Application to a Driving Study with Kinematic Events,” *Journal of the American Statistical Association*, 108, 495–503. [1993,995]

- Kushida C, Chang A, Gadkary C, Guilleminault C, Carrillo O, and Dement W (2001), "Comparison of Actigraphic, Polysomnographic, and Subjective Assessment of Sleep Parameters in Sleep-disordered Patients," *Sleep Medicine*, 2, 389–396. [992] [PubMed: 14592388]
- Lawton WH, and Sylvestre EA (1971), "Self-Modeling Curve Resolution," *Technometrics*, 12, 617–633. [993]
- Lawton WH, Sylvestre EA, and Maggio MS (1972), "Self-Modeling Linear Regression," *Technometrics*, 13, 513–532. [993]
- Lindstrom MJ (1995), "Self Modeling with Random Shift and Scale Parameters and Free-knot Spline Shape Function," *Statistics in Medicine*, 14, 2009–2021. [993] [PubMed: 8677401]
- Liu JS (1994), "The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a Gene Regulation Problem," *Journal of the American Statistical Association*, 89, 985–966. [997]
- McCulloch CE, R. S, and Neuhaus JM (2008), *Generalized, Linear, and Mixed Models*, 2nd ed New York: Wiley [993]
- Ogbagaber S, Albert P, Lewin D, and Iannotti RJ (2012), "Summer Activity Patterns Among Teenage Girls: Harmonic Shape Invariant Modeling to Estimate Circadian Cycles," *Journal of Circadian Rhythms*, 10, 2 [992,994] [PubMed: 22559328]
- Ramsay JO, and Silverman BW (2005), *Functional Data Analysis (Second Edition)*, New York: Springer [993]
- Smith JQ (1979), "A Generalization of the Bayesian Forecasting Model," *Journal of Royal Statistical Society, Series B*, 41, 375–387. [993]
- Spiegelhalter DJ, Best NG, Carlin BP, and van der Linde A (2002), "Bayesian Measures of Model Complexity and Fit" (with discussion), *Journal of Royal Statistical Society, Series B*, 64, 583–639. [996]
- Wang Y, and Brown MB (1996), "A Flexible Model for Human Circadian Rhythms," *Biometrics*, 52, 588–596. [992,993] [PubMed: 8672704]
- Wang Y, Ke C, and Brown MB (2003), "Shape-Invariant Modeling of Circadian Rhythms with Random Effects and Smoothing Spline ANOVA Decompositions," *Biometrics*, 59, 804–812. [992,993,995,1000] [PubMed: 14969458]
- Zeger SL (1988), "A Regression Model for Time Series of Counts," *Biometrika*, 75, 621–629. [993]
- Zeger SL, and Diggle PJ (1994), "Semi-parametric Models for Longitudinal Data with Application to CD4 cell Numbers in HIV Seroconverters," *Biometrics*, 50, 689–699. [993] [PubMed: 7981395]
- Zhang D, Lin X, Raz J, and Sowers MF (1998), "Semiparametric Stochastic Mixed Models for Longitudinal Data," *Journal of the American Statistical Association*, 93, 710–719. [993]

Figure 1.

Individual plots of three students showing their physical activity over 7 days: the solid lines represent a lowess smooth curve. Each tick represents a four-hour period. Each point reflects a 15 minute with the first point representing the count from 12:00 am to 12:15 am on the first day.

Figure 2.

Plots of the overall mean trajectory and autocorrelation. Each tick represents a 4-hour period, and each point reflects a 15 minute with the first point representing the count from 12:00 am to 12:15 am on the first day in (a). The lag in (b) is in units of 15 minutes.

Figure 3.

Plots of the posterior mean for five individual circadian patterns, $\sigma_i h_i(k)$, for the model with $L = 10$. The overall curve defined by setting the random effects in $h_i(k)$ to zero, is also presented. Each tick represents a 4-hour period. Each point reflects a 15 minute with the first point representing the count from 12:00 am to 12:15 am.

Figure 4.

Residual plots for the model with $L = 10$: (a) residuals versus fitted values (the solid line represent a lowess smooth curve). Fitted values are defined as the posterior mean of $\exp(\lambda_{ij})$; (b) autocorrelation where the lag term is in units of 15 minutes.

Figure 5.

Estimated overall mean trajectory over 7 days for the model with $L = 10$. Each tick represents a 4-hour period. Each point reflects a 15 minute with the first point representing the count from 12:00 am to 12:15 am on the first day.

Figure 6.
Plots of individual estimated mixing parameters with $L = 10$.

Table 1.

Simulation results for the model with $L = 7$.

Component	Variable	Parameter	True	Mean	SD	CPo95%HPD
Mean	Intercept	α_1	6.73	6.729	0.052	95.6
	Obesity	α_2	-0.12	-0.120	0.055	95.2
Scale (σ_i)	Intercept	θ_1	1.28	1.275	0.006	94.6
	Obesity	θ_2	0.02	0.020	0.006	94.8
	Race1	θ_3	-0.03	-0.031	0.008	96.2
	Race2	θ_4	0.01	0.010	0.008	95.0
	Family affluence	θ_5	0.03	0.030	0.007	96.4
	Parent education	θ_6	-0.03	-0.030	0.008	96.2
	Intercept	β_1	-0.03	-0.042	0.040	95.4
Mixing (p_i) Harmonic ($h_i(k)$)	Intercept	δ_1	-0.90	-0.892	0.026	95.4
		δ_2	0.17	0.174	0.027	95.2
		δ_3	-0.48	-0.470	0.037	95.0
		δ_4	0.05	0.054	0.029	95.6
		δ_5	0.31	0.312	0.030	95.4
		δ_6	-0.11	-0.096	0.043	96.6
		δ_7	0.10	0.084	0.036	96.6
	Obesity	ϕ_1	-0.10	-0.102	0.026	95.6
	Parent education	ϕ_2	0.13	0.130	0.028	96.0
		ϕ_1	0.80	0.806	0.007	95.4
		ϕ_2	0.158	0.156	0.007	97.4
		ϕ_3	0.009	0.010	0.002	96.8
		ϕ_4	0.019	0.018	0.002	94.8
		ϕ_5	0.009	0.008	0.002	96.0
		ϕ_6	0.003	0.003	0.001	94.4
		ϕ_7	0.002	0.002	0.001	95.0
		σ_γ	1.11	1.110	0.009	95.8

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Component	Variable	Parameter	True	Mean	SD	CPo95%HPD
		σ_{τ}	0.47	0.472	0.043	95.6
		σ_{ξ}	0.39	0.397	0.032	95.2
		σ_{η}	0.25	0.253	0.022	94.2
		ρ	0.79	0.790	0.004	97.8
		σ_{δ}^2		8.278	1.612	

Table 2.Descriptive statistics of the physical activity count data ($I = 89$).

Characteristic	N (%) or Mean (SD)
Obesity	
0: No obese (BMI < 30)	74 (83.15)
1: Obese (BMI ≥ 30)	15 (16.85)
Race	
1: Hispanic	33 (37.08)
2: African American	23 (25.84)
3: Other	33 (37.08)
Family Affluence Score	5.05 (1.57)
Parental Education	3.71 (1.78)

Table 3.

Posterior estimates of the parameters for the model with $L = 10$.

Component	Variable	Parameter	Posterior mean	Posterior SD	95% HPD interval
Mean	Intercept	α_1	6.7298	0.0551	(6.6201, 6.8362)
	Obese	α_2	-0.1244	0.0572	(-0.2379, -0.0126)
	Race1	α_3	0.0577	0.0662	(-0.0708, 0.1885)
	Race2	α_4	0.0381	0.0706	(-0.0963, 0.1813)
	Family affluence	α_5	-0.0152	0.0626	(-0.1347, 0.1111)
	Parent education	α_6	0.0410	0.0651	(-0.0880, 0.1655)
Scale (σ_i)	Intercept	θ_1	1.2779	0.0076	(1.2625, 1.2924)
	Obese	θ_2	0.0172	0.0074	(0.0032, 0.0322)
	Race1	θ_3	-0.0279	0.0086	(-0.0445, -0.0108)
	Race2	θ_4	0.0053	0.0091	(-0.0128, 0.0228)
	Family affluence	θ_5	0.0309	0.0080	(0.0155, 0.0468)
	Parent education	θ_6	-0.0305	0.0082	(-0.0462, -0.0146)
Mixing (p_i)	Intercept	β_1	-0.0283	0.0442	(-0.1139, 0.0580)
	Obese	β_2	-0.0597	0.0458	(-0.1490, 0.0313)
	Race1	β_3	-0.0430	0.0540	(-0.1471, 0.0642)
	Race2	β_4	0.0062	0.0568	(-0.1079, 0.1163)
	Family affluence	β_5	0.0032	0.0515	(-0.0971, 0.1042)
	Parent education	β_6	0.0193	0.0527	(-0.0825, 0.1263)
Harmonic ($h_i(k)$)	Obese	ϕ_1	-0.0948	0.0284	(-0.1509, -0.0387)
	Race1	ϕ_2	0.0148	0.0331	(-0.0491, 0.0805)
	Race2	ϕ_3	0.0538	0.0348	(-0.0154, 0.1209)
	Family affluence	ϕ_4	-0.0520	0.0317	(-0.1147, 0.0098)
	Parent education	ϕ_5	0.1248	0.0329	(0.0602, 0.1903)
		σ_γ^2	1.1109	0.0124	(1.0867, 1.1349)
		σ_τ	0.4688	0.0445	(0.3865, 0.5590)

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Component	Variable	Parameter	Posterior mean	Posterior SD	95% HPD interval
		σ_{ξ}	0.3920	0.0350	(0.3275, 0.4631)
		σ_{ψ}	0.2488	0.0214	(0.2071, 0.2900)
		ρ	0.7895	0.0059	(0.7778, 0.8008)
		σ_{δ}^2	9.0746	7.1222	(1.8304, 21.5375)

Table 4. Posterior estimates of the parameters for the harmonic function, $h_j(k)$, for the model with $L = 10$.

Para- meter	Posterior mean	Posterior SD	95% HPD interval	Para- meter	Posterior mean	Posterior SD	95% HPD interval
δ_1	-0.90308	0.03017	(-0.96065, -0.84191)	ϕ_1	0.80039	0.00950	(0.78096, 0.81817)
δ_2	0.17105	0.02982	(0.11465, 0.23100)	ϕ_2	0.15755	0.00847	(0.14109, 0.17412)
δ_3	-0.48327	0.03980	(-0.56168, -0.40469)	ϕ_3	0.00942	0.00220	(0.00524, 0.01370)
δ_4	0.04540	0.03172	(-0.01762, 0.10663)	ϕ_4	0.01866	0.00331	(0.01228, 0.02508)
δ_5	0.30807	0.03308	(0.24583, 0.37524)	ϕ_5	0.00944	0.00234	(0.00497, 0.01405)
δ_6	-0.11175	0.04004	(-0.19221, -0.03294)	ϕ_6	0.00249	0.00124	(0.00044, 0.00499)
δ_7	0.09726	0.04458	(0.00758, 0.18547)	ϕ_7	0.00170	0.00109	(1.73×10^{-11} , 0.00374)
δ_8	0.00326	0.07515	(-0.20146, 0.20229)	ϕ_8	0.00018	0.00037	(9.11×10^{-11} , 0.00096)
δ_9	0.00477	0.02816	(-0.04444, 0.06882)	ϕ_9	0.00009	0.00017	(7.58×10^{-11} , 0.00041)
δ_{10}	0.00283	0.03929	(-0.08496, 0.08826)	ϕ_{10}	0.00008	0.00018	(3.01×10^{-11} , 0.00037)

Table 5.DIC Values for the model with $L = 10$.

Model	Dev ($\bar{\Omega}$)	p_D	DIC
Overdispersion	513686.81	57659.52	629005.85
No overdispersion	513777.54	57655.72	629088.97