

# **HHS Public Access**

Author manuscript *Stat Med.* Author manuscript; available in PMC 2018 November 10.

Published in final edited form as:

Stat Med. 2017 November 10; 36(25): 4028–4040. doi:10.1002/sim.7401.

# A Joint Modeling and Estimation Method for Multivariate Longitudinal Data with Mixed Types of Responses to Analyze Physical Activity Data Generated by Accelerometers

Haocheng Li<sup>a,\*</sup>, Yukun Zhang<sup>b</sup>, Raymond J. Carroll<sup>c,d</sup>, Sarah Kozey Keadle<sup>e</sup>, Joshua N. Sampson<sup>f</sup>, and Charles E. Matthews<sup>f</sup>

<sup>a</sup>Departments of Oncology and Community Health Sciences, University of Calgary

<sup>b</sup>Department of Oncology, University of Calgary

<sup>c</sup>Department of Statistics, Texas A&M University

<sup>d</sup>School of Mathematical and Physical Sciences, University of Technology Sydney

<sup>e</sup>Kinesiology Department, California Polytechnic State University

<sup>f</sup>National Cancer Institute

## Abstract

A mixed effect model is proposed to jointly analyze multivariate longitudinal data with continuous, proportion, count and binary responses. The association of the variables is modeled through the correlation of random effects. We use a quasilikelihood type approximation for non-linear variables, and transform the proposed model into a multivariate linear mixed model framework for estimation and inference. Via an extension to the EM approach, an efficient algorithm is developed to fit the model. The method is applied to physical activity data, which uses a wearable accelerometer device to measure daily movement and energy expenditure information. Our approach is also evaluated by a simulation study.

### Keywords

Accelerometers; Longitudinal data; Mixed effects model; Multivariate longitudinal data; Penalized Quasilikelihood

## 1. Introduction

We propose a new framework to estimate multivariate longitudinal data that consists of multiple types of variables. Our methodology is motivated by physical activity data recorded by wearable accelerometer devices [1]. These devices have several advantages over traditional questionnaire-based measures of physical activity and are increasingly used in

<sup>\*</sup>Correspondence to: 2500 University Drive NW, Calgary, T2N 1N4, Canada. haocheng.li@ucalgary.ca.

Supplementary Materials

Supporting tables referenced in Sections 4 and 6 are available with this paper at the Statistics in Medicine website on Wiley Online Library.

studies investigating physical activity and health. One clear advantage is the ability to measure activity continuously over days or weeks at relatively high frequencies (e.g., 80 Hz). However, since this data can be summarized into many metrics that are potentially associated, the increased information presents new analytic challenges.

Recent literature [2] shows that a complete description of an individual's pattern of physical activity requires specification of multiple variables, such as those listed in Figure 1 (a)–(h). Among these eight variables, we note that two are continuous measures, two are proportions, two are counts, and two are binary. Our motivating example is an experimental study that examined the impact of 12-weeks of exercise training on health. The 63 participants wore monitors during baseline and weeks 3, 6, 9 and 12 of the intervention and we extracted the eight variables for one day's wear in each week.

The main purposes of this work are to develop statistical methodology for jointly modeling the longitudinal pattern of the eight factors, which postulates their association structures. The association pattern can be widely used in this physical activity data analysis, including but not limit to, study daily sedentary time and energy expenditure levels under certain conditions. For example, we could explore the trend of the two outcomes for participants who are less likely to have sedentary activity but prefer to take more daily steps, more moderate to vigorous physical activity (MVPA) time, and higher intensity in activities at baseline. Although one may simply select the individuals who meet all criteria for the inference, the results would have a severe loss of information and there could be very few or even zero subjects satisfying all criteria in our small sample. On the other hand, as we will show in Section 5, the association structure obtained from our joint modeling uses the information from all participants and is flexible to handle user-specified conditions for different research purposes.

In this manuscript, we propose a flexible statistical model that handles a large number of response variables mixed with multiple types in longitudinal studies. A naive solution is to ignore the correlation among multiple outcomes and fit them separately as if they are independent, but this potentially losses important information from the data. The other simple approach is to consider the association only among the same type of variables (e.g. bivariate normal), while ignore the association across different types of responses. As we will show, our estimation for outcomes under specified conditions needs to understand the association patterns among all variables, while ignoring such correlation would lead to biased conclusions. For joint modeling approach, most of the current multivariate methods focus on one or two types of outcomes [3]. For example, Fieuws and Verbeke [4] and Fieuws et al. [5] respectively discuss the methods to handle multiple normal and binary outcomes, Gueorguieva and Agresti [6] develop a modeling strategy for one normal and one binary responses, and Buu et al. [7] propose a joint model for one count and one binary variables to handle zero-inflated count data. These models are limited in practice. Instead, we develop a multivariate mixed effect model for various types of outcomes, which incorporates the within individual correlation across variables by random effects.

The difficulty for this study is finding a method that can fit the proposed model. With the increasing number of variables of different types, the number of random effects also

increases. Current maximum likelihood methods for joint modeling are only feasible for low dimensional random effects. For multivariate linear mixed models, where the analytical formulation of marginal density is known, the model fitting is infeasible when the number of variables exceed four as the parameters in the covariance matrix for random effects are rapidly increasing [4]. For models of non-linear variables, large dimensional integration problems are more difficult to handle in the presence of multiple random effects. For example, the maximum likelihood approach using Monte Carlo or Gauss-Hermite quadrature is computationally extremely difficult for the integration of seven random effects in Fieuws et al. [5]. For our accelerometer data, we respectively set a random intercept and a random slope for eight variables. Therefore, our model involves  $2 \times 8 = 16$  dimensional random effects, and it exceeds the limit for maximum likelihood.

An alternative model fitting approach is to use a pairwise likelihood strategy by splitting joint estimation into a series of bivariate joint mixed models [4, 5]. However, it could still be intractable as the number of pairs quadratically increases with the number of variables. Moreover, as our proposed model involves four types of variables, the paired structures involve different formulations (i.e. bivariate normal, normal and proportion, binary and count, etc), which is difficult to implement in practice.

We use an upgraded penalized quasilikelihood approach to fit the data [8, 9]. This method extends the previous approach [10] by accommodating the data with continuous, proportion, count, and binary variables. Based on the penalized quasilikelihood method, all of the non-linear variables are transformed and approximately postulated by linear mixed models. Therefore, four types of measurements can be postulated by an approximated multivariate linear mixed model. In addition, the idea of an ECME algorithm [11] is extended to our joint model to solve the estimation problem for a large dimension of random effects. Our method avoids the numerical integrations for non-linear variables, and it only has modest computational workload. In addition, the updating of parameters in maximum likelihood method requires tedious computation of the first and second order derivatives with various formulations from different types of outcomes [12, 13], while our algorithm is easy to implement since it only involves the first derivatives of log-likelihood functions for linear mixed model.

The paper is organized as follows. Section 2 describes the models for multiple types of variables, and Section 3 describes our developed algorithm to fit the model. Section 4 gives results from our simulation studies. Section 5 analyzes the motivating physical activity data obtained from wearable accelerometer-based device. Concluding remarks are in Section 6.

#### 2. Mixed Effects Model

#### 2.1. Model Specification

Let  $Y_{ij}^{(\ell)}$  be the  $\ell^{th}$  outcome at occasion  $j = 1, ..., J_i$  for subject i = 1, ..., n. To accommodate the data, we set  $\ell = 1, 2$  for continuous responses,  $\ell = 3, 4$  for proportional responses,  $\ell = 5, 6$ for count responses and  $\ell = 7, 8$  for binary responses, respectively. We shall assume that  $Y_{ij}^{(\ell)}$ comes from a specified distribution with mean  $\mu_{ij}^{(\ell)}$ , linear predictor  $\eta_{ij}^{(\ell)}$ , and link functions to

connect  $\mu_{ij}^{(\ell)}$  and  $\eta_{ij}^{(\ell)}$  in various forms for different types of outcomes. Then we define linear predictor  $\eta_{ij}^{(\ell)} = X_{ij}^{(\ell)} \beta^{(\ell)} + Z_{ij}^{(\ell)} u_i^{(\ell)}$ , where  $X_{ij}^{(\ell)}$  and  $Z_{ij}^{(\ell)}$  are covariate vectors for fixed and random effects, respectively,  $\beta^{(\ell)}$  is a vector of fixed effect coefficients,  $u_i^{(\ell)}$  is a vector of correlated random effects with Normal( $\mathbf{0}, \Psi^{(\ell)}$ ).

For continuous data (l = 1, 2), we assume that the response follows a normal distribution and let  $\eta_{ij}^{(l)} = \mu_{ij}^{(l)}$  to have

$$Y_{ij}^{(\ell)} = \mu_{ij}^{(\ell)} + \varepsilon_{ij}^{(\ell)} = X_{ij}^{(\ell)} \beta^{(\ell)} + Z_{ij}^{(\ell)} u_i^{(\ell)} + \varepsilon_{ij}^{(\ell)}, \ \ell = 1, 2,$$
(1)

where  $\varepsilon_{ii}^{(\ell)}$  is independent random noise with Normal(0,  $\sigma^{2(\ell)}$ ).

For proportional data, based on the Beta regression framework proposed by Ferrari and Cribari-Neto [14], the density function for proportional outcomes  $Y_{ij}^{(\ell)}$  given random effects  $u_i^{(\ell)}$  is assumed to have

$$\frac{\Gamma\{\phi^{(\ell)}\}}{\Gamma\{\mu_{ij}^{(\ell)}\phi^{(\ell)}\}\Gamma[\{1-\mu_{ij}^{(\ell)}\}\phi^{(\ell)}]}\{y_{ij}^{(\ell)}\}^{\mu_{ij}^{(\ell)}\phi^{(\ell)}-1}\{1-y_{ij}^{(\ell)}\}^{\{1-\mu_{ij}^{(\ell)}\}\phi^{(\ell)}-1}, \ \ell=3,4,$$

where  $\Gamma(\cdot)$  is the gamma function, and we reparameterize  $\sigma^{2(b)} = 1/\{1 + \phi^{(b)}\}$  (l = 3, 4) to unify the notations with continuous outcomes. We also have

 $\operatorname{var}\{Y_{ij}^{(\ell)}|u_{i}^{(\ell)}\} = \sigma^{2(\ell)}\mu_{ij}^{(\ell)}\{1-\mu_{ij}^{(\ell)}\} \text{ and use logit link function to connect } \mu_{ij}^{(\ell)} \text{ and } \eta_{ij}^{(\ell)} \text{ as} \\ \operatorname{logit}\{\mu_{ij}^{(\ell)}\} = \eta_{ij}^{(\ell)} = X_{ij}^{(\ell)}\beta^{(\ell)} + Z_{ij}^{(\ell)}u_{i}^{(\ell)}, \ \ell=3,4.$  (2)

Poisson distributions with log link function and binomial distributions with logit link function are employed to model the count and binary data, respectively. They lead to

$$\log\{\mu_{ij}^{(\ell)}\} = \eta_{ij}^{(\ell)} = X_{ij}^{(\ell)} \beta^{(\ell)} + Z_{ij}^{(\ell)} u_i^{(\ell)}, \ \ell = 5, 6,$$
(3)

$$logit\{\mu_{ij}^{(\ell)}\} = \eta_{ij}^{(\ell)} = X_{ij}^{(\ell)} \beta^{(\ell)} + Z_{ij}^{(\ell)} u_i^{(\ell)}, \ \ell = 7, 8.$$
(4)

We further assume that given  $u_i^{(\ell)}(\ell=1,\ldots,8)$ , the observations are independent across all occasions, subjects, and different types of responses. Therefore, the random effects model the association pattern across visit time points and also postulate the correlation structure across all variables.

For ease of exposition, we combine notations as  $Y_i^{(\ell)} = \{Y_{i1}^{(\ell)}, \dots, Y_{iJ_i}^{(\ell)}\}^{\mathrm{T}}$  and  $Y_i = \{Y_i^{(1)^{\mathrm{T}}}, \dots, Y_i^{(8)^{\mathrm{T}}}\}^{\mathrm{T}}$  for responses,  $X_i^{(\ell)} = \{X_{i1}^{(\ell)^{\mathrm{T}}}, \dots, X_{iJ_i}^{(\ell)^{\mathrm{T}}}\}^{\mathrm{T}}$  and  $X_i$  to be a block diagonal matrix with elements  $X_i^{(\ell)}(\ell=1,\dots,8), Z_i^{(\ell)} = \{Z_{i1}^{(\ell)^{\mathrm{T}}},\dots, Z_{iJ_i}^{(\ell)^{\mathrm{T}}}\}^{\mathrm{T}}$  and  $Z_i$  to be a block diagonal matrix with elements  $Z_i^{(\ell)}(\ell=1,\dots,8)$ .

We also denote  $\beta = {\beta^{(1)}}^T, ..., \beta^{(8)}^T {}^T$  for fixed effect coefficients, and

 $u_i = \{u_i^{(1)^{\mathrm{T}}}, \dots, u_i^{(8)^{\mathrm{T}}}\}^{\mathrm{T}}$  for random effect variables. Thus, the random effect variables have  $u_i = \operatorname{Normal}(\mathbf{0}, \Psi)$  with block matrix  $\Psi$  including diagonal elements  $\Psi^{(\mathbf{0})}$  and off-diagonal element  $\Psi^{(\mathbf{0})}(\mathbf{0}, \mathbf{1}, 2, \dots, 8; \mathbf{0}, \mathbf{0})$ . The element  $\Psi^{(\mathbf{0})}$  denotes the covariance of  $u_i^{(\ell)}$ . The

element  $\Psi^{(\bar{l})}$  determines the covariance of  $u_i^{(\ell)}$  and  $u_i^{(\bar{\ell})}$ , which measures the association level across two variables. We further denote  $\sigma^2 = \{\sigma^{2(1)}, ..., \sigma^{2(4)}\}^{\mathrm{T}}$ .

Therefore, the multivariate longitudinal data model for various types of variables involves three sets of parameters to be estimated: (1) fixed effect coefficients  $\beta$ ; (2)  $u_i$ 's covariance matrix  $\Psi$ ; (3) dispersion parameters  $\sigma^2$ .

#### 2.2. Approximated Linear Mixed Model

We approximate the proportion, count and binary variables using the penalized quasilikelihood method (PQL) proposed by Breslow and Clayton [8]. The approximation is upgraded by Goldstein and Rasbash [9] with second order approximation terms to improve performance. As we will show in simulation study, the upgraded approach employed in this manuscript leads to smaller bias comparing to the regular PQL method without using second order approximation. We briefly describe the formula for proportional variables here, and the formulas for count and binary data are displayed in the Appendix A.1. Let  $H(\cdot)$  be our inverse function of logit link with  $H(\cdot) = \exp(\cdot)/\{1 + \exp(\cdot)\}$ , and  $H'(\cdot)$  and  $H''(\cdot)$  be its first and second derivatives. Given specified values of  $(\hat{\beta}, \hat{u}_i)$ , the proportional data model in (2)

has  $\hat{\eta}_{ij}^{(\ell)} = X_{ij}^{(\ell)} \hat{\beta}^{(\ell)} + Z_{ij}^{(\ell)} \hat{u}_i^{(\ell)} (\ell=3, 4)$ . Following the approximation methods discussed in Lindstrom and Bates [15], Wolfinger and O'Connell [16], Goldstein and Rasbash [9] and Molenberghs and Verbeke [17], the formulation is displayed as follows with detailed derivations described in the Appendix A.2,

$$Y_{ij}^{(\ell)*} \equiv [1/H'\{\hat{\eta}_{ij}^{(\ell)}\}] [Y_{ij}^{(\ell)} - H\{\hat{\eta}_{ij}^{(\ell)}\}] + \hat{\eta}_{ij}^{(\ell)} - [1/2H'\{\hat{\eta}_{ij}^{(\ell)}\}] H''\{\hat{\eta}_{ij}^{(\ell)}\} [Z_{ij}^{(\ell)} \widehat{\operatorname{var}}\{u_i^{(\ell)} - \hat{u}_i^{(\ell)}\} Z_{ij}^{(\ell)^{\mathrm{T}}}] \approx X_{ij}^{(\ell)} \beta^{(\ell)} + Z_{ij}^{(\ell)} u_i^{(\ell)} + \varepsilon_{ij}^{(\ell)}, \ \ell = 3, 4,$$
(5)

where  $\varepsilon_{ij}^{(\ell)}$  has mean zero and variance  $\sigma^{2(\ell)}/H'\{\hat{\eta}_{ij}^{(\ell)}\}$ .

After performing similar transformations for the binary and count variables, we now have

six transformed outcomes  $Y_{ij}^{(\ell)*}(\ell=3,\ldots,8)$  that have approximated mean structure  $E\{Y_{ij}^{(\ell)*}|u_i^{(\ell)}\}=X_{ij}^{(\ell)}\beta^{(\ell)}+Z_{ij}^{(\ell)}u_i^{(\ell)}$  and variance structure  $var\{Y_{ij}^{(\ell)*}|u_i^{(\ell)}\}=var\{\varepsilon_{ij}^{(\ell)}\}$ . The estimation of model parameters can thus be implemented as fitting a linear mixed model. As these variables can be respectively fitted as approximated linear mixed model and two continuous responses are also postulated by linear mixed models, a multivariate linear mixed model can be used to jointly fit all variables. Therefore, despite the various types of outcomes, they are transformed into a unified framework which can be conveniently estimated by the algorithm described in Section 3.

### 3. Model Estimation

We estimate the parameters by extending the idea of an ECME algorithm [11]. The ECME algorithm updates the fixed effects parameters ( $\beta$ ,  $\sigma^2$ ) by the Newton-Raphson method and updates the random effects parameters  $\Psi$  by the EM algorithm. Liu and Rubin [18] study the convergence properties of this algorithm. We provide detailed procedures for the model fitting here. Similar to the notation of  $Y_i$ , we define the approximated response vector as

 $Y_i^{(\ell)*} = \{Y_{i1}^{(\ell)*}, \dots, Y_{iJ_i}^{(\ell)*}\}^{\mathrm{T}} (\ell=3, 4, \dots, 8), \text{ and combine these vectors with continuous}$ outcomes as  $Y_i^* = \{Y_i^{(1)^{\mathrm{T}}}, Y_i^{(2)^{\mathrm{T}}}, Y_i^{(3)*^{\mathrm{T}}}, \dots, Y_i^{(8)*^{\mathrm{T}}}\}^{\mathrm{T}}.$  Let  $\Sigma_i$  be the diagonal matrix with the variance for  $\varepsilon_{1i}^{(\ell)}$  as

$$\sum_{i} \doteq \text{diag}[\sigma^{2(1)}\mathbf{I}_{J_{i}}, \sigma^{2(2)}\mathbf{I}_{J_{i}}, \sigma^{2(3)}\mathbf{G}_{i}^{(3)}, \sigma^{2(4)}\mathbf{G}_{i}^{(4)}, \mathbf{F}_{i}^{(5)}, \mathbf{F}_{i}^{(6)}, \mathbf{G}_{i}^{(7)}, \mathbf{G}_{i}^{(8)}],$$
(6)

where  $\mathbf{I}_{J_i}$  is  $J_i \times J_i$  identity matrix,  $\mathbf{F}_i^{(\ell)} = diag[\exp\{-\hat{\eta}_{i1}^{(\ell)}\}, \dots, \exp\{-\hat{\eta}_{iJ_i}^{(\ell)}\}]$  and  $\mathbf{G}_i^{(\ell)} = diag[1/H'\{\hat{\eta}_{i1}^{(\ell)}\}, \dots, 1/H'\{\hat{\eta}_{iJ_i}^{(\ell)}\}]$ . We also let  $V_i$  to be the approximated covariance for  $Y_i^*$  as

$$V_i \doteq \sum_i + Z_i \Psi Z_i^{\mathrm{T}}.$$
 (7)

The approximated conditional covariance for  $u_i$  given  $Y_i^*$  by assuming known  $\beta$  as

$$U_i \doteq \Psi - \Psi Z_i^{\mathrm{T}} V_i^{-1} Z_i \Psi, \quad (8)$$

and for unknown  $\beta$  as

$$\widehat{\operatorname{var}}(u_i - \hat{u}_i) = U_i + \Psi Z_i^{\mathrm{T}} V_i^{-1} X_i (\sum_{i=1}^n X_i^{\mathrm{T}} V_i^{-1} X_i)^{-1} X_i^{\mathrm{T}} V_i^{-1} Z_i \Psi, \quad (9)$$

where the estimate of  $\widehat{var}(u_i^{(\ell)} - \hat{u}_i^{(\ell)})$  for (5), (A.1) and (A.2) can be obtained by extracting the corresponding elements.

The algorithm has

- 1. Based on the current working values  $\{\beta_{curr}, \Psi_{curr}, \sigma_{curr}^2, \hat{u}_{i,curr}^{(\ell)}(\ell=3,\ldots,8)\},\$ calculate  $\hat{\eta}_{i,curr}^{(\ell)}(\ell=3,\ldots,8)$ , update  $\sum_{i,curr}, V_{i,curr}, U_{i,curr}$  and  $\widehat{var}_{curr}(u_i - \hat{u}_i)$ sequentially by Equations (6) to (9), and then update  $Y_{i,curr}^*$  by Equations (5), (A. 1) and (A.2).
- 2. update  $\beta$  as

$$\beta_{\text{new}} = \left(\sum_{i=1}^{n} X_{i}^{\text{T}} V_{i,\text{curr}}^{-1} X_{i}\right)^{-1} \left(\sum_{i=1}^{n} X_{i}^{\text{T}} V_{i,\text{curr}}^{-1} Y_{i,\text{curr}}^{*}\right), \quad (10)$$

- 3. Based on  $\{\beta_{\text{new}}, \Psi_{\text{curr}}, \sigma_{\text{curr}}^2, \hat{u}_{i,\text{curr}}^{(\ell)} (\ell=3,\ldots,8)\}$ , repeat step 1 to calculate  $\hat{\eta}_{i,\text{new}}^{(\ell)} (\ell=3,\ldots,8), \Sigma_{i,\text{new}}, V_{i,\text{new}}, \widehat{\text{var}}_{\text{new}} (u_i \hat{u}_i) \text{ and } Y_{i,\text{new}}^*$ , sequentially.
- 4. update  $\sigma^{2(b)}$  as

$$\begin{split} &\sigma_{\text{new}}^{2(\ell)} = N^{-1} \sum_{i=1}^{n} \{Y_{i,\text{new}}^{(\ell)*} - X_{i}\beta_{\text{new}}^{(\ell)}\}^{^{\text{T}}} \{W_{i}^{(\ell)}\}^{^{-1}} \{Y_{i,\text{new}}^{(\ell)*} - X_{i}^{(\ell)}\beta_{\text{new}}^{(\ell)}\}, \text{ where } \\ &N = \sum_{i=1}^{n} J_{i}, W_{i}^{(\ell)} = \mathbf{I}_{J_{i}} + Z_{i}^{(\ell)} \Psi_{\text{curr}}^{(\ell)} Z_{i}^{(\ell)^{^{\text{T}}}} / \sigma_{\text{curr}}^{2(\ell)} (\ell = 1, 2), \text{ and } \\ &W_{i}^{(\ell)} = \text{diag}[1/H'\{\hat{\eta}_{i,\text{new}}^{(\ell)}\}] + Z_{i}^{(\ell)} \Psi_{\text{curr}}^{(\ell)} Z_{i}^{(\ell)^{^{\text{T}}}} / \sigma_{\text{curr}}^{2(\ell)} (\ell = 3, 4). \end{split}$$

5. update 
$$\hat{u}_i$$
 as  $\hat{u}_{i,\text{new}} = \Psi_{\text{curr}} Z_i^{\text{T}} V_{i,\text{new}}^{-1} (Y_{i,\text{new}}^* - X_i \beta_{\text{new}})$ .

6. update  $\Psi$  as  $\Psi_{\text{new}} = n^{-1} \sum_{i=1}^{n} (\hat{u}_{i,\text{new}} \hat{u}_{i,\text{new}}^{\text{T}} + U_{i,\text{new}}).$ 

We keep  $\hat{u}_{i,\text{new}}^{(\ell)}(\ell=3,\ldots,8)$  for the next iteration by selecting subvectors from  $\hat{u}_{i,\text{new}}$ . The procedure is iterated until convergence of  $\{\beta, \Psi, \sigma^2, \hat{u}_i^{(\ell)}(\ell=3,\ldots,8)\}$ .

According to equation (10), the covariance matrix for  $\hat{\beta}$  can be calculated by

 $\left(\sum_{i=1}^{n} X_i^{\mathrm{T}} V_i^{-1} X_i\right)^{-1}$ . This variance estimator leads to satisfactory results in our simulation studies and data analysis. However, the estimator is based on the approximation for  $Y_i^*$  and assuming other parameters are known. The improved estimators discussed in [8] can be used to reduce bias.

#### Remark 1

We fit the data by a generalized linear model without random effects to obtain an initial

value of  $\beta$ . We also set all elements in  $\sigma^2$  to be 0.01 and all elements in  $\hat{u}_i^{(\ell)}(\ell=3,\ldots,8)$  to be 0 as starting values. Following the discussion from [8], we set the diagonal elements in the covariance matrix  $\Psi$  to be small positive values and off-diagonal elements to be 0. In a preliminary simulation study (not reported), the diagonal variances in  $\Psi$  are all set to be 0.1.

However, during estimation iteration, this initial value could lead to fairly extreme values in  $\Sigma_i$  for count outcomes, and thus  $V_i$  could be singular and may not be invertible. In the preliminary study, 36% of the simulation runs have this problem for 1 to 3 subjects. We solve the issue by resetting the variances in  $\Psi$  corresponding to the count outcomes as 0.001. Based on the new initial value settings, none of the runs in moderate sample scenario has such issue in the study reported in Section 4. There are less than 1% of the runs in large sample scenario have one subject with singular  $V_i$ . For those rare runs with singular  $V_i$  under new starting values, we remove the corresponding subjects from iteration procedures, and then all of the runs successfully achieve convergence.

#### 4. Simulation Studies

In this section, we use a simulation study to demonstrate the performance of our proposed approach (labeled as JOINT-PQL2). As a comparison, two naive approaches are also explored, where the first method fits eight variables by assuming them to be independent (labeled as NAIVE1), and the other approach only models the association among the same type of variables (labeled as NAIVE2). We also study the performance of the regular PQL method which models the complete association structure but not using second order approximations for our proportion, count and binary outcomes (labeled as JOINT-PQL1).

In the simulation study of 500 runs, we have n = 200 subjects and each subject has  $J_i = 5$  visits. We also consider a larger sample size with n = 400 and  $J_i = 9$ . At visit time *j*, subject *i* has eight observations  $\{Y_{ij}^{(1)}, \ldots, Y_{ij}^{(8)}\}$  as described in Section 2. The continuous variables  $Y_{ij}^{(1)}$  and  $Y_{ij}^{(2)}$  are generated according to

where  $t_{ij} = j - 3$  (j = 1, ..., 5) for the moderate sample,  $t_{ij} = (j - 5)/2$  (j = 1, ..., 9) for the large sample, and  $x_{ij}^{(\ell)}$  are generated from standard normal distribution and independent across *i*, *j*, *l* Other types of variables have similar linear predictor formulation with

$$\eta_{ij}^{(\ell)} = \beta_0^{(\ell)} + \beta_1^{(\ell)} t_{ij} + \beta_2^{(\ell)} x_{ij}^{(\ell)} + \beta_3^{(\ell)} t_{ij} x_{ij}^{(\ell)} + u_{i,0}^{(\ell)} + t_{ij} u_{i,1}^{(\ell)}, \ \ell = 3, 4, \dots, 8$$

We set  $\beta_0^{(\ell)} = 0.5$ ,  $\beta_1^{(\ell)} = 0.2$ ,  $\beta_2^{(\ell)} = 0.2$ ,  $\beta_3^{(\ell)} = -0.1$  for  $\ell = 1, 3, 5, 7$ , and  $\beta_0^{(\ell)} = -0.5$ ,  $\beta_1^{(\ell)} = -0.2$ ,  $\beta_2^{(\ell)} = -0.2$ ,  $\beta_3^{(\ell)} = 0.1$  for  $\ell = 2, 4, 6, 8$ , and the covariance matrix  $\Psi$  for  $u_i$  has  $\Psi^{(\ell)}(\ell = 1, 2, ..., 8)$  in  $2 \times 2$  matrix with diagonal elements to be  $\Psi_{[1,1]}^{(\ell\ell)} = \Psi_{[2,2]}^{(\ell\ell)} = 0.5$  and off-diagonal elements to be  $\Psi_{[1,2]}^{(\ell\ell)} = 0.25$ , and  $\Psi^{(\ell)}$  in  $2 \times 2$  matrix with all elements to be 0.1. We also set  $\sigma^{2(1)} = \sigma^{2(2)} = 1$  and  $\sigma^{2(3)} = \sigma^{2(4)} = 1/30$ .

As the discussion in Section 1, we are interested in the estimation of model parameters as well as the longitudinal pattern of outcomes under certain conditions. The conditional patterns can be used to study the physical activity outcomes for participants who meet

particular criteria, and they will be discussed in Section 5. As an example, we use three approaches to fit simulated data, and the estimates are used to calculate the following conditional expectations by averaging the results from one million times of Monte Carlo samplings:

$$E\{Y_{\rm ij}^{(\ell)}|x_{\rm ij}^{(\ell)}\!=\!0,Y_{i1}^{(3)}\geq 0.5,Y_{i1}^{(4)}\geq 0.5,Y_{i1}^{(5)}\geq 15,Y_{i1}^{(6)}\geq 5,Y_{i1}^{(7)}\!=\!1,Y_{i1}^{(8)}\!=\!1\},$$

where l = 1, 2, j = 1, ..., 5 for the moderate sample, and j = 1, ..., 9 for the large sample.

Tables 1–2 present the relative bias percent (Bias%), the average model-based standard error (ASE), the empirical standard error (ESE) and 95% coverage rate (CR) for  $\beta$  estimates in count and binary outcomes, respectively. The  $\beta$  estimates for continuous and proportional outcomes are listed in Online Supplementary Material. The results show that our proposed JOINT-PQL2 method yield small biases and the coverage rates are appropriate for  $\beta$ . ASE and ESE agree reasonably well, suggesting the variance estimates for  $\beta$  is appropriate. The JOINT-PQL1 method, however, has larger bias and poor coverage rate for some parameters in  $\beta$ . For the estimation of  $\beta$ , the NAIVE1 and NAIVE2 approaches lead to acceptable conclusions as the proposed JOINT-PQL2 method. We also report the estimates for  $\sigma^2$  and diagonal elements in  $\Psi$  using our JOINT-PQL2 approach in Online Supplementary Material, which indicate fairly good performance in the estimation of the model parameters.

Figure 2 shows the true and the averaged estimates of the conditional expectations for l = 1, 2 across *j*. Our JOINT-PQL2 approach generally captures the true patterns, while the NAIVE1 and NAIVE2 methods lead to obvious bias to describe the trends of conditional expectations. We do not present the results from the JOINT-PQL1 method in Figure 2 because it leads to similar patterns as the JOINT-PQL2 method.

Therefore, the simulation results suggest that the JOINT-PQL1 method without using second order approximation, may have biased parameter estimation in  $\beta$ . The NAIVE1 and NAIVE2 approaches, which either completely or partially ignore the correlation among responses, may provide misleading conclusions in the estimation of conditional expectations with respect to multiple outcomes. On the other hand, our proposed JOINT-PQL2 method has satisfactory estimation results for both model parameters and conditional expectations.

#### 5. Application to Physical Activity Data

In this section the proposed method is applied to the physical activity dataset collected by wearable devices. The project is to investigate the metabolic effects of exercise interventions to increase activity and reduce sedentary time in a group of office workers [1]. The raw activity data is obtained by a device named the ActivPAL<sup>TM</sup> (www.paltech.plus.com), which is taped in front of the thigh and uses a vertical-axis accelerometer to measure the angle of the thigh and the frequency of body movement. For this project, Zhang et al. [19] develop an R package "PAactivPAL" to handle the ActivPAL device's raw records, which summarizes the dense time activity information into daily averages. We focus on the following eight variables: daily sedentary hours, daily energy expenditure (measured by METs hours), the proportion of time for sedentary bout greater than 20 minutes, the proportion of time for

active bout greater than 5 minutes, daily number of standing up activities, daily number of steps, whether daily MVPA time is greater than one hour, and whether the highest METs in 10 minutes is greater than 3.

According to the data description in Section 1, the dataset has n = 63 subjects and

 $Y_{ij}^{(\ell)}(\ell=1,\ldots,8)$  corresponds to eight selected responses. All individuals are scheduled to have one day's measurement across five weeks with  $t_{ij} = -2, -1, 0, 1, 2$ , respectively. Incomplete daily observations are removed from our analysis. Moreover, the individual *i* enrolled in exercise group has  $X_{ij}^{(\ell)}=1$  for all *j* and  $\ell$  while others in control group have  $X_{ij}^{(\ell)}=0$ .

The model used in simulation Section 4 are introduced to fit the data. This model includes intercept, group, time and a group-time interaction for fixed effects, and intercept and time for random effects. Our proposed JOINT-PQL2 approach is applied to estimate parameters. Table 3 presents  $\beta$  estimates for eight variables. The table illustrates that four outcomes, daily energy expenditure levels, the daily number of steps, the probability of daily MVPA time is greater than one hour, and the probability of highest METs in 10 minutes greater than 3, have significantly higher levels in exercise treatment group than control group. It implies that the treatment group will gain health benefits associated with increased physical activity.

Based on the model of the association structure among eight activity factors, it is of interest to study the pattern of daily sedentary hour and energy expenditure among particular subgroups across five weeks. We select two subsets of subjects with active and inactive behaviors at Week 0, respectively. The criteria for active performance have: (1) the proportion of sedentary bout greater than 20 minutes is less than 20%; (2) the active bouts greater than 5 minutes are more than 30%; (3) daily standing up behaviors are more than 40 times; (4) daily steps are more than 6000; (5) daily MVPA time is greater than one hour; (6) the highest METs in 10 minutes is greater than 3. The individual who reaches all six criteria at Week 0 is defined as active participant, while those who do not meet any term is taken into inactive group. Therefore, the conditional expectations to be estimated here are

$$E\{Y_{\rm ij}^{(\ell)}|Y_{i1}^{(3)}{<}0.2,Y_{i1}^{(4)} \ge 0.3,Y_{i1}^{(5)} \ge 40,Y_{i1}^{(6)} \ge 6000,Y_{i1}^{(7)}{=}1,Y_{i1}^{(8)}{=}1\}$$

and

$$E\{Y_{\rm ij}^{(\ell)}|Y_{i1}^{(3)} \ge 0.2, Y_{i1}^{(4)} < \! 0.3, Y_{i1}^{(5)} < \! 40, Y_{i1}^{(6)} < \! 6000, Y_{i1}^{(7)} = \! 0, Y_{i1}^{(8)} = \! 0\},$$

where l = 1, 2 and j = 1, ..., 5. We study the situations with  $X_{ij}^{(\ell)} = 0$  and  $X_{ij}^{(\ell)} = 1$ , respectively.

Figure 3(a)(b) shows the estimated conditional expectations for active/inactive participants with/without exercise training across five weeks. For daily sedentary hours, the estimates suggest that active participants have lower sedentary time than inactive subjects across

weeks. In addition, the sedentary time in both exercise and control groups is decreasing with increasing time, and the exercise group has faster decreasing rate. A reasonable explanation is that many subjects in both groups receive several lifestyle suggestions to increase spontaneous activities (e.g. walk the stairs), which reduces the sedentary time. The supervised structured exercise training for the exercise group led to further reductions in sedentary behaviors. For energy expenditure levels, in both exercise and control groups, the active participants have higher outcome than the inactive ones at Week 0 but the difference is fairly small at Week 12. Active subjects at baseline in both the exercise and control groups decreased energy expenditure by Week 12, while the less active subjects increased energy expenditure. In this study, all participants in the exercise groups completed the same amount of exercise each week (~ 200 min) regardless of their activity status at Week 0. Although this is standard practice in such trials to ensure all participants complete the same dose, these data suggest that more active participants at baseline decreased their energy expenditure as a result of the standard intervention and future studies could consider tailoring exercise recommendations based on activity status at Week 0 to promote increases in energy expenditure for all participants.

#### 6. Discussion

We have proposed a joint modeling and estimation strategy for longitudinal data with continuous, proportion, count and binary variables. To avoid the computational difficulty resulting from the large dimension of random effects, our algorithm uses an improved quasilikelihood approximation to handle non-linear outcomes, while employs an efficient estimation method to fit multivariate linear mixed model data. The simulation results are promising and suggest that the proposed method has little bias and outperforms naive approaches which ignore the association among multiple responses. The data analysis on the physical activity data collected by using wearable accelerometer device proves the utility of our method in applications.

The penalized quasilikelihood (PQL) method is often criticized for its biased estimates [20]. However, Goldstein and Rasbash [9] suggest that the second order approximation greatly improves the PQL approach. Vonesh et al. [21] study the asymptotic results for the PQL method and prove that it provides a consistent estimator if the number of subjects and the number of measurements per subject go to infinity. The simulation studies in this manuscript agree with these conclusions. We noted that our proposed approach has a small bias in the estimation for both  $\beta$  and conditional expectations with moderate sample sizes, and the bias become negligible with larger sample sizes.

To compare our method with commonly used model fitting approaches, we further investigate the performance of two maximum likelihood approaches, which are based on Laplace approximation (implemented by the R package "Ime4" [22, 23]) and Gauss-Hermite quadrature approximation (11 quadrature points are used), respectively. Both methods are

employed only to fit *one* binary outcome  $Y_i^{(8)}$ . The estimates for the parameter  $\beta^{(8)}$  are briefly reported in Online Supplementary Material and they are comparable to the estimates produced by our newly proposed method. However, in our multivariate data structure with eight responses, these maximum likelihood approaches are computationally intractable.

Finally, we have discussed the approach to select initial values in Section 3. Based on the suggestion from [8], we use small positive values for diagonal elements in  $\Psi$  (0.001 for count outcomes and 0.1 for other types of responses). This method works well in our simulation and application studies. Moreover, it would be interesting to consider selecting the initial  $\Psi$  from original data. This could be implemented by using either a maximum likelihood or restricted maximum likelihood estimation procedure, for example, see [8] and [16].

#### Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

#### Acknowledgments

Li was supported by discovery grants program 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 Sarah Kozey Keadle for making the physical activity data available to them.

#### References

- Kozey-Keadle S, Libertine A, Lyden K, Staudenmayer J, Freedson PS. Changes in sedentary time and spontaneous physical activity in response to an exercise training and/or lifestyle intervention. Journal of Physical Activity and Health. 2014; 11:1324–1333. [PubMed: 24184493]
- Keadle SK, Sampson J, Li H, Lyden K, Matthews CE, Carroll RJ. An evaluation of accelerometerderived metrics to assess daily behavioral patterns. Medicine & Science in Sports & Exercise. 2017; 49:54–63. [PubMed: 27992396]
- 3. Verbeke G, Fieuws S, Molenberghs G, Davidian M. The analysis of multivariate longitudinal data: A review. Statistical methods in medical research. 2014; 23:42–59. [PubMed: 22523185]
- Fieuws S, Verbeke G. Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. Biometrics. 2006; 62:424–431. [PubMed: 16918906]
- Fieuws S, Verbeke G, Boen F, Delecluse C. High dimensional multivariate mixed models for binary questionnaire data. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2006; 55:449–460.
- 6. Gueorguieva RV, Agresti A. A correlated probit model for joint modeling of clustered binary and continuous responses. Journal of the American Statistical Association. 2001; 96:1102–1112.
- 7. Buu A, Li R, Tan X, Zucker RA. Statistical models for longitudinal zero-inflated count data with applications to the substance abuse field. Statistics in Medicine. 2012; 31:4074–4086. [PubMed: 22826194]
- Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association. 1993; 88:9–25.
- Goldstein H, Rasbash J. Improved approximations for multilevel models with binary responses. Journal of the Royal Statistical Society. Series A. 1996; 159:505–513.
- Li H, Staudenmayer J, Carroll RJ. Hierarchical functional data with mixed continuous and binary measurements. Biometrics. 2014; 70:802–811. [PubMed: 25134936]
- 11. Schafer, JL. Tech. rep. The Methodological Center, The Pennsylvania State University; 1998. Some improved procedures for linear mixed models.
- Lindstrom MJ, Bates DM. Newton-raphson and em algorithms for linear mixed-effects models for repeated-measures data. Journal of the American Statistical Association. 1988; 83:1014–1022.
- Broström G, Holmberg H. Generalized linear models with clustered data: Fixed and random effects models. Computational Statistics and Data Analysis. 2011; 55:3123–3134.
- Ferrari S, Cribari-Neto F. Beta regression for modelling rates and proportions. Journal of Applied Statistics. 2004; 31:799–815.

- Lindstrom MJ, Bates DM. Nonlinear mixed effects models for repeated measures data. Biometrics. 1990; 46:673–687. [PubMed: 2242409]
- Wolfinger R, O'Connell M. Generalized linear mixed models a pseudo-likelihood approach. Journal of Statistical Computation and Simulation. 1993; 48:233–243.
- 17. Molenberghs, G., Verbeke, G. Models for Discrete Longitudinal Data. Springer; 2005.
- Liu C, Rubin DB. The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. Biometrika. 1994; 81:633–648.
- Zhang Y, Li H, Kozey-Keadle S, Matthews CE, Carroll RJ. PAactivPAL: Summarize Daily Physical Activity from 'activPAL' Accelerometer Data. R package version 1.0. 2015
- Rodriguez G, Goldman N. Improved estimation procedures for multilevel models with binary response: A case-study. Journal of the Royal Statistical Society. Series A (Statistics in Society). 2001; 164:339–355.
- Vonesh EF, Wang H, Nie L, Majumdar D. Conditional second-order generalized estimating equations for generalized linear and nonlinear mixed-effects models. Journal of the American Statistical Association. 2002; 97:271–283.
- 22. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2017.
- 23. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. Journal of Statistical Software. 2015; 67:1–48.

#### Appendix

#### A.1. Second Order Approximation for Count and Binary Variables

For count variables, we have

#### (A.1)

where  $\varepsilon_{ii}^{(\ell)}$  has mean zero and variance  $\exp\{-\hat{\eta}_{ii}^{(\ell)}\}$ .

For binary variables, we have

$$Y_{ij}^{(\ell)*} \equiv [1/H'\{\hat{\eta}_{ij}^{(\ell)}\}][Y_{ij}^{(\ell)} - H\{\hat{\eta}_{ij}^{(\ell)}\}] + \hat{\eta}_{ij}^{(\ell)} - [1/2H'\{\hat{\eta}_{ij}^{(\ell)}\}]H''\{\hat{\eta}_{ij}^{(\ell)}\}[Z_{ij}^{(\ell)}\widehat{\operatorname{var}}\{u_i^{(\ell)} - \hat{u}_i^{(\ell)}\}Z_{ij}^{(\ell)^{\mathrm{T}}}] \approx X_{ij}^{(\ell)}\beta^{(\ell)} + Z_{ij}^{(\ell)}u_i^{(\ell)} + \varepsilon_{ij}^{(\ell)}, \ \ell = 7, 8,$$
(A.2)

where  $\varepsilon_{\rm ij}^{(\ell)}$  has mean zero and variance  $1/H'\{\hat{\eta}_{\rm ij}^{(\ell)}\}$ .

#### A.2. Second Order Approximation in Penalized Quasilikelihood

The derivation of penalized quasilikelihood (PQL) based approximations for proportion outcomes follows the techniques in Molenberghs and Verbeke [17]. For simplicity, we

remove ( $\hat{l}$ ) from notations. Given known values of ( $\hat{\beta}$ ,  $\hat{u}_i$ ) which lead to  $\hat{\eta}_{ij} = X_{ij}\hat{\beta} + Z_{ij}\hat{u}_i$ . A Taylor series expansion of  $H(\eta_{ij})$  around  $\hat{\eta}_{ij}$  for  $Y_{ij}$  has

$$Y_{ij} \approx H(\hat{\eta}_{ij}) + H'(\hat{\eta}_{ij})(\eta_{ij} - \hat{\eta}_{ij}) + \varepsilon_{ij}^*, \quad (A.3)$$

where  $\varepsilon_{ij}^*$  has mean zero and variance  $\sigma^2 H'(\hat{\eta}_{ij})$ . However, the approximation in (A.3) may lead to a severely biased estimator. To solve this problem, we could use a quadratic Taylor expansion with

$$Y_{ij} \approx H(\hat{\eta}_{ij}) + H'(\hat{\eta}_{ij})(\eta_{ij} - \hat{\eta}_{ij}) + (1/2)H''(\hat{\eta}_{ij})(\eta_{ij} - \hat{\eta}_{ij})^2 + \varepsilon_{ij}^*.$$
 (A.4)

In practice the term  $(\eta_{ij} - \hat{\eta}_{ij})^2$  may be intractable. Goldstein and Rasbash [9] suggest to work on a simplified formulation of (A.4) with the second order expansion only for  $u_{ji}$  which has

$$Y_{ij} \approx H(\hat{\eta}_{ij}) + H'(\hat{\eta}_{ij})(\eta_{ij} - \hat{\eta}_{ij}) + (1/2)H''(\hat{\eta}_{ij})(Z_{ij}u_i - Z_{ij}\hat{u}_i)^2 + \varepsilon_{ij}^*, \quad (A.5)$$

and the term  $(Z_{ij}u_i - Z_{ij}\hat{u}_i)^2$  can be replaced by  $Z_{ij}\widehat{\text{var}}(u_i - \hat{u}_i)Z_{ij}^{\text{T}}$  in computation. We take a transformation for (A.5) to obtain

$$\{1/H'(\hat{\eta}_{ij})\}\{Y_{ij}-H(\hat{\eta}_{ij})\}+\hat{\eta}_{ij}-\{1/2H'(\hat{\eta}_{ij})\}H''(\hat{\eta}_{ij})\{Z_{ij}\widehat{\operatorname{var}}(u_i-\hat{u}_i)Z_{ij}^{\mathrm{T}}\} \approx \eta_{ij}+\varepsilon_i(t)=X_{ij}\beta+Z_{ij}u_i+\varepsilon_{ij}\alpha+Z_{ij}u_i+\varepsilon_{ij}\alpha+Z_{ij}u_i+\varepsilon_{ij}\alpha+Z_{ij}u_i+\varepsilon_{ij}\alpha+Z_{ij}u_i+\varepsilon_{ij}\alpha+Z_{ij}u_i+\varepsilon_{ij}\alpha+Z_{ij}u_i+\varepsilon_{ij}\alpha+Z_{ij}u_i+\varepsilon_{ij}\alpha+Z_{ij}u_i+\varepsilon_{ij}\alpha+Z_{ij}u_i+\varepsilon_{ij}\alpha+Z_{ij}u_i+\varepsilon_{ij}\alpha+Z_{ij}u_i+\varepsilon_{ij}\alpha+Z_{ij}u_i+\varepsilon_{ij}\alpha+Z_$$

#### (A.6)

where  $\boldsymbol{\varepsilon}_{ij}$  has mean zero and variance  $\sigma^2/H'\{\hat{\eta}_{ij}\}$ . Define the term on the left side of (A.6) as variable  $Y_{ij}^*$ 

$$Y_{ij}^* \equiv \{1/H'(\hat{\eta}_{ij})\}\{Y_{ij} - H(\hat{\eta}_{ij})\} + \hat{\eta}_{ij} - [1/2H'(\hat{\eta}_{ij})]H''(\hat{\eta}_{ij})\{Z_{ij}\widehat{\operatorname{var}}(u_i - \hat{u}_i)Z_{ij}^{\mathrm{T}}\},$$

and (A.6) implies

$$Y_{ij}^* \approx X_{ij}\beta + Z_{ij}u_i + \varepsilon_{ij}$$

Li et al.



#### Figure 1.

Sample data from two subjects on weeks 0, 3, 6, 9 and 12. Solid lines and "X" labels display the observations from individual with ID 4. Dashed lines and "O" labels represent the outcomes from individual with ID 5. (a)  $Y^{(1)}$ : continuous variable for daily sedentary hours; (b)  $Y^{(2)}$ : continuous variable for energy expenditure; (c)  $Y^{(3)}$ : proportion of sedentary time greater than 20 minutes, (d)  $Y^{(4)}$ : proportion of active time greater than 5 minutes; (e)  $Y^{(5)}$ : count number of daily standing up behaviors; (f)  $Y^{(6)}$ : count number of daily steps; (g)  $Y^{(7)}$ : binary variable for whether daily moderate to vigorous physical activity (MVPA) time is

greater than one hour; (h)  $Y^{(8)}$ : binary variable for whether the highest energy expenditure rate measured by metabolic equivalents (METs) in 10 minutes is greater than 3.

Li et al.



#### Figure 2.

Simulation results for the conditional expectations for l = 1, 2 defined in Section 4. (a)(b) moderate sample size scenario with n = 200 and  $J_i = 5$ , (c)(d) large sample size scenario with n = 400 and  $J_i = 9$ . Dotted lines denote the true conditional expectation values. Solid lines represent the averaged values of the estimates from our JOINT-PQL2 method. Shadowed areas display the 10% to 90% quantiles of the estimated values in 500 simulation runs. Thick and thin dashed lines represent the averaged estimates by the NAIVE1 and NAIVE2 methods, respectively.

Li et al.



#### Figure 3.

The estimates of conditional expectations across five weeks for daily sedentary hours  $(Y_{ij}^{(1)})$ and energy expenditure levels  $(Y_{ij}^{(2)})$  defined in Section 5. The individual who reaches the criteria in Section 5 at Week 0 is defined as active participant, while those who do not meet any term at Week 0 is defined as inactive participant. Thick and thin lines represent the

estimates of active and inactive participants, respectively. Solid and dotted lines display the exercise treatment group and the control group, respectively.

# Table 1

Simulation results for  $\beta^{(5)}$  and  $\beta^{(6)}$  in count outcomes. As defined in Section 4, "JOINT-PQL2" denotes our proposed method, "NAIVE1" and "NAIVE2" association pattern but not using second order approximation. Displayed are the relative bias percent (Bias%), the average model-based standard error represent the approaches completely or partially ignoring the association structure among responses, and "JOINT-PQL1" is the method involving full (ASE), the empirical standard error (ESE) and 95% coverage rate (CR) for the estimates. Relative bias greater than 10% is highlighted in bold.

			<i>u</i> = 200,	$J_i = 5$		'	<i>i</i> = 400,	$J_i = 9$	
		Bias%	ASE	ESE	CR	Bias%	ASE	ESE	CR
(E)	<b>NAIVE1</b>	-1.42	0.06	0.06	0.95	-0.90	0.04	0.04	0.93
$\beta_0^{(c)}$	NAIVE2	-1.62	0.06	0.06	0.96	-0.93	0.04	0.04	0.93
	JOINT-PQL1	18.66	0.05	0.05	0.59	11.53	0.04	0.04	0.67
	JOINT-PQL2	-1.22	0.06	0.06	0.95	-1.23	0.04	0.04	0.96
$_{\circ}(5)$	<b>NAIVE1</b>	-1.49	0.05	0.05	0.96	-0.92	0.04	0.04	0.96
$\beta_1^{(2)}$	NAIVE2	-2.67	0.05	0.05	0.94	-0.98	0.04	0.04	0.96
	JOINT-PQL1	1.26	0.05	0.05	0.94	-0.64	0.04	0.04	0.94
	JOINT-PQL2	-1.68	0.05	0.05	0.94	-1.43	0.04	0.04	0.95
<u>_(5)</u>	<b>NAIVE1</b>	-0.50	0.03	0.03	0.93	0.05	0.01	0.01	0.94
$\beta_2^{(2)}$	NAIVE2	-0.05	0.03	0.03	0.94	0.08	0.01	0.01	0.94
	JOINT-PQL1	-1.44	0.03	0.03	0.95	-0.40	0.01	0.01	0.95
	JOINT-PQL2	0.27	0.03	0.03	0.95	0.34	0.01	0.01	0.94
$\mathcal{L}(\mathcal{B})$	<b>NAIVE1</b>	0.22	0.02	0.02	0.97	0.01	0.01	0.01	0.95
$\beta_3^{(c)}$	NAIVE2	-0.38	0.02	0.02	0.95	0.09	0.01	0.01	0.94
	JOINT-PQL1	-0.91	0.02	0.02	0.95	0.13	0.01	0.01	0.95
	JOINT-PQL2	-0.62	0.02	0.02	0.96	0.30	0.01	0.01	0.94
<sup>0</sup> (6)	<b>NAIVE1</b>	4.83	0.07	0.08	0.90	1.95	0.04	0.05	0.93
$\beta_0^{(u)}$	NAIVE2	4.99	0.07	0.07	0.91	1.86	0.04	0.05	0.93
	JOINT-PQL1	-34.49	0.06	0.06	0.19	-25.18	0.04	0.04	0.12
	JOINT-PQL2	4.55	0.07	0.07	0.93	1.36	0.04	0.05	0.94
$\beta_1^{(6)}$	<b>NAIVE1</b>	5.61	0.06	0.06	0.95	0.11	0.04	0.04	0.94

~
₽
-
=
<u> </u>
<u></u>
0
<
~
മ
an
anu
anu:
anus
anusc
anuscr
anuscri
anuscrip

Li et al.

		r	<i>i</i> = 200,	$J_i = 5$		r	<i>t</i> = 400,	$J_i = 9$	
		Bias%	ASE	ESE	CR	Bias%	ASE	ESE	CR
	NAIVE2	5.57	0.06	0.06	0.94	0.05	0.04	0.04	0.94
	JOINT-PQL1	-9.75	0.05	0.05	0.94	-9.25	0.04	0.04	0.93
	JOINT-PQL2	4.23	0.06	0.06	0.95	0.56	0.04	0.04	0.95
(9)	NAIVE1	-0.23	0.04	0.04	0.97	0.07	0.02	0.02	0.95
$\beta_2^{(c)}$	NAIVE2	-0.32	0.04	0.04	0.97	-0.00	0.02	0.02	0.95
	JOINT-PQL1	-4.25	0.04	0.04	0.95	-1.09	0.02	0.02	0.95
	JOINT-PQL2	0.80	0.04	0.04	0.97	0.26	0.02	0.02	0.96
(9)°	<b>NAIVE1</b>	1.04	0.03	0.03	0.97	-0.16	0.01	0.01	0.97
$\beta_3^{(c)}$	NAIVE2	-1.31	0.03	0.03	0.97	-0.06	0.01	0.01	0.97
	JOINT-PQL1	-5.04	0.03	0.03	0.94	0.37	0.01	0.01	0.95
	JOINT-PQL2	0.42	0.03	0.03	0.97	0.99	0.01	0.01	0.97

"NAIVE2" represent the approaches completely and partially ignoring the association structure among responses, respectively, and "JOINT-PQL1" is the method involving full association pattern but not using second order approximation. Displayed are the relative bias percent (Bias%), the average modelbased standard error (ASE), the empirical standard error (ESE) and 95% coverage rate (CR) for the estimates. Relative bias greater than 10% is Simulation results for  $\beta^{(7)}$  and  $\beta^{(8)}$  in binary outcomes. As defined in Section 4, "JOINT-PQL2" denotes our proposed method, "NAIVE1" and highlighted in bold.

		n	<i>i</i> = 200,	$J_i = 5$		1	t = 400,	$J_i = 9$	
		Bias%	ASE	ESE	CR	Bias%	ASE	ESE	CR
°(2)	<b>NAIVE1</b>	1.20	0.09	0.10	0.91	0.92	0.05	0.05	0.93
$\beta_0^{(.)}$	NAIVE2	2.39	0.09	0.10	0.92	0.74	0.05	0.06	0.92
	JOINT-PQL1	-15.06	0.08	0.08	0.84	-13.18	0.05	0.04	0.70
	JOINT-PQL2	1.63	0.09	0.10	0.92	0.17	0.05	0.05	0.94
(2)	NAIVE1	3.57	0.07	0.08	0.92	2.83	0.05	0.05	0.93
$\beta_1^{(\cdot)}$	NAIVE2	6.71	0.07	0.08	0.91	2.56	0.05	0.05	0.93
	JOINT-PQL1	-21.84	0.06	0.06	0.87	-19.72	0.04	0.04	0.83
	JOINT-PQL2	2.13	0.07	0.08	0.92	1.44	0.05	0.05	0.95
(2)	NAIVE1	-0.73	0.08	0.08	0.94	-1.40	0.04	0.04	0.94
$\beta_2^{(1)}$	NAIVE2	0.26	0.08	0.08	0.93	-1.26	0.04	0.04	0.93
	JOINT-PQL1	-9.87	0.07	0.07	0.92	-7.92	0.04	0.04	0.94
	JOINT-PQL2	-0.45	0.08	0.08	0.94	-0.54	0.04	0.04	0.95
$(2)^{-1}$	<b>NAIVE1</b>	-0.97	0.06	0.06	0.93	-1.79	0.03	0.03	0.96
$\beta_3^{(1)}$	NAIVE2	-3.75	0.06	0.06	0.94	-1.72	0.03	0.03	0.95
	JOINT-PQL1	-11.40	0.05	0.05	0.96	-12.47	0.03	0.03	0.95
	JOINT-PQL2	-2.35	0.06	0.07	0.93	-2.28	0.03	0.03	0.94
(8)	<b>NAIVE1</b>	3.31	0.09	0.10	0.93	0.68	0.05	0.05	0.95
$\beta_0^{(2)}$	NAIVE2	3.47	0.09	0.10	0.93	0.50	0.05	0.05	0.95
	JOINT-PQL1	-14.35	0.08	0.08	0.85	-13.04	0.05	0.05	0.72
	JOINT-PQL2	2.98	0.09	0.09	0.95	0.62	0.05	0.06	0.94

Autho	
r Manu	
script	

Li et al.

β <sup>(8)</sup> NAIV NAIV JOINT-I JOINT-I									
β <sub>1</sub> <sup>(8)</sup> NAIV NAIV JOINT-I JOINT-I		Bias%	ASE	ESE	CR	Bias%	ASE	ESE	CR
Bin NAIV JOINT-I JOINT-I	/E1	6.16	0.07	0.08	0.92	3.23	0.05	0.05	0.93
I-TNIOL JOINT-I	/E2	6.43	0.07	0.08	0.90	3.20	0.05	0.05	0.93
I-LNIOf	PQL 1	-21.24	0.06	0.06	0.90	-19.48	0.04	0.04	0.82
	PQL2	5.89	0.07	0.08	0.92	2.71	0.05	0.05	0.93
olan NAIV	/E1	2.95	0.08	0.08	0.94	-0.57	0.04	0.04	0.95
$\beta_2^{(c)}$ NAIV	/E2	1.14	0.08	0.08	0.93	-0.32	0.04	0.04	0.94
I-LNIOf	PQL 1	-10.23	0.07	0.07	0.94	-9.68	0.04	0.04	0.92
I-LNIOf	PQL2	1.53	0.08	0.08	0.93	0.23	0.04	0.04	0.94
ر(8) NAIV	Æ1	4.15	0.06	0.06	0.93	0.32	0.03	0.03	0.94
$\beta_3^{(c)}$ NAIV	/E2	0.70	0.06	0.06	0.93	0.17	0.03	0.03	0.94
I-LNIOf	PQL 1	-12.04	0.05	0.05	0.94	-12.39	0.03	0.03	0.92
I-LNIOf	PQL2	6.81	0.06	0.07	0.92	0.02	0.03	0.03	0.93

# Table 3

Physical activity data analysis results for  $\beta$  using our proposed JOINT-PQL2 approach. Displayed are the estimates (Est.), the standard error (SE), and pvalues.

		Continuc	sne			Proporti	uo
	Est.	SE	<i>p</i> -value		Est.	SE	<i>p</i> -value
$\beta_0^{(1)}$	9.77	0.30	< 0.01	$\beta_0^{(3)}$	-1.27	0.09	< 0.01
$\beta_1^{(1)}$	-0.13	0.14	0.38	$\beta_1^{(3)}$	-0.02	0.04	0.55
$\beta_2^{(1)}$	-0.05	0.40	0.91	$\beta_2^{(3)}$	-0.01	0.12	0.92
$\beta_3^{(1)}$	-0.17	0.19	0.38	$\beta_3^{(3)}$	-0.01	0.05	0.92
$\beta_0^{(2)}$	20.98	0.47	< 0.01	$\beta_0^{(4)}$	-0.75	0.07	< 0.01
$\beta_1^{(2)}$	0.13	0.25	0.59	$\beta_1^{(4)}$	0.01	0.03	0.80
$\beta_2^{(2)}$	1.67	0.64	0.01	$\beta_2^{(4)}$	0.10	0.10	0.30
$\beta_3^{(2)}$	0.12	0.34	0.73	$\beta_3^{(4)}$	0.04	0.05	0.36
		Count				Binary	
	Est.	SE	<i>p</i> -value		Est.	SE	<i>p</i> -value
$\beta_0^{(5)}$	3.75	0.05	< 0.01	$\beta_0^{(7)}$	-3.70	0.54	< 0.01
$\beta_1^{(5)}$	-0.01	0.02	0.70	$\beta_1^{(7)}$	0.29	0.37	0.43
$\beta_2^{(5)}$	-0.00	0.07	0.97	$\beta_2^{(7)}$	4.79	0.69	< 0.01

< 0.01

0.60

4.08

0.00

0.09

0.49

 $\beta_2^{(6)}$ 

 $\beta_2^{(8)}$ 

0.03

0.46

1.00

0.25

0.04

0.05

 $\beta_3^{(6)}$ 

 $\beta_3^{(8)}$ 

0.08	< 0.01	0.10
0.48	0.41	0.31
0.83	-1.58	0.52
$\beta_3^{(7)}$	$\beta_0^{(8)}$	$\beta_1^{(8)}$
	-	
0.49	< 0.01	0.11
0.03 0.49	0.06 < 0.01	0.03 0.11

 $\beta_3^{(5)} \\ \beta_0^{(6)} \\ \beta_1^{(6)}$ 

Stat Med. Author manuscript; available in PMC 2018 November 10.

Author Manuscript