Probabilistic principal component analysis to identify profiles of physical activity behaviours in the presence of nonignorable missing data

Marco Geraci†

University of South Carolina, Columbia, South Carolina, USA

Alessio Farcomeni

Sapienza – University of Rome, Rome, Italy

Summary. Our paper is motivated by an accelerometer-based study of physical activity (PA) behaviours in a large cohort of UK school-aged children. Advances in research on PA are accompanied by a growing number of results that are contributing to form a complex picture of PA behaviours in children. One source of such complexity is intimately related to the multiplicity of dimensions associated with PA. Currently a comprehensive individual accelerometer summary can include a large number of outcomes and this clearly poses challenges for the analysis. We explore the application of principal component analysis (PCA) to accelerometer measurements that are aggregated daily over several days of the week and are affected by missingness. The probabilistic approach to PCA with latent scores is extended to include nonignorable missing data. The extended likelihood is maximised through a Monte Carlo EM algorithm via adaptive rejection Metropolis sampling. Our findings suggest that physical activity and inactivity are two dimensions over which children aggregate into distinct behavioural profiles, characterised by gender and season but not by anthropometric factors.

Keywords: body mass index; latent scores; Millennium Cohort Study; obesity; sedentary behaviour

1. Introduction

Physical inactivity is an important modifiable risk factor for a number of diseases, including cardiovascular and chronic diseases such as colorectal cancer, depression and diabetes mellitus (Warburton et al., 2006). Physical inactivity in England is estimated to cost more than eight billion British pounds a year. This includes both the direct costs of treating major, lifestyle-related diseases and the indirect costs of sickness absence (National Institute for Health and Clinical Excellence, 2008). It is also estimated that 54,000 premature deaths a year are linked to a sedentary lifestyle (Department for Culture, Media and Sport, 2002).

In July 2011, the United Kingdom (UK) physical activity guidelines were changed to meet new scientific evidence on the benefits of physical activity and the potential harms of inactivity across the entire life course. Promotion of physical activity in children and young people (National Institute for Health and Clinical Excellence, 2009) has therefore gained a strategically central place in the public health agenda since acquiring active lifestyles at early ages may impact not only current but also future health of individuals. However, scientific evidence relating

†*Address for correspondence:* Department of Epidemiology and Biostatistics, Arnold School of Public Health, University of South Carolina, 915 Greene Street, Columbia SC 29208, USA.
E-mail: geraci@mailbox.sc.edu

to a number of important aspects of the UK physical activity policy is lacking (Bull and the Expert Working Groups, 2010). A recent study (Griffiths et al., 2013) showed that only half of UK seven-year-olds achieve recommended levels of physical activity, with girls far less active than boys, calling for population-wide interventions.

To this end, data for research and surveillance purposes are being increasingly obtained on a large scale using accelerometers, devices capable of providing an objective measure of the intensity, duration and frequency of physical activities. Accelerometer-based measurements can also be used to identify sedentary behaviours, which are often distinguished from activity absence since they consist of purposeful engagement in activities that involve minimal movement and low energy expenditure. It has been suggested that sedentary behaviour comprises the majority of young children's time and that children establish a sedentary lifestyle at an early age (Reilly et al., 2004). Sedentariness is often considered as an independent risk factor for a number of metabolic disorders (Ekelund et al., 2006) with its own patterns and determinants (Biddle et al., 2004), rather than simply an extreme of the physical activity continuum. Interestingly, physical activity and inactivity seem to be two dimensions over which adolescents aggregate into distinct clusters (Marshall et al., 2002; Biddle et al., 2004) and these clusters can be identified and characterised by demographic, anthropometric and lifestyle factors. For example, distinct classes of activity and sedentary behaviours were identified among a cohort of children and adolescents by Heitzler et al. (2011) using data from accelerometers and questionnaires. Moreover, these classes differed by school grade, ethnicity, family living arrangements, and overweight status of the study participants (Heitzler et al., 2011). It is also worth mentioning that type and context of physical activity behaviours may vary among adolescents with different activity levels (Koorts et al., 2011).

The data collected for the project on physical activity determinants (Griffiths et al., 2013) in seven-year-old children of the Millennium Cohort Study (MCS), a UK-wide longitudinal multi-purpose survey, represent a major and unique epidemiological resource. In fact, there are currently no equivalent studies of accelerometer-determined activity on this scale ($n \sim 9{,}000$) in pre-pubertal children from across the UK. Figure 1 shows the time series plot of data collected at 15-second epochs (i.e., the interval at which data samples are accumulated before being stored in memory) for one child of the MCS. Accelerometers produce an output known as 'acceleration counts' which is dimensionless and thus requires calibration in order to be converted into physiologically more relevant units, usually based on energy expended in units of time (e.g., metabolic equivalent of task). Physical activity is typically classified into categories of varying levels of intensity, with sedentary behaviour on one end of the spectrum, followed by light, moderate and vigorous activity. Classification thresholds obtained from a calibration study in 7-year old children (Pulsford et al., 2011) are shown in Figure 1.

Accelerometer data are typically collected at a high sampling frequency over a relatively extended period of time. In large-scale epidemiological studies, the temporal framework of the survey protocol can extend to several days. For example, MCS children participating to the study were asked to wear the monitor for seven consecutive days during waking hours, with the purpose of observing a representative sample of weekly patterns of activity in free-living conditions. The availability of measurements in such abundance, which can easily exceed millions in number, offers opportunities for an increased understanding of issues of public health importance as well as challenges for the development of statistical methods capable of linking, analysing and extracting information from large complex data sets. The need for appropriate statistical, computational and database management strategies is discussed by Staudenmayer et al. (2012).

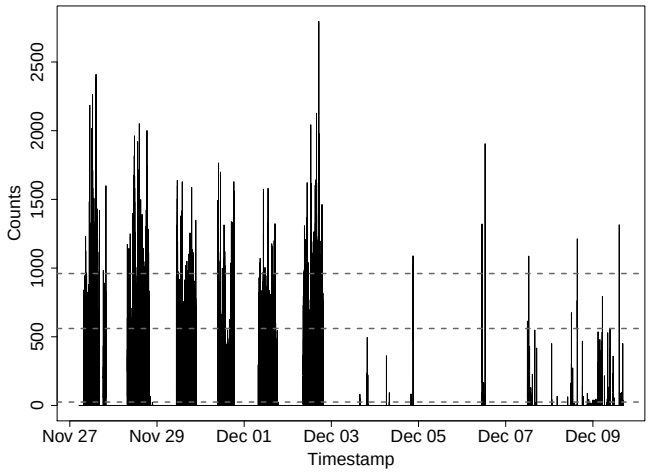

Fig. 1. Plot of accelerometer counts collected in 15-second epochs for one child of the Millennium Cohort Study. Dashed grey lines mark thresholds for sedentary behaviour (0–25 counts), light (25–560 counts), moderate (560–960 counts) and vigorous activity ($>$ 960 counts).

Yet, in epidemiological analyses data are often aggregated at the individual level. On the one hand, this allows one to use a few numbers to summarise the data and to capture *large-scale* effects; on the other hand, high levels of aggregation may determine an important loss of information on *small-scale* effects. Methods for time-series analysis like spectral analysis and functional PCA aim at discovering small-scale structures using data sampled at high frequencies. Therefore, they have been designed to cope with temporally correlated observations. For example, a long run of accelerometer measurements on an epoch-by-epoch basis over a number of days can be opportunely analysed by smoothing mean levels of activity over time using functional data methods (Morris et al., 2006; Sera et al., 2011). Other methods such as standard PCA or regression analysis can be applied only to independent observations and therefore require data aggregation. The present study moves within a middle ground in which it is possible to retain some of the structure in the data while reducing the modelling and computing effort when information at a small-scale is not requested. In particular, we consider data aggregated by day of the week. While this level of the analysis preserves part of the longitudinal structure of the data, it is also in tune with the objectives of the MCS accelerometer study protocol. A lower level of aggregation could be, for example, the hour of the day.

Advances in research on physical activity are accompanied by a growing number of results that are contributing to form a complex picture of the relationship between physical activity and mechanisms of growth and development in children. One source of such complexity is intimately related to the multiplicity of dimensions associated with physical activity. A comprehensive physical activity profile can be made of a large number of different outcomes, including daily summaries of amount and accumulation of physical activity, classified by levels of energy expenditure, time of the day and modality (Esliger and Tremblay, 2007), and this clearly poses challenges for the analysis. Table 1 illustrates an example of an individual summary by day of the week. The amount of activity is typically summarised by counts and time spent at different levels of intensity, while accumulation of physical activity is measured in terms of bouts, i.e. short bursts of intense activity. Table 1 also reports the total number of steps which were collected simultaneously with acceleration counts in the MCS accelerometer study.

Table 1. Example of weekly physical activity profile summary with seven outcome variables for one child of the Millennium Cohort Study.

		Mon	*Tue*	...	*Sun*
Total activity					
(1)	counts (×1000)	287.7	564.7	...	305.7
(2)	steps (×1000)	7.4	13.3	...	6.5
Proportion of time (%) spent					
(3)	in sedentary behaviour	61.4	51.5	...	57.9
(4)	in light activity	33.6	37.5	...	36.7
(5)	in moderate-to-vigorous activity	5.0	11.0	...	5.4
Sporadic (<10 minutes) moderate-to-vigorous bouts					
(6)	total time (minutes)	37.5	92.75	...	38
(7)	frequency	91	169	...	96

These outcomes are strongly correlated and although independent and interactive effects of physical activity and inactivity on child development have been acknowledged (Ekelund et al., 2006; Hamer et al., 2009; Mitchell et al., 2009; Drenowatz et al., 2012), there is very little account of statistical methods that jointly analyse multiple physical activity outcomes. A rare example is offered by Eberth and Smith (2010) who used copula models to analyse the propensity to participate in sporting activities and the duration of sporting activity undertaken. The application of a latent class analysis (Heitzler et al., 2011) represents a more recent example of joint modelling. In contrast, univariate analyses of physical activity and inactivity outcomes ignore their correlation structure and may potentially lead to a loss of information.

In this study we consider the application of principal component analysis (PCA) to identify profiles of active and sedentary behaviours using accelerometer data. PCA, originally introduced by Pearson (1901), is arguably one the most popular multivariate analysis techniques often described as a tool for dimensionality reduction. Some authors consider PCA as a descriptive method, which needs not introducing distributional assumptions (see for example Jolliffe, 2002), whereas others provide probabilistic justifications in terms of, for example, multivariate normality (Tipping and Bishop, 1999). Here we take a probabilistic view as it allows a seamless extension of standard methods to deal with some of the statistical problems that arise in population-based studies such as the MCS physical activity survey.

In fact, central to this paper is the problem of missing data in accelerometer-based studies. Noncompliance with study protocol accounts for a large fraction of the missing data (Catellier et al., 2005), but other reasons may include, for example, removal of non-waterproof monitors during aquatic activities or during sport activities in which these devices are prohibited by safety regulations. In a recent study on predictors of nonresponse using the MCS data (Rich et al., 2013), it was found that children who exercised once a week or less according to the MCS questionnaire-based data, were less likely to provide reliable accelerometer measurements for all days of the week. Concerns about the informativeness of the missing data process are therefore justifiable and statistical methods to reduce associated bias (Little and Rubin, 2002) are warranted.

A number of statistical methods have been developed to cope with nonignorable missing mechanisms (i.e., processes where the probability of being missing depends on the unobserved data after conditioning on observed information) and such methods have been applied in different analytic frameworks. However, approaches to missing data in PCA have typically focussed on assumptions of ignorability (Tipping and Bishop, 1999; Ilin and Raiko, 2010; Josse and

Husson, 2012a).

Our modelling strategy will address the following issues: (a) standard PCA requires independence between observations (i.e., between rows), however observations in different days of the week are clustered within child. Repeated observations will be treated as variables (i.e., columns); (b) children contribute different number of days which may or may not include weekend days, when physical activity tends to follow distinct patterns from those in the rest of the week; the problem of nonignorable missing data will be addressed by extending methods developed for random-effects models to PPCA, which represents the novel methodological contribution of this paper. We refer the reader to Tipping and Bishop (1999) for a discussion on the rationale of PPCA and its links with least-squares and factor analysis. We stress, however, that the adoption of PPCA here is justified by its computational advantages and this will be shown in the next sections; (c) finally a quantile regression analysis is carried out to study the association between PPCA scores and a number of covariates of interest.

In Section 2 we introduce data and methods, with further technical details in Appendix A. We also assess the performance of our proposed methods in a simulation study and report the results in Appendix B. In Section 3 we present the results of the analysis using the MCS data. In Section 4 we discuss avenues for future research and in Section 5 we conclude with final remarks.

2. Statistical methods for accelerometer data

2.1. The MCS data

The MCS is a longitudinal UK-wide study of children born in the new century (Smith and Joshi, 2002). The first survey was conducted when the children were aged 9 months old, followed by four surveys at ages 3, 5, 7 and 11 years. Physical activity monitoring was carried out at age 7 years. All MCS children (14043) were invited to wear an accelerometer upon written consent from parents or guardians. Accelerometers were distributed between May 2008 and August 2009 by post. Children were asked to wear an ActiGraph GT1M accelerometer (ActiGraph, Florida, USA) the morning after they received it for seven consecutive days during all waking hours, but to remove it during aquatic activities as the accelerometers are not waterproof. The devices were initialised to start collecting acceleration counts and steps two days after posting at 5:00 a.m. The accelerometers were returned after the monitoring period in a pre-paid envelope (Griffiths et al., 2013). The accelerometer files were downloaded only if the parents reported that the accelerometer had been worn by their child. The analysis that follows is restricted to 5682 singleton children of white ethnic background.

The MCS accelerometer files were processed following a number of data quality procedures using the package `pawacc` (Geraci, 2014c) for the statistical environment `R` (R Development Core Team, 2014). As full details are given elsewhere (Geraci et al., 2012), here we report selected processing criteria. Non-wear time status was defined as 20 minutes or more of consecutive zero-counts on an epoch-by-epoch basis, while accelerometer counts were classified using the cut-offs reported in Figure 1. In addition, a small number of extreme counts were removed before data aggregation (Rich et al., 2014). Only participants contributing with at least two valid unique days of the week were retained in the dataset, where 'valid day' here is defined as a day including at least ten hours of wear time. This reliability threshold was determined in a previous study using the same data (Rich et al., 2013).

The outcome variables were derived for each day of the week as shown in Table 1. Since a small number of children provided data recorded for over a week, the outcomes were averaged

Table 2. Number of children by number of valid days, overall (n) and with at least one weekend day (n^*) in the Millennium Cohort Study data. The expected number of child-days is given by $n \cdot 7$, while the proportion (%) of missing child-days is given by $1 - d/7$.

Number of days d	*Number of children* n	n^*	*Expected child-days* (% *missing*)
2	309	89	2163 (71)
3	396	156	2772 (57)
4	696	358	4872 (43)
5	1072	726	7504 (29)
6	1569	1569	10983 (14)
7	1640	1640	11480 (0)
Overall	5682	4538	39774 (22)

out if these were available for the same day of the week on two or more calendar dates. Table 2 reports the distribution of children by number of valid days (d). There is clearly a loss of information as determined by the large proportion of children (4042 out of 5682) who did not provide valid observations for all seven days of the week. The quantity $1 - n \cdot d/n \cdot 7 = 1 - d/7$ defines the proportion of missing expected child-days ($n \cdot 7$) for a given number of valid days, and this ranged from 14% ($d = 6$) to 71% ($d = 2$). Overall, 22% of expected child-days were missing. Moreover, the proportion (%) of missing values by day of the week was equal to 17 (Monday), 18 (Tuesday), 16 (Wednesday), 15 (Thursday), 15 (Friday), 32 (Saturday), and 44 (Sunday).

2.2. Gaussian PPCA

Let $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})^\top$ be the $p \times 1$ random vector collecting the variables of interest for the ith subject, $i = 1, \ldots, n$. The probabilistic representation of PCA (Tipping and Bishop, 1999) is given by the model

$$\mathbf{y}_i = \boldsymbol{\mu} + W\mathbf{u}_i + \boldsymbol{\epsilon}_i, \qquad i = 1, \ldots, n, \tag{1}$$

where $\mathbf{u}_i = (u_{i1}, \ldots, u_{iq})^\top$ is a $q \times 1$, $q \leq p$, vector of (latent) principal components and W is a $p \times q$ matrix of loadings with elements w_{jh}, $j = 1, \ldots, p$, $h = 1, \ldots, q$. Furthermore, it is assumed that $\mathbf{u}$ is stochastically independent from $\boldsymbol{\epsilon}$ (implicitly assumed by Tipping and Bishop, 1999). Conventionally, $\mathbf{u}_i \sim \mathcal{N}(\mathbf{0}, I_q)$, where I_q denotes the identity matrix of order q. If in addition the error is assumed to be zero-centered Gaussian with covariance matrix Ψ, $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \Psi)$, we obtain the multivariate normal distribution $\mathbf{y}_i \sim \mathcal{N}(\boldsymbol{\mu}, C)$, $C = WW^\top + \Psi$. We also assume that $\Psi = \psi I_p$, $\psi \in \mathbb{R}_+$, so that the elements of $\mathbf{y}_i$ are conditionally independent, given $\mathbf{u}_i$. The parameter $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^\top$ allows for a location-shift fixed effect.

The marginal log-likelihood is thus given by

$$l(\boldsymbol{\theta}; Y) \equiv \sum_{i=1}^{n} \log \int_{\mathbb{R}^q} f(\mathbf{y}_i, \mathbf{u}_i)\, d\mathbf{u}_i = -\frac{n}{2} \left\{ p \log(2\pi) + \log|C| + \operatorname{tr}\left(C^{-1}S\right) \right\}, \tag{2}$$

where $f(\mathbf{y}_i, \mathbf{u}_i) = f(\mathbf{y}_i|\mathbf{u}_i)\, f(\mathbf{u}_i)$ is the joint density of the response and the random effects and $S = {}^1\!/_n \sum_{i=1}^{n} = (\mathbf{y}_i - \boldsymbol{\mu})(\mathbf{y}_i - \boldsymbol{\mu})^\top$. The parameter $\boldsymbol{\theta} = (\boldsymbol{\mu}, W, \psi)$ can be estimated by using

the maximum likelihood estimation (MLE) equations

$$\begin{aligned}\hat{\boldsymbol{\mu}} &= \frac{1}{n} Y^\top \mathbf{1}_n, \\ \hat{W} &= H_q \left(\Delta_q - \psi I_q\right)^{1/2} R, \\ \hat{\psi} &= \frac{1}{p-q} \sum_{j=q+1}^{p} \delta_j, \end{aligned} \tag{3}$$

where $Y = \left[\mathbf{y}_i^\top\right]_{1 \le i \le n}$ is the $n \times p$ response matrix, $\mathbf{1}_n$ denotes the $n \times 1$ vector of ones, H_q is a $p \times q$ matrix that collects the principal eigenvectors of the sample covariance matrix (obtained by replacing $\boldsymbol{\mu}$ with $\hat{\boldsymbol{\mu}}$ in S), with corresponding eigenvalues $\delta_1, \ldots, \delta_q$ in the diagonal matrix Δ_q. R is an arbitrary $q \times q$ orthogonal rotation matrix, e.g. $R = I_q$. Note that since $\boldsymbol{\Psi}$ is restricted, we avoid the trivial situation in which the likelihood (2) is optimised for $\hat{W} = \mathbf{0}$ which would obviously be uninformative about the latent structure. Finally, the individual scores can be predicted via best linear prediction $\hat{\mathbf{u}} \equiv \mathrm{E}\left(\mathbf{u}|\mathbf{y}\right) = \left(W^\top W + \Psi\right)^{-1} W^\top (\mathbf{y} - \boldsymbol{\mu})$.

2.3. EM algorithm for PPCA

In their seminal paper, Dempster et al. (1977) discussed the use of the EM algorithm in factor analysis. Tipping and Bishop (1999) also proposed an EM algorithm to estimate the parameter $\boldsymbol{\theta}$, where the incomplete part of the data is represented by the latent variables $\mathbf{u}$. Briefly, given the complete data log-likelihood

$$l\left(\boldsymbol{\theta}; Y, U\right) = \sum_{i}^{n} \log f\left(\mathbf{y}_i | \mathbf{u}_i, \boldsymbol{\theta}\right) + \log f\left(\mathbf{u}_i\right), \tag{4}$$

where $U = \left[\mathbf{u}_i^\top\right]_{1 \le i \le n}$, the EM approach alternates between an

(i) expectation step (E-step) $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathrm{E}_{\mathbf{u}|\mathbf{y},\boldsymbol{\theta}^{(t)}} \left\{l\left(\boldsymbol{\theta}; Y, U\right)\right\}$; and a

(ii) maximisation step (M-step) $\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$,

where $\boldsymbol{\theta}^{(t)}$ is the estimate of the parameter after t cycles. The expectation in step (i) is taken with respect to the conditional distribution $f\left(\mathbf{u}|\mathbf{y}, \boldsymbol{\theta}^{(t)}\right)$ which, in the Gaussian PPCA, is normal with parameters depending on $\boldsymbol{\theta}^{(t)}$ only. Closed-form solutions to steps (i) and (ii) are provided by Tipping and Bishop (1999).

The EM estimation approach might have little immediate appeal as opposed to the usual diagonalisation of the sample covariance matrix. However, it provides a computationally efficient strategy in analysing high-dimensional large datasets (Roweis, 1998) and it is particularly enticing when $\mathbf{y}$ has missing values. An excellent overview on the vast literature of the EM algorithm is given in McLachlan and Krishnan (2008).

We now consider extending Tipping and Bishop's (1999) EM approach to the case in which the vector $\mathbf{y}$ is partially observed and the missing data mechanism is nonignorable.

2.4. Nonignorable missing data

Suppose that $\mathbf{y}_i$ contains s_i, $s_i < p$, missing values. Let us denote with $\mathbf{m}_i$, $i = 1, \ldots, n$, the $p \times 1$ missing data indicator whose jth element m_{ij} equals 1 if y_{ij} is missing and 0 otherwise, and let $M = \left[\mathbf{m}_i^\top\right]_{1 \le i \le n}$ be the corresponding $n \times p$ matrix. Here we use the convenient notation

$\mathbf{y}_i = (\mathbf{z}_i, \mathbf{x}_i)$ which partitions the response vector in missing ($\mathbf{z}_i$) and observed ($\mathbf{x}_i$) components of dimensions $s_i \times 1$ and $(p - s_i) \times 1$, respectively. The modelling approach that follows is an adaptation of Ibrahim et al.'s (2001) methods for missing responses in random-effects models with nonmonotone patterns of missing data.

The ith contribution to the complete data density of $(\mathbf{y}_i, \mathbf{u}_i, \mathbf{m}_i)$ is given by

$$f(\mathbf{y}_i, \mathbf{u}_i, \mathbf{m}_i | \boldsymbol{\theta}, \boldsymbol{\eta}) = f(\mathbf{y}_i | \mathbf{u}_i, \boldsymbol{\theta}) f(\mathbf{u}_i) f(\mathbf{m}_i | \mathbf{y}_i, \boldsymbol{\eta}), \tag{5}$$

where the additional factor $f(\mathbf{m}_i | \mathbf{y}_i, \boldsymbol{\eta})$, indexed by the parameter $\boldsymbol{\eta}$, is the missing data mechanism (MDM) (Little and Rubin, 2002), which we assume to be independent from $\mathbf{u}_i$. This assumption simplifies the subsequent steps of the estimation algorithm, although it can be relaxed at the cost of increased computational time (see Ibrahim et al., 2001; Ibrahim and Molenberghs, 2009, for a discussion).

Note that if the data are missing at random (MAR), i.e. $f(\mathbf{m}_i | \mathbf{y}_i, \boldsymbol{\eta}) = f(\mathbf{m}_i | \mathbf{x}_i, \boldsymbol{\eta})$, and $\boldsymbol{\eta}$ is separate from the parameter of interest $\boldsymbol{\theta}$, then the MDM can be ignored when concerned with inference about $\boldsymbol{\theta}$. The assumption of ignorability cannot be tested from within the dataset used in the analysis. However, different models for $f(\mathbf{m}_i | \mathbf{y}_i, \boldsymbol{\eta})$ can be investigated and sensitivity of the results to different modelling choices can be assessed formally.

Estimation of $\boldsymbol{\theta}$ would in general require marginalizing the log-likelihood based on (5) over the unobserved data, which however leads to a rather intractable integral of dimension $s_i + q$. Instead, we develop an EM algorithm where the E-step at the $(t+1)$th iteration is defined as follows:

$$Q(\boldsymbol{\lambda} | \boldsymbol{\lambda}^{(t)}) = \mathrm{E}_{\mathbf{z},\mathbf{u}|\mathbf{x},\mathbf{m},\boldsymbol{\lambda}^{(t)}} \left\{ l(\boldsymbol{\lambda}; Y, U, M) \right\}, \tag{6}$$

with $\boldsymbol{\lambda} = (\boldsymbol{\theta}, \boldsymbol{\eta})$ and $l(\boldsymbol{\lambda}; Y, U, M) = \sum_i^n \log f(\mathbf{y}_i | \mathbf{u}_i, \boldsymbol{\theta}) + \log f(\mathbf{u}_i) + \log f(\mathbf{m}_i | \mathbf{y}_i, \boldsymbol{\eta})$, and where the expectation is taken with respect to the conditional distribution of $\mathbf{z}_i$ and $\mathbf{u}_i$, given the observed data, evaluated at $\boldsymbol{\lambda}^{(t)}$.

The E-step (6), however, does not yet offer a computational advantage since it is not easy to solve analytically. We therefore apply a Monte Carlo E-step as suggested by Ibrahim et al. (2001). Sampling from $f\left(\mathbf{z}_i, \mathbf{u}_i | \mathbf{x}_i, \mathbf{m}_i, \boldsymbol{\lambda}^{(t)}\right)$ is carried out efficiently using the full conditionals

$$f\left(\mathbf{z}_i | \mathbf{x}_i, \mathbf{u}_i, \mathbf{m}_i, \boldsymbol{\lambda}^{(t)}\right) \propto f\left(\mathbf{y}_i | \mathbf{u}_i, \boldsymbol{\lambda}^{(t)}\right) f\left(\mathbf{m}_i | \mathbf{y}_i, \boldsymbol{\lambda}^{(t)}\right), \tag{7}$$

$$f\left(\mathbf{u}_i | \mathbf{x}_i, \mathbf{z}_i, \mathbf{m}_i, \boldsymbol{\lambda}^{(t)}\right) \propto f\left(\mathbf{y}_i | \mathbf{u}_i, \boldsymbol{\lambda}^{(t)}\right) f(\mathbf{u}_i). \tag{8}$$

Here we consider an adaptive rejection Metropolis sampling (ARMS) algorithm (Gilks and Wild, 1992), as implemented in the `R` package `HI` (Petris and Tardella, 2013), although in principle other samplers can be considered.

As for the MDM, Ibrahim et al. (2001) proposed a binomial model of the type

$$f(\mathbf{m}_i | \mathbf{y}_i, \boldsymbol{\eta}) = \prod_{j=1}^{p} p_{ij}^{m_{ij}} (1 - p_{ij})^{1 - m_{ij}}, \tag{9}$$

where p_{ij} is the probability that y_{ij} is missing, conditional on the response $\mathbf{y}_i$. The ARMS algorithm is convenient as, basically, no tuning is needed. In addition, it is run in parallel for each row of the data matrix, which therefore greatly speeds up the computation.

Finally, we obtain a sample $\boldsymbol{\xi}_{i1}, \ldots, \boldsymbol{\xi}_{iK}$ for $i = 1, \ldots, n$ at each EM iteration t, where the $(s_i + q) \times 1$ vector $\boldsymbol{\xi}_{ik} = (\tilde{\mathbf{z}}_{ik}, \tilde{\mathbf{u}}_{ik})$, $k = 1, \ldots, K$, contains 'imputed' values for $\mathbf{z}_i$ and $\mathbf{u}_i$ (with the understanding that $\boldsymbol{\xi}_{ik} = \tilde{\mathbf{u}}_{ik}$ if $s_i = 0$). Here the Monte Carlo sample size K is kept constant

Table 3. Correlation matrix of physical activity outcomes averaged over the week in the Millennium Cohort Study (complete cases only). See Table 1 for correspondence between number codes and variables names.

	(2)	(3)	(4)	(5)	(6)	(7)
(1)	0.82	−0.75	0.24	0.91	0.95	0.89
(2)		−0.65	0.22	0.74	0.78	0.68
(3)			−0.48	−0.69	−0.66	−0.73
(4)				0.17	0.16	0.24
(5)					0.97	0.85
(6)						0.90

throughout. Alternative strategies with varying $K^{(t)}$ that may increase the speed of the EM algorithm can be considered (Booth and Hobert, 1999; Ibrahim et al., 2001). The E-step (6) is approximated by

$$Q(\boldsymbol{\lambda}|\boldsymbol{\lambda}^{(t)}) = \frac{1}{K}\sum_{i=1}^{n}\sum_{k=1}^{K} l\left(\boldsymbol{\lambda}; \boldsymbol{\xi}_{ik}, \mathbf{x}_i, \mathbf{m}_i\right). \tag{10}$$

The maximisation of (10) with respect to $\boldsymbol{\lambda}$ is straightforward. Define $\tilde{\mathbf{y}}_{ik} = (\tilde{\mathbf{z}}_{ik}, \mathbf{x}_i)$ if $s_i > 0$ or $\tilde{\mathbf{y}}_{ik} = \mathbf{y}_i$ if $s_i = 0$, $i = 1, \ldots, n$, $k = 1, \ldots, K$. The maximum likelihood solution of the M-step at the $(t+1)$th iteration is given by

$$\hat{\boldsymbol{\mu}}^{(t+1)} = \frac{1}{nK}\sum_{i=1}^{n}\sum_{k=1}^{K}\left(\tilde{\mathbf{y}}_{ik} - \hat{W}^{(t)}\tilde{\mathbf{u}}_{ik}\right), \tag{11}$$

$$\hat{W}^{(t+1)} = \left\{\sum_{i=1}^{n}\sum_{k=1}^{K}\left(\tilde{\mathbf{y}}_{ik} - \hat{\boldsymbol{\mu}}^{(t+1)}\right)\tilde{\mathbf{u}}_{ik}^{\top}\right\}\left(\sum_{i=1}^{n}\sum_{k=1}^{K}\tilde{\mathbf{u}}_{ik}\tilde{\mathbf{u}}_{ik}^{\top}\right)^{-1}, \tag{12}$$

$$\hat{\psi}^{(t+1)} = \frac{1}{nKp}\sum_{i=1}^{n}\sum_{k=1}^{K}\|\tilde{\mathbf{y}}_{ik} - \hat{\boldsymbol{\mu}}^{(t+1)} - \hat{W}^{(t+1)}\tilde{\mathbf{u}}_{ik}\|_2^2. \tag{13}$$

Analogously, the MLE of $\boldsymbol{\eta}$ can be easily obtained using standard results for generalised linear models.

Note that the computational burden can be alleviated by first integrating out the random effects in (6) and then sampling from $f\left(\mathbf{z}_i|\mathbf{x}_i, \mathbf{m}_i, \boldsymbol{\lambda}^{(t)}\right)$ during the Monte Carlo E-step. Further details are provided in Appendix A, while a simulation study to evaluate the proposed methods is presented in Appendix B.

3. Results

3.1. Complete-case PPCA

In this section we restrict our attention on subjects that have been observed on all seven days of the week ($n = 1640$) and we carry out a PPCA as described in Section 2.2, that is, by ignoring the missing data mechanism (complete-case analysis). The results that follow cannot be used to draw conclusions since they are possibly biased. However, such results can be used to inform the next modelling step and for assessing sensitivity to assumptions on the missing data mechanism (see Section 3.2).

Table 3 shows the correlation matrix of the variables reported in Table 1. As expected, acceleration counts (1) and steps (2) were positively correlated one to each other since they

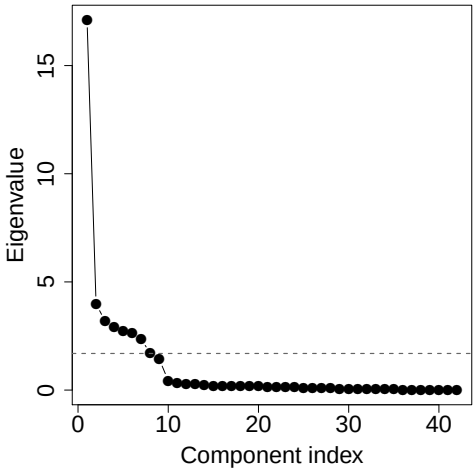

Fig. 2. Screeplot of eigenvalues obtained from the probabilistic principal component analysis of physical activity outcomes in the Millennium Cohort Study (complete cases only).

both quantify the amount of movement. As such, they showed a negative correlation with sedentary behaviour (3) but positive with all other variables measuring activity. Time spent at light physical activity level (4) had the smallest (in magnitude) correlation with all other variables.

Let us define $\mathbf{y}_i = \left(y_{i1}^{(Mon)}, \ldots, y_{i6}^{(Mon)}, y_{i7}^{(Tue)}, \ldots, y_{i12}^{(Tue)}, \ldots, y_{i42}^{(Sun)}\right)^\top$ as the ith response of dimensions 42×1 made up of the variables (1-3) and (5-7) listed in Table 1 on each day of the week (proportion of time spent in light activity was excluded as variables 3, 4, and 5 sum to one). Essentially, repeated measurements for each child are treated as columns, analogously to multivariate methods for time series data such as singular spectrum analysis (SSA) (Jolliffe, 2002). However, unlike SSA, the temporal correlation is not explicitly modelled.

An alternative approach to account for temporal effects in PPCA using longitudinal measurements has been proposed by Nyamundanda et al. (2014). They considered a model where the parameters, including the loadings, were let to vary with time. As a result, this approach yields a highly parameterised model (Nyamundanda et al., 2014), which, if followed in our case, would likely lead to nontrivial optimisation challenges given that the extension to incorporate the MDM in equation (5) already presents a computational onus. Moreover, our model, in contrast to Nyamundanda et al.'s (2014) approach, has the advantage of allowing for the temporal patterns to be associated with specific principal components and thus to be assessed in terms of their relative importance (an illustration is given further below in this section).

Given the differences between scales, PPCA (Tipping and Bishop, 1999) was applied to standardised variables. The screeplot in Figure 2 shows the eigenvalues for all 42 principal components. About 87% of the variability was explained by the first eight components (to which we limit the ensuing discussion) with estimated residual standard deviation $\hat{\psi} = 0.159$.

The barplot of the estimated loadings in Figure 3 aids interpretation of the first eight principal components. The first and most important component (40.7%) was driven by the negative correlation between sedentary and active behaviours. In other words, children with higher scores on this dimension tend to have higher levels of moderate-to-vigorous physical activity (MVPA) and lower levels of sedentary behaviour. Therefore, the first component relates to

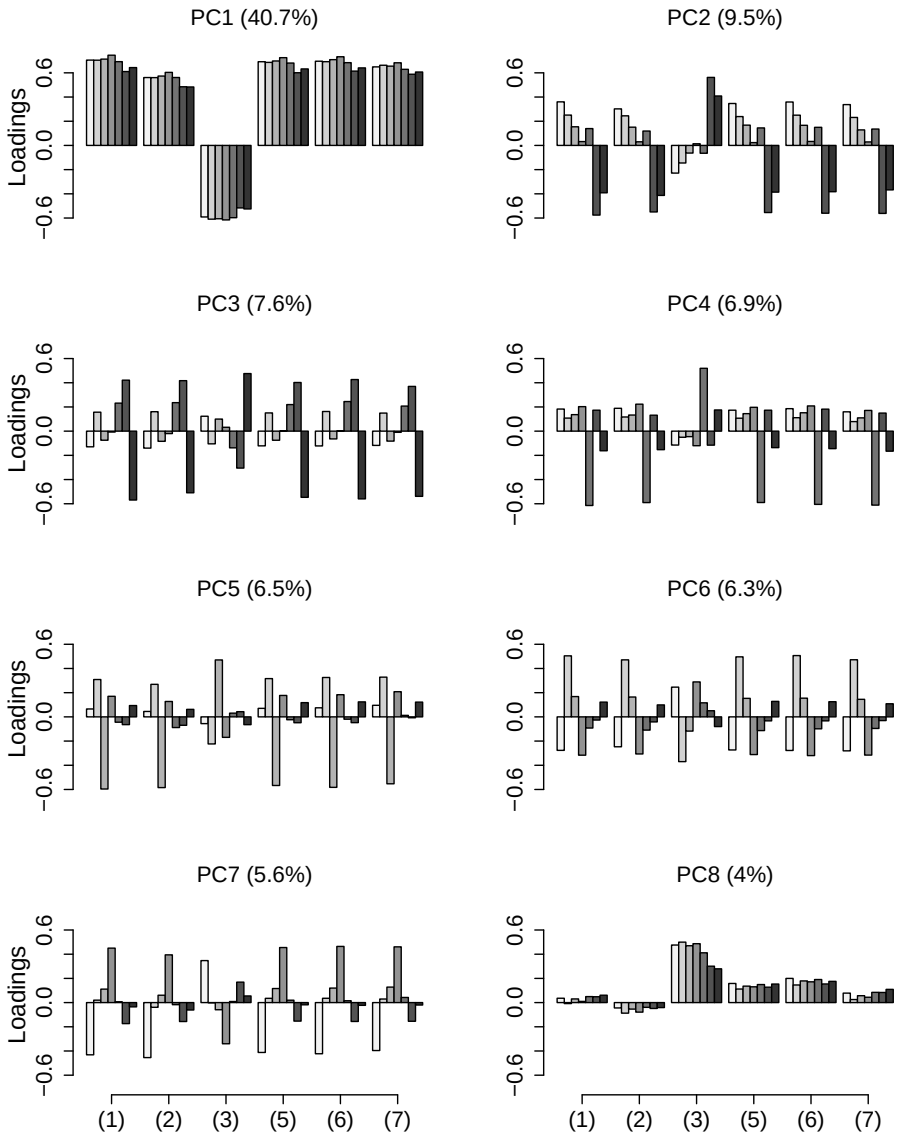

Fig. 3. Results of the principal component analysis of physical activity outcomes in the Millennium Cohort Study (complete cases only): barplots of the loadings for the first eight components and proportion (%) of variability explained. Bars for total counts (1) and steps (2), sedentary behaviour (3), moderate-to-vigorous activity (5), duration (6) and frequency (7) of bouts are colour-coded by day of the week starting from Monday (lightest grey) to Sunday (darkest grey).

the 'predominant behaviour'. The second component (9.5%) contrasted weekday and weekend activity patterns, while the third (7.6%) contrasted Saturday and Sunday patterns. Note also that, along the third component, higher activity levels on Saturday are paralleled by higher levels of sedentary behaviour on Sunday, and vice versa. Hence, it is reasonable to associate the second and third dimensions with 'weekend behaviours'.

Components four to seven, each accounting between 5.6% and 6.9% of the variability, presented correlations with activities during distinct days of the week. Finally, the eighth component (4.0%) specifically related to sedentary behaviour but was otherwise minimally or not correlated with the other outcomes. It is important to stress that, while the first component establishes a trade-off between sedentary and active behaviours, the eighth dimension determines the relative location of children in terms of sedentariness, independently from their predominant behaviour.

3.2. Modelling the missing data mechanism

In this section we analyse the MCS data using the model introduced in equation (5). For the MDM, we considered four models of decreasing level of complexity, with the first (i) and last (iv) models representing the 'extremes' of such complexity:

(i) the logistic regression model $\text{logit}\,\{p_{ij}\} = \mathbf{t}_i^\top \boldsymbol{\eta}$, where $\mathbf{t}_i$ is a 36×1 vector made up of the variables (1-3, 5, 7) listed in Table 1 on each of the seven days of the week and $\boldsymbol{\eta}$ is a 36×1 parameter vector, including an intercept.

(ii) a logistic regression as in (i) with each variable averaged over weekdays and weekend days, resulting in a 11×1 parameter vector, including an intercept;

(iii) a logistic regression with a combined measure of activity (total counts (1), total steps (2), time spent in MVPA (5), and duration of MVPA bouts (6) summed together), time spent in SB (3), and frequency of MVPA bouts (7), with each variable averaged over weekdays and weekend days, resulting in a 7×1 parameter vector, including an intercept;

(iv) the basic intercept model $\text{logit}\,\{p_{ij}\} = \eta_0$;

These models take into account the results obtained in the previous sections. In particular, we would expect to see a contrast between sedentary behaviour and MVPA due to their negative correlation. Also, as seen in Section 2.1, the proportion of missing values differed by day of the week, with fractions for Saturday and Sunday twice as much (or more) as those for weekdays, suggesting that there might be distinct missing data processes at play. Therefore, the predictors were either entered for each day of the week or aggregated by weekday/weekend. Duration of MVPA bouts was included as part of the composite variable in model (iii) above, but not separately in any of the other models to avoid identifiability issues due to its near-unity correlation with time spent in MVPA (Table 3).

We fixed to eight the number q of principal components to be estimated. This choice was motivated by the results obtained in the complete-case PCA (Section 3.1). We also investigated a generalised cross-validation (GCV) approach (Josse and Husson, 2012b) which gave $q = 11$ as optimal number of components. However, the value of the GCV criterion for $q = 11$ was not substantially different from that for $q = 8$. Finally, we estimated the simplified E-step (15) as described in Appendix A using a Monte Carlo sample size $K = 100$. As in the complete-case analysis, we took into account the different scales of the variables. The standardisation was carried out at each step of the EM algorithm, whereby variables were divided by the

Table 4. Approximated observed log-likelihoods, Akaike (AIC) and Bayesian (BIC) information criteria, and residual standard deviations (ψ) from the Millennium Cohort Study data for the probabilistic principal component analysis models based on four different missing data mechanisms defined in Section 3.2.

	(i)	(ii)	(iii)	(iv)
Log-likelihood	−234720.1	−234983.4	−236897.7	−252842.6
AIC	470814.3	471290.8	475111.4	506989.1
BIC	475379.4	475689.8	479483.9	511321.7
ψ	0.156	0.155	0.159	0.145

current standard deviation estimates (however, during the sampling step all the variables were transformed back).

Firstly, we report the approximated observed log-likelihood, Akaike (AIC) and Bayesian (BIC) information criteria (Ibrahim et al., 2008) for the four estimated PPCA models. Here the goodness of fit is defined as the value of the approximated observed log-likelihood at convergence, i.e. calculated for $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}$, while the penalty is calculated using the dimension of $\boldsymbol{\lambda} = (\boldsymbol{\theta}, \boldsymbol{\eta})$ (the latter differed between models clearly only due to the different size of the MDM's parameter $\boldsymbol{\eta}$). The information criteria provided more support to the models based on MDMs (i) and (ii), with a slight preference for the former (Table 4). Henceforth we focus on the results based on the simpler MDM model (ii) since the parameter of interest $\boldsymbol{\theta}$ did not significantly differ between the two models (results not shown).

During weekdays, the predicted probability of data being missing was lower with higher volumes of activity as measured by total counts (Table 5). Intuitively, this could be explained by the lower occurrence of non-wear periods when more activity is recorded. The negative coefficient for sedentary time on the one hand, but positive for steps and MVPA on the other, are perhaps a consequence of higher compliance rates observed between Monday and Friday, hence when children might be less active because they are involved in day-to-day routines.

In contrast, opposite associations were observed during weekend days, when the fraction of missing values tend to be much higher. These results could be interpreted as a consequence of a process by which children are more likely not to follow the study protocol if they do not participate in moderate-to-vigorous activities. This clearly leads to underestimate the volume of physical activity and the proportion of sedentary time, but only during weekend days. This finding might also explain the association found by Rich et al. (2013) between lack of exercise and nonresponse in the MCS data.

We computed the standard errors for $\boldsymbol{\eta}$ using the method of Goetghebeur and Ryan (2000) which is essentially based on Rubin's rules for multiple imputation. The coefficients in the MDM were highly significant, with p-values less than 0.001, thus supporting the importance of including the physical activity outcomes in the MDM.

We also note in Table 4 that $\hat{\psi}$ is comparable among the first three models, but somewhat lower than the others in the intercept model (iv). This suggests that including information about the physical activity outcomes in the MDM produces a shift of the relative weight of the principal components. Indeed, Figure 4 shows the contribution of the first eight eigenvalues relative to the sum of all 42 eigenvalues when ignoring the MDM as compared to that observed in the nonignorable model. There is indication that the weights are redistributed when accounting for the missing data, with a substantial reduction in weight for the first component.

Table 5. Maximum likelihood estimates and standard errors (SE) from the Millennium Cohort Study data for the missing data mechanism (ii) defined in Section 3.2.

	Weekdays		Weekends	
	Estimate	SE	Estimate	SE
Total counts	−4.682	0.162	4.239	0.090
Total steps	0.289	0.026	−0.817	0.024
Time in sedentary behaviour	−0.818	0.036	0.202	0.022
Time in MVPA	2.794	0.105	−1.759	0.054
Frequency of MVPA bouts	0.852	0.035	−1.671	0.032

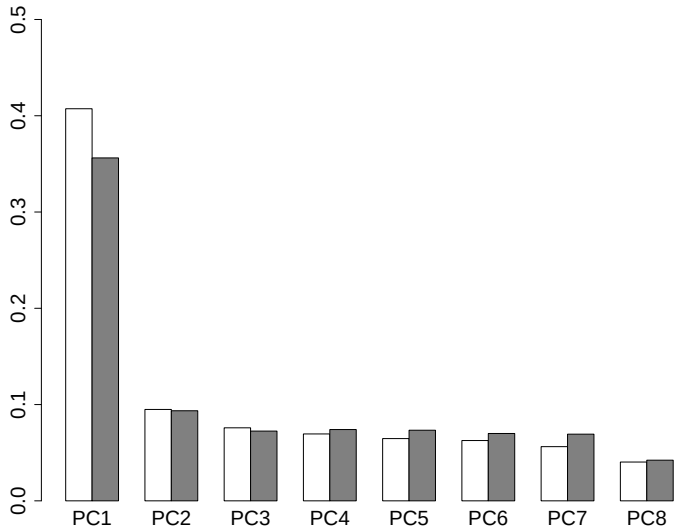

Fig. 4. Proportion of variability explained by the first eight principal components from the Millennium Cohort Study data for the ignorable (white) and nonignorable (grey) models.

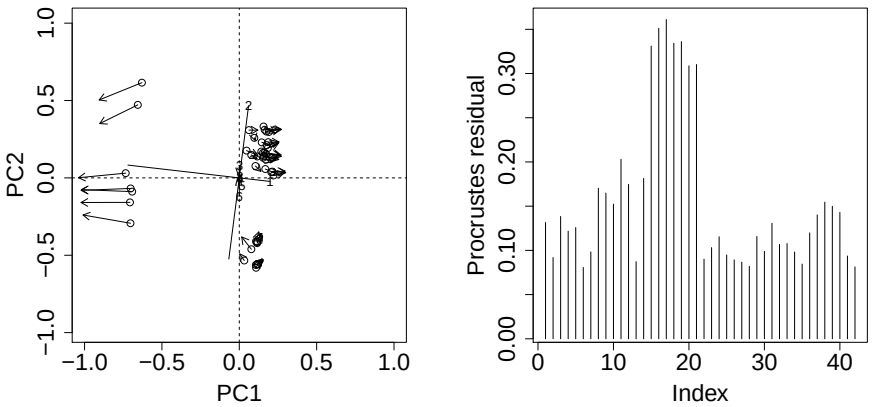

Fig. 5. Comparison between the ignorable and nonignorable models based on a Procrustes analysis of the estimated loadings in the Millennium Cohort Study: The left panel shows the shifts in configurations observed between pairs of the first (PC1) and second (PC2) principal components, while the right panel shows the diagram of Procrustes residuals.

Table 6. Average proportion of time (%) spent in sedentary behaviour (SB) and in moderate-to-vigorous physical activity (MVPA) for children grouped by predicted individual scores on the first (PC1) and second (PC2) principal components. The number of children (n) in each group is also reported.

Predominant behaviour	Sedentary (PC1 < 0)			Active (PC1 > 0)		
	SB	MVPA	n	SB	MVPA	n
Weekdays behaviour						
Less active (PC2 < 0)	55.9	6.4	1869	49.1	9.9	1064
More active (PC2 > 0)	54.6	7.4	1278	47.4	11.1	1471
Weekend behaviour						
Less sedentary (PC2 < 0)	50.7	8.0	1869	43.6	12.2	1064
More sedentary (PC2 > 0)	56.9	5.5	1278	50.2	8.7	1471

The estimated loadings, $\hat{W}$, were similar in sign as those obtained in the complete-case analysis, therefore offering a similar interpretation to that given in Section 3.1. Differences in magnitude between loadings in the nonignorable (NI) and the ignorable (IGN) models were assessed using the relative Euclidean difference, averaged over the variables space

$$\frac{1}{p}\sum_{j=1}^{p}\frac{\|\hat{\mathbf{w}}_j^{(NI)}-\hat{\mathbf{w}}_j^{(IGN)}\|_2}{\|\hat{\mathbf{w}}_j^{(IGN)}\|_2}, \tag{14}$$

after checking that corresponding axes had the same orientation. Analogously, we calculated the relative distance between the predicted scores, averaged over the individuals space. The former was equal to 1.12 and the latter to 1.26, which provide further evidence of differences between the two approaches. A graphical illustration of these differences for the first and second dimensions based on a Procrustes analysis (Oksanen et al., 2013) is given in Figure 5. The latter shows that some of the differences between the $\hat{\mathbf{w}}_j^{(NI)}$'s and the $\hat{\mathbf{w}}_j^{(IGN)}$'s were conspicuous, especially in relation to the first dimension. Moreover, the residuals for daily measurements of proportion of sedentary time (i.e., those with index between 15 and 21) were particularly large as compared to all other residuals (Figure 5). It is worth mentioning that the similarity between estimated loadings could also be assessed through a generalisation of the Pearson's correlation coefficient, like the RV coefficient (Robert and Escoufier, 1976).

3.3. Physical activity behaviours

In this section, we regress the individual predicted scores, as provided by the nonignorable model, on astronomical season, children's sex and selected anthropometric variables.

We start with some observations on groups of children with distinct physical activity behaviours. As noted in Section 3.1, the first principal component (PC1) is associated with the predominant behaviour, with scores that arrange children according to increasing levels of activity, while the second component (PC2) characterises weekend behaviours. Table 6 shows the average sedentary and MVPA time for four groups of children, one for each of the four quadrants determined by the first two components. By comparing the values across columns, it is evident that the average proportion of time spent sedentarily is larger in children with negative PC1 scores, with values ranging between 50.7% and 56.9%; in contrast, children with positive PC1 scores spend, on average, more time in MVPA, with average proportions varying between 8.7% and 12.2%. These two groups of children can be further classified based on their behaviour during weekdays or weekends: those with lower proportions of sedentary time during weekend

Table 7. Average proportion of time (%) spent in sedentary behaviour (SB) and in moderate-to-vigorous physical activity (MVPA) for children of the Millennium Cohort Study grouped by predicted individual scores on the first (PC1) and eighth (PC8) principal components. The number of children (n) in each group is also reported.

Predominant behaviour	Sedentary (PC1 < 0)			Active (PC1 > 0)		
	SB	MVPA	n	SB	MVPA	n
Sedentary behaviour						
Less sedentary (PC8 < 0)	51.9	6.6	1629	45.7	9.9	1148
More sedentary (PC8 > 0)	57.4	7.1	1518	50.5	11.2	1387

days but lower levels of activity during the rest of the week (negative PC2 scores), and, vice versa, those with symmetrically opposite behaviours (positive PC2 scores).

Table 7 shows four groups of children based on their scores on the first and the eight (PC8) principal components. As seen previously, the latter is associated with the sedentary behaviour of the children, regardless of their predominant behaviour. On average, children with positive scores on the first dimension had comparable levels of MVPA but rather dissimilar sedentary behaviours depending on their PC8 score. This was also true for children with negative PC1 scores, who were on average more sedentary than the others. These results support previous findings (Marshall et al., 2002; Biddle et al., 2004) and point to the existence of well-defined dimensions over which children aggregate into distinct sedentary and active behavioural groups.

We use a quantile regression approach (Koenker and Bassett, 1978) to investigate the effect of the covariates on the tails of the scores distributions. The PPCA results suggest that lower or higher scores reflect different levels of children's activity and, in general, classify children into distinct behavioural groups. It is therefore informative to investigate whether lower or higher score centiles are also differently related to the covariates under study. It should be stressed that our regression analysis is conditional on the PPCA results and, thus, it does not take into account the uncertainty associated with the score predictions. Alternatively, one could consider a unified strategy whereby the covariates effects are estimated within the PPCA model. This is the approach followed by, for example, Nyamundanda et al. (2010) who extended Tipping and Bishop's (1999) model to include location-shift effects in the distribution of the latent principal components u. However, the estimation of quantile effects more complex than location-shift would require abandoning the normality assumption for u adopted in a standard PPCA framework.

Seasons were defined according to calendar equinoxes and solstices. Anthropometric measurements comprised height, weight, waist circumference and body fat percentage. Since these four variables present substantial correlations, an additional PCA was carried out to reduce the dimensionality. As a result two components were derived which together accounted for 92% of the variability of the original variables (Figure 6). The interpretation was straightforward: the first component (72%) was positively correlated with all anthropometric variables. This component was interpreted as 'size'; the second component (20%) showed positive correlation with height, but negative correlation with waist and body fat. This component was interpreted as how 'slender' children are. It is interesting to note that there is a 'gradient' in overweight/obesity status (based on cut-offs in Cole et al., 2000) which runs diagonally from the upper left to the lower right quadrant, i.e. children classified as overweight and obese tend to position in the 'large and not slim' quadrant of the individuals factor map.

We estimated a sequence of 19 regression quantiles ($\tau \in \{0.05, 0.1, \ldots,$

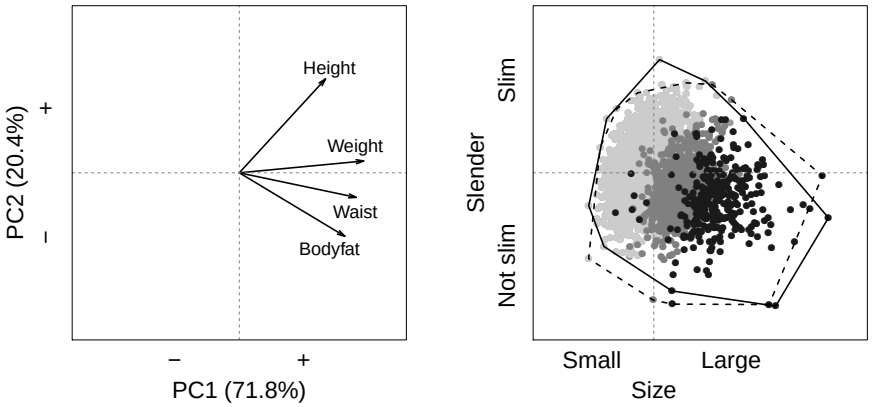

Fig. 6. Principal component analysis of body measurements in children of the Millennium Cohort Study: variables factor map (left) and individuals factor map (right) with convex hulls for boys (dashed) and girls (solid) superimposed. Weight status is displayed in light grey (normal weight and underweight), medium grey (overweight) and dark grey (obese).

0.95}) for the predicted PPCA scores on the first, the second, and the eighth components, along with the ordinary least squares (OLS) estimates. Following the interpretation of the PPCA results given in Section 3.1, lower and higher quantiles of the first component denote, respectively, predominantly sedentary and predominantly active children. Lower and higher quantiles of the second component indicate, respectively, lower and higher levels of activity (sedentariness) during the weekdays (weekend days). Finally, lower and higher quantiles of the eighth component indicate, respectively, less and more sedentary children regardless of their predominant behaviour. We also carried out a regression analysis for each of the other components but, for the sake of brevity, we report only a brief summary of the results.

Overall, boys showed higher levels of MVPA than girls (Figure 7). However, predominantly active boys (upper tail) were more active than predominantly active girls to a greater extent than predominantly sedentary girls (lower tail) were more sedentary than predominantly sedentary boys, as shown by the larger magnitude of the estimates for higher quantiles of PC1. Similarly, the estimates of the quantile effects for sex were larger on the right tail of PC2 scores, suggesting that differences between levels of activity in boys and girls were stronger during the weekdays than in the weekends. There were also significant effects of sex on PC8 scores, with boys showing higher scores than girls on average. However, all quantile effects were close to the OLS estimate.

Anthropometric measurements had little or no effect on individual scores (Figure 7). Similar results (not shown) were obtained when using height, weight, waist circumference and body fat percentage as covariates. This seems to disagree with other findings (Heitzler et al., 2011). However, in a literature review, Prentice-Dunn and Prentice-Dunn (2011) discuss the difficulties in assessing the association between childhood obesity and physical activity or sedentary behaviour when using cross-sectional data or data lacking accurate measurements of body composition and diet.

There were striking seasonal effects and these differed by quantile (Figure 8). For example, in autumn the predominant behaviour score decreased markedly as compared to summer as shown by the estimated quantile effects for PC1. Predominantly active children (upper tail) were affected to a greater extent than predominantly sedentary children (lower tail). That is,

there are changes in predominant behaviours in autumn, and these changes are stronger at higher activity levels. The estimates of the quantile effects decreased from positive to negative values across the distribution of PC2, suggesting that weekly patterns were also affected, with decreased levels of activity during the weekdays but higher during the weekends, which likely reflect the beginning of the academic year. Sedentary behaviour, as determined by PC8 scores, was on average higher in autumn than in summer, but all quantiles effects were close to the OLS estimate.

Winter and spring had, respectively, negative and positive effects on activity levels as compared to, respectively, autumn and winter. In contrast, there was little evidence of changes in weekly patterns (PC2) or sedentary behaviour (PC8) during winter or spring. Winter and spring effects were approximately constant across quantiles and equal to the OLS estimates.

Finally, the regression analyses of the other components (results not shown) revealed sex and seasonal effects, which differed by quantile of the score distributions, but, again, no effects associated with anthropometric factors.

In summary, the lesson that we learn from these observations is clear: (a) children have distinct sedentary and active behaviours, with weekdays and weekend patterns; (b) different groups of children can be identified based on their scores using a few principal components rather than a large number of physical activity outcomes; (c) these scores are strongly associated with factors such as sex and season (and presumably others as well); and (d) the strength and direction of these associations may depend on whether the scores are below or above the average. On a more practical level, these results could be used, for example, to inform interventions aimed at increasing UK children's levels of physical activity during autumn and winter, especially on weekdays, or to support initiatives for girls to stimulate their participation into vigorous physical activities.

As we restricted our analysis to UK white singletons, the conclusions should be confined to this sub-population. Future studies may look at principal component scores by ethnicity which, in the MCS data, has been found to be a strong predictor of the children's physical activity (Griffiths et al., 2013). This type of analysis, of course, would pose additional methodological issues. The extension of our methods to the full MCS sample should take into account its complex design and, where appropriate, the correlation between data from siblings.

4. Further extensions

The probabilistic formulation of the PCA provides the means to work with different distributions in connection to different norms. For example, as a robust alternative to the Gaussian specification of the error term in model (1), one could consider the Laplacian error's law $f(a) = \frac{1}{2\psi} e^{-|a|/\psi}$, $\psi \in \mathbb{R}_+$. If we maintain the normality assumption for $\mathbf{u}$, the marginal log-likelihood of $\mathbf{y}$ will be given by the log-likelihood of a multivariate normal-Laplace convolution, that is

$$\begin{aligned} l(\boldsymbol{\theta}; Y) &\equiv \sum_{i=1}^{n} \log \int_{\mathbb{R}^q} f(\mathbf{y}_i | \mathbf{u}_i, \boldsymbol{\theta}) f(\mathbf{u}_i) \, d\mathbf{u}_i \\ &= \sum_{i=1}^{n} \log \int_{\mathbb{R}^q} \left(\frac{1}{2\psi}\right)^p e^{-\frac{1}{\psi}\|\mathbf{y}_i - \boldsymbol{\mu} - W\mathbf{u}_i\|_1} 2\pi^{-\frac{q}{2}} e^{-\frac{1}{2}\|\mathbf{u}_i\|_2^2} d\mathbf{u}_i. \end{aligned}$$

It is easy to show that the above expression is equivalent to the log-likelihood of a median mixed model (Geraci and Bottai, 2014) in which the random effects are associated with the p columns of Y and the clustering variable is defined according to the row index. By imposing

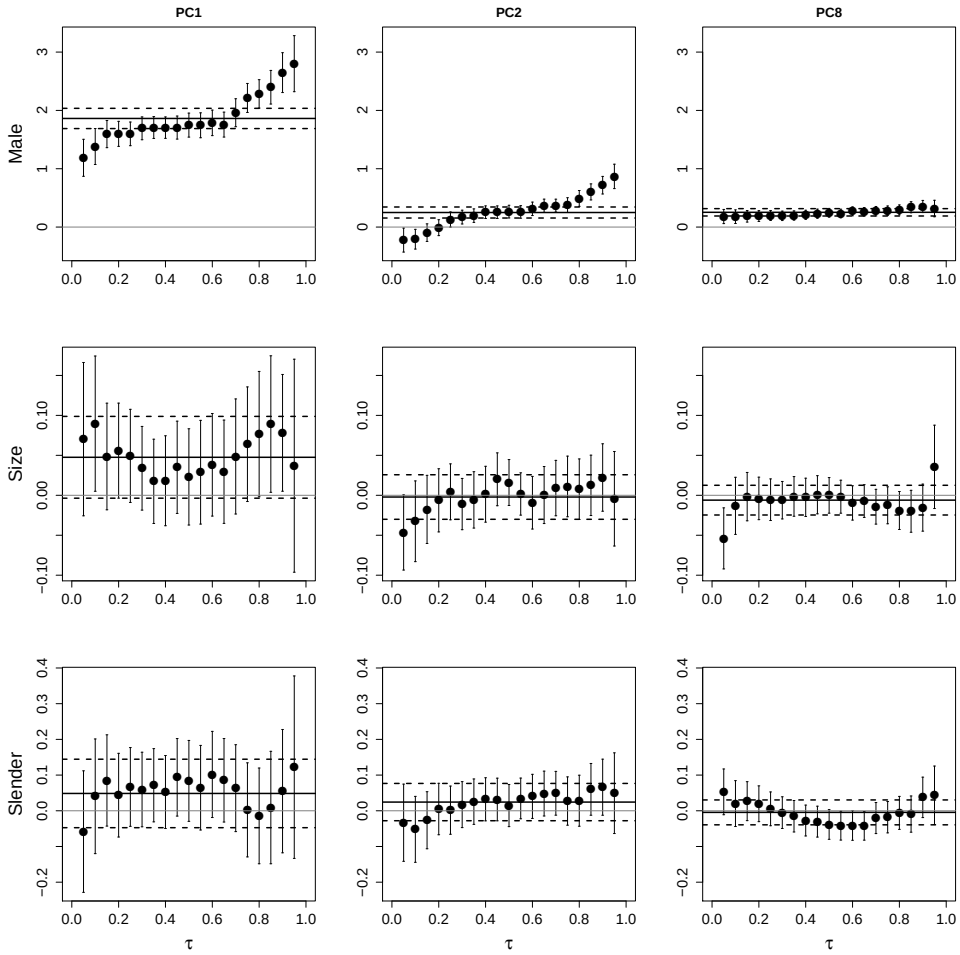

Fig. 7. Estimated demographic and anthropometric quantile effects on individual scores from the Millennium Cohort Study data for the first, second and eighth principal components with 95% confidence intervals (error bars). The ordinary least squares effect (solid lines) and its 95% confidence interval (dashed lines) are superimposed.

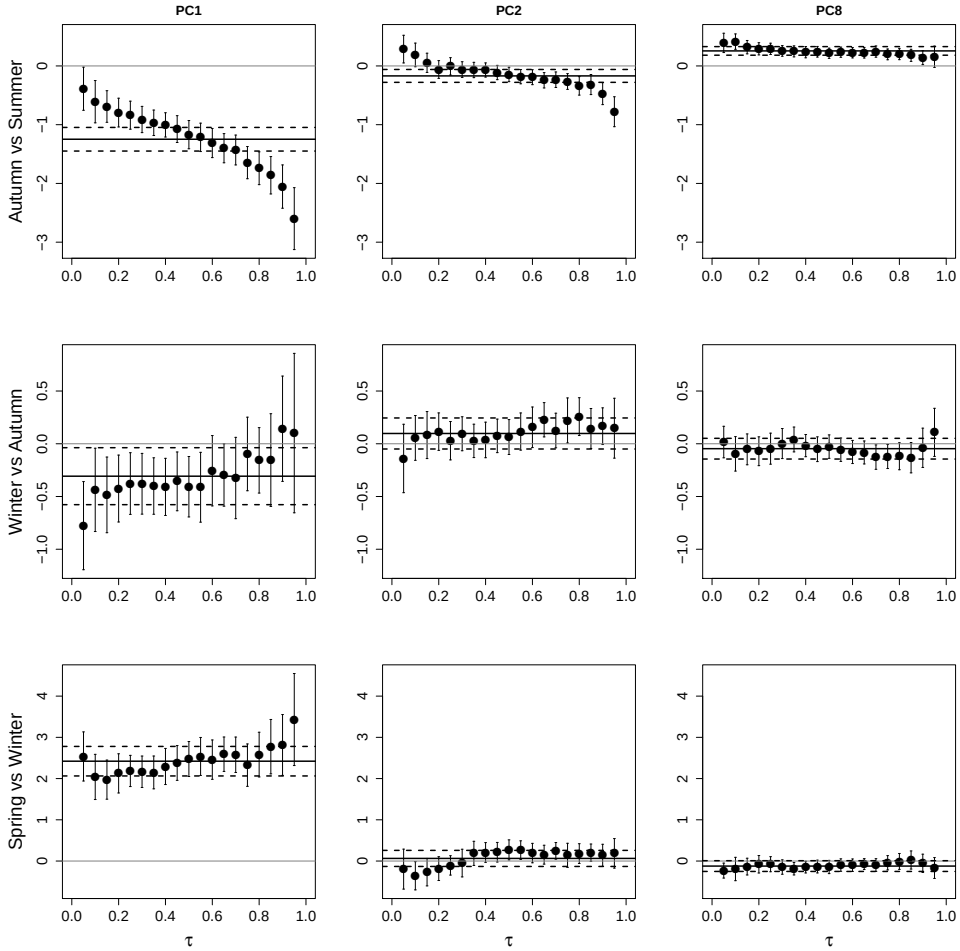

Fig. 8. Estimated seasonal quantile effects on individual scores from the Millennium Cohort Study data for the first, second and eighth principal components with 95% confidence intervals (error bars). The ordinary least squares effect (solid lines) and its 95% confidence interval (dashed lines) are superimposed.

a symmetric covariance matrix for the random effects, say D, it is straightforward to estimate $W \equiv D^{1/2}$ using existing software (Geraci, 2014a,b). Unfortunately, being estimation based on numerical integration (Geraci and Bottai, 2014), this strategy becomes computationally prohibitive for, say, $p > 4$ which obviously is somewhat unattractive given that the goal of PCA is to deal with high-dimensional data. A computational viable alternative would consist in approximating the L_1-norm function using, for example, a Huber M-estimator (Li and Swetits, 1998). This would allow reformulating the minimisation problem as a quadratic programming problem (Mangasarian and Musicant, 2000) that can be solved using efficient algorithms.

The idea of using the Laplace distribution in a probabilistic framework for PCA can be traced back to Baccini et al. (1996). However, in their setting the principal components $\mathbf{u}_i$ were introduced as fixed effects. The main advantage of the L_1-norm variant of PCA is its robustness to outliers (see for example Section 10.4 in Jolliffe, 2002, for an overview on robust PCA). However, this approach is not rotational invariant. See Ding et al. (2006) for an extension of the Laplace distribution.

5. Concluding remarks

The London 2012 Olympic Games and related initiatives such as the School Games competitions have created a unique opportunity in Britain to inspire today's children and young people (Department for Culture, Media and Sport, 2012) by encouraging participation into sports and by nurturing a sporting culture for generations to come. This also offers a timely opportunity to increase the efforts of the scientific community towards the understanding of why and how children develop different physical activity behaviours. In the UK as well as in many other countries, large-scale surveys are contributing to build precious data resources with impressive amounts of accelerometer measurements which can often be linked to databases with demographic, socioeconomic, cognitive and anthropometric variables. Yet, physical activity studies that aim at obtaining representative samples of behaviours in free-living conditions are cursed by noncompliance to study protocol which results in major data loss. More importantly, the risk of introducing bias in the analysis as a consequence of informative missingness is very real. This paper offers a novel and elegant approach to the analysis of multivariate outcomes with nonignorable missing data as well as a step forward to the understanding of physical activity and inactivity behaviours.

Acknowledgements

The work of the first author was supported in part by funding received for the MRC Centre of Epidemiology for Child Health (grant G0400546) while at University College London. Accelerometer data collection was funded by the Wellcome Trust (grant 084686/Z/08/A). We thank two anonymous referees for helpful comments on the manuscript.

Appendix A

In this Appendix, we derive a simplified E-step for the Monte Carlo EM algorithm (Section 2.4) where the random effects are integrated out from the complete data log-likelihood. Indeed, the Gaussian model provides scope for reductions in computational time. In particular, by

integrating the random effects out from (6), the E-step for the ith subject can be written as

$$\begin{aligned}
Q_i(\boldsymbol{\lambda}|\boldsymbol{\lambda}^{(t)}) &= \iint \left\{\log f\left(\mathbf{y}_i, \mathbf{u}_i|\boldsymbol{\theta}\right) + \log f\left(\mathbf{m}_i|\mathbf{y}_i, \boldsymbol{\eta}\right)\right\} f\left(\mathbf{z}_i, \mathbf{u}_i|\mathbf{x}_i, \mathbf{m}_i, \boldsymbol{\lambda}^{(t)}\right) d\mathbf{u}_i d\mathbf{z}_i \\
&= \iint \left\{\log f\left(\mathbf{y}_i|\mathbf{u}_i, \boldsymbol{\theta}\right) + \log f\left(\mathbf{u}_i\right)\right\} f\left(\mathbf{z}_i, \mathbf{u}_i|\mathbf{x}_i, \mathbf{m}_i, \boldsymbol{\lambda}^{(t)}\right) d\mathbf{u}_i d\mathbf{z}_i \\
&\quad + \int \log f\left(\mathbf{m}_i|\mathbf{y}_i, \boldsymbol{\eta}\right) f\left(\mathbf{z}_i|\mathbf{x}_i, \mathbf{m}_i, \boldsymbol{\lambda}^{(t)}\right) d\mathbf{z}_i \\
&= \int \left\{-\frac{p}{2}\log(\psi) - \frac{1}{2\psi}\operatorname{tr}\left(W^\top W B^{(t)}\right) - \frac{1}{2\psi}\|\mathbf{y}_i - \boldsymbol{\mu} - W\mathbf{v}_i^{(t)}\|_2^2 \right. \\
&\quad \left. + \log f\left(\mathbf{m}_i|\mathbf{y}_i, \boldsymbol{\eta}\right)\right\} f\left(\mathbf{z}_i|\mathbf{x}_i, \mathbf{m}_i, \boldsymbol{\lambda}^{(t)}\right) d\mathbf{z}_i \\
&\equiv \mathrm{E}_{\mathbf{z}|\mathbf{x},\mathbf{m},\boldsymbol{\lambda}^{(t)}}\left\{l\left(\boldsymbol{\lambda}; \mathbf{y}_i, \mathbf{m}_i\right)\right\} \qquad (15)
\end{aligned}$$

where $\mathbf{v}_i^{(t)} = B^{(t)}W^{(t)^\top}\left(\mathbf{y}_i - \boldsymbol{\mu}^{(t)}\right)/\psi^{(t)}$ and $B^{(t)} = \left\{W^{(t)^\top}W^{(t)}/\psi^{(t)} + I_q\right\}^{-1}$. Note that by assumption $\mathbf{m}_i$ is independent from $\mathbf{u}_i$. The expectation above is now taken with respect to

$$f\left(\mathbf{z}_i|\mathbf{x}_i, \mathbf{m}_i, \boldsymbol{\lambda}^{(t)}\right) \propto e^{-\frac{1}{2}\left(\mathbf{y}_i - \boldsymbol{\mu}^{(t)}\right)^\top C^{(t)^{-1}}\left(\mathbf{y}_i - \boldsymbol{\mu}^{(t)}\right)} f\left(\mathbf{m}_i|\mathbf{y}_i, \boldsymbol{\eta}^{(t)}\right),$$

$C^{(t)} = W^{(t)}W^{(t)^\top} + \Psi^{(t)}$.

Again, we obtain a sample $\tilde{\mathbf{z}}_{ik}$, $i = 1, \ldots, n$, $k = 1, \ldots, K$ and calculate the approximate E-step

$$Q(\boldsymbol{\lambda}|\boldsymbol{\lambda}^{(t)}) = \frac{1}{K}\sum_{i=1}^{n}\sum_{k=1}^{K} l\left(\boldsymbol{\lambda}; \tilde{\mathbf{z}}_{ik}, \mathbf{x}_i, \mathbf{m}_i\right). \qquad (16)$$

The MLE equations of the M-step which follow from maximising the log-likelihood in (16) are similar to equations (27) and (28) in Tipping and Bishop (1999) and they do not require explicit computation of the covariance matrix. We omit them for the sake of brevity.

We conclude this section by noting that, based on the linear predictions

$$\hat{\mathbf{u}}_{ik} = \left(\hat{W}^\top\hat{W} + \hat{\Psi}\right)^{-1}\hat{W}^\top(\tilde{\mathbf{y}}_{ik} - \hat{\boldsymbol{\mu}}), \qquad (17)$$

where $\tilde{\mathbf{y}}_{ik} = (\tilde{\mathbf{z}}_{ik}, \mathbf{x}_i)$ is the complete data vector at convergence, we can calculate the element-wise variances of $\frac{1}{K}\sum_{k=1}^{K}\hat{\mathbf{u}}_{ik}$ over the individuals space as estimates of $\delta_1, \ldots, \delta_q$. The quantity $(p-q)\cdot\hat{\psi}$ provides the portion of the total variability associated with the 'discarded' components.

Appendix B

We ran a Monte Carlo simulation to assess the performance of the proposed approach. As basis of our simulation, we considered the subset of 1640 MCS children who had valid data for all days of the week (Table 2). Using this subset, we estimated the PPCA parameter $\boldsymbol{\theta}$, which we now define to be the 'true' parameter $\boldsymbol{\theta}_0 = (\boldsymbol{\mu}_0, W_0, \psi_0)$ for our simulation purposes.

Let $y_{i1}^{(d)}$, $y_{i2}^{(d)}$, and $y_{i3}^{(d)}$ be, respectively, the total counts, proportion of sedentary time and of MVPA time for the ith child on the dth day, $i = 1, \ldots, 1640$ and $d = 1, \ldots, 7$. At each Monte Carlo replication, we removed the physical activity outcomes measured on a given day d from a random sample of children; the latter was drawn with probability equal to

$$\left[1 + \exp\left\{-\left(\eta_0 + \eta_1\bar{y}_{i1} + \eta_2\bar{y}_{i2} + \eta_3\bar{y}_{i3}\right)\right\}\right]^{-1},$$

Table 8. Simulation study results for the probabilistic principal component analysis (PPCA) with missing data: Performance of complete-case (CC), single imputation (SI) and nonignorable (NI) PPCA under missing-completely-at-random (MCAR) or missing-not-at-random (MNAR) mechanisms.

	Distance of $\hat{W}$ from W_0		*Distance of $\hat{\psi}$ from ψ_0*	
	MCAR data	*MNAR data*	*MCAR data*	*MNAR data*
CC	0.42	0.35	0.007	0.014
SI	0.62	0.65	0.113	0.110
NI	0.39	0.28	0.005	0.003

where $\bar{y}_{i1} = \frac{1}{7}\sum_{d=1}^{7} y_{i1}^{(d)}$, $\bar{y}_{i2} = \frac{1}{7}\sum_{d=1}^{7} y_{i2}^{(d)}$, and $\bar{y}_{i3} = \frac{1}{7}\sum_{j=d}^{7} y_{i3}^{(d)}$ are, respectively, the total counts, proportion of sedentary time and of MVPA time averaged over the week.

We considered ignorable and nonignorable mechanisms by setting $\boldsymbol{\eta} = (-1.5, 0, 0, 0)'$ and $\boldsymbol{\eta} = (-2, 0.5, -1, 1)'$, respectively. Note that the former model produces data missing completely at random (MCAR). The resulting overall proportions of missing data were 18% and 15% in the ignorable and nonignorable scenarios, respectively.

For each simulated dataset, we conducted three analyses: a complete-case PPCA; a PPCA applied to a singly imputed dataset, the latter obtained using the EM-based algorithm by Josse and Husson (2012a) and available in the `R` package `missMDA` (Husson and Josse, 2013); and our proposed PPCA with MDM as specified above. Note that the first two approaches are appropriate for, respectively, MCAR and MAR data. We report the average distance (after 100 replicates for each setting) between the 'true' W_0 and the estimate $\hat{W}$ using a measure along the lines of (14), where axes are taken in absolute value to avoid orientation issues; and the average Euclidean distance between ψ_0 and the estimate $\hat{\psi}$.

Among the three methods, the single imputation approach had the least favourable performance when estimating the parameters of interest W_0 and ψ_0 for either of the two scenarios (Table 8). The nonignorable approach had a slight advantage over the complete-case in the MCAR scenario and, as expected, it became more competitive in the MNAR scenario.

References

Baccini, A., P. Besse, and A. de Falguerolles (1996). A L_1-norm PCA and a heuristic approach. In E. Diday, Y. Lechevalier, and O. Optiz (Eds.), *Ordinal and Symbolic Data Analysis*. Berlin, Germany: Springer-Verlag.

Biddle, S. J. H., T. Gorely, S. J. Marshall, I. Murdey, and N. Cameron (2004). Physical activity and sedentary behaviours in youth: Issues and controversies. *Journal of the Royal Society for the Promotion of Health 124*(1), 29–33.

Biddle, S. J. H., T. Gorely, and D. J. Stensel (2004). Health-enhancing physical activity and sedentary behaviour in children and adolescents. *Journal of Sports Sciences 22*(8), 679–701.

Booth, J. G. and J. P. Hobert (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society, B 61*, 265–285.

Bull, F. C. and the Expert Working Groups (2010). Physical activity guidelines in the UK: Re-

view and recommendations. Technical report, School of Sport, Exercise and Health Sciences. Loughborough University.

Catellier, D. J., P. J. Hannan, D. M. Murray, C. L. Addy, T. L. Conway, S. Yang, and J. C. Rice (2005). Imputation of missing data when measuring physical activity by accelerometry. *Medicine and Science in Sports and Exercise 37*(11), S555–S562.

Cole, T. J., M. C. Bellizzi, K. M. Flegal, and W. H. Dietz (2000). Establishing a standard definition for child overweight and obesity worldwide: International survey. *BMJ 320*(7244), 1240–1243.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B 39*(1), 1–38.

Department for Culture, Media and Sport (2002). Game plan: A strategy for delivering government's sport and physical activity objectives. Technical report, Cabinet Office, London, UK.

Department for Culture, Media and Sport (2012). Beyond 2012: The London 2012 legacy story. Technical report, Cabinet Office, London, UK.

Ding, C., D. Zhou, X. He, and H. Zha (2006). R_1-PCA: Rotational invariant L_1-norm principal component analysis for robust subspace factorization. In *Proceedings of the 23rd international conference on Machine learning*, pp. 281–288. ACM.

Drenowatz, C., S. Kobel, S. Kettner, D. Kesztys, and J. M. Steinacker (2012). Interaction of sedentary behaviour, sports participation and fitness with weight status in elementary school children. *European Journal of Sport Science 14*(1), 100–105.

Eberth, B. and M. D. Smith (2010). Modelling the participation decision and duration of sporting activity in scotland. *Economic Modelling 27*(4), 822–834.

Ekelund, U., S. Brage, K. Froberg, M. Harro, S. A. Anderssen, L. B. Sardinha, C. Riddoch, and L. B. Andersen (2006). TV viewing and physical activity are independently associated with metabolic risk in children: The European Youth Heart Study. *PLoS Medicine 3*(12), e488.

Esliger, D. W. and M. S. Tremblay (2007). Physical activity and inactivity profiling: The next generation. *Canadian Journal of Public Health. Revue Canadienne de Sante Publique 98 Suppl 2*, S195–207.

Geraci, M. (2014a). Linear quantile mixed models: The lqmm package for Laplace quantile regression. *Journal of Statistical Software 57*(13), 1–29.

Geraci, M. (2014b). lqmm: Linear quantile mixed models. R package version 1.5.

Geraci, M. (2014c). pawacc: Physical activity with accelerometers. R package version 1.2.1.

Geraci, M. and M. Bottai (2014). Linear quantile mixed models. *Statistics and Computing 24*(3), 461–479.

Geraci, M., C. Rich, F. Sera, M. Cortina-Borja, L. J. Griffiths, and C. Dezateux (2012). Technical report on accelerometry data processing in the Millennium Cohort Study. Technical report, University College London.

Gilks, W. R. and P. Wild (1992). Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistical Society, C 41*(2), 337–348.

Goetghebeur, E. and L. Ryan (2000). Semiparametric regression analysis of interval-censored data. *Biometrics 56*(4), 1139–1144.

Griffiths, L. J., M. Cortina-Borja, F. Sera, T. Pouliou, M. Geraci, C. Rich, T. J. Cole, C. Law, H. Joshi, A. R. Ness, S. A. Jebb, and C. Dezateux (2013). How active are our children? Findings from the Millennium Cohort Study. *BMJ Open 3*(8).

Griffiths, L. J., C. Rich, M. Geraci, F. Sera, M. Cortina-Borja, T. Pouliou, L. Platt, J. Johnson, and C. Dezateux (2013). Technical report on the enhancement of Millennium Cohort Study data with accelerometer-derived measures of physical activity and sedentary behaviour in seven year olds. Technical report, Centre for Longitudinal Studies.

Hamer, M., E. Stamatakis, and G. Mishra (2009). Psychological distress, television viewing, and physical activity in children aged 4 to 12 years. *Pediatrics 123*(5), 1263–1268.

Heitzler, C., L. Lytle, D. Erickson, J. Sirard, D. Barr-Anderson, and M. Story (2011). Physical activity and sedentary activity patterns among children and adolescents: A latent class analysis approach. *Journal of Physical Activity and Health 8*(4), 457.

Husson, F. and J. Josse (2013). missMDA: Handling missing values with/in multivariate data analysis (principal component methods). R package version 1.7.2.

Ibrahim, J. G., M. H. Chen, and S. R. Lipsitz (2001). Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. *Biometrika 88*(2), 551–564.

Ibrahim, J. G. and G. Molenberghs (2009). Missing data methods in longitudinal studies: A review. *TEST 18*(1), 1–43.

Ibrahim, J. G., H. T. Zhu, and N. S. Tang (2008). Model selection criteria for missing-data problems using the EM algorithm. *Journal of the American Statistical Association 103*(484), 1648–1658.

Ilin, A. and T. Raiko (2010). Practical approaches to principal component analysis in the presence of missing values. *Journal of Machine Learning Research 11*, 1957–2000.

Jolliffe, I. T. (2002). *Principal component analysis* (Second ed.). New York, NY: Springer-Verlag.

Josse, J. and F. Husson (2012a). Handling missing values in exploratory multivariate data analysis methods. *Journal de la Société Française de Statistique 153*(2), 79–99.

Josse, J. and F. Husson (2012b). Selecting the number of components in principal component analysis using cross-validation approximations. *Computational Statistics and Data Analysis 56*(6), 1869–1879.

Koenker, R. and G. Bassett (1978). Regression quantiles. *Econometrica 46*, 33–50.

Koorts, H., C. Mattocks, A. R. Ness, K. Deere, S. N. Blair, R. R. Pate, and C. Riddoch (2011). The association between the type, context, and levels of physical activity amongst adolescents. *Journal of Physical Activity and Health 8*(8), 1057–1065.

Li, W. and J. Swetits (1998). The linear l_1 estimator and the Huber M-estimator. *SIAM Journal on Optimization 8*(2), 457–475.

Little, R. and D. Rubin (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley.

Mangasarian, O. L. and D. R. Musicant (2000). Robust linear and support vector regression. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 22*(9), 950–955.

Marshall, S. J., S. J. H. Biddle, J. F. Sallis, T. L. McKenzie, and T. L. Conway (2002). Clustering of sedentary behaviors and physical activity among youth: A cross-national study. *Pediatric Exercise Science 14*(4), 401–417.

McLachlan, G. and T. Krishnan (2008). *The EM algorithm and extensions.* Hoboken, New Jersey: John Wiley & Sons.

Mitchell, J. A., C. Mattocks, A. R. Ness, S. D. Leary, R. R. Pate, M. Dowda, S. N. Blair, and C. Riddoch (2009). Sedentary behavior and obesity in a large cohort of children. *Obesity 17*(8), 1596–1602.

Morris, J. S., C. Arroyo, B. A. Coull, L. M. Ryan, R. Herrick, and S. L. Gortmaker (2006). Using wavelet-based functional mixed models to characterize population heterogeneity in accelerometer profiles: A case study. *Journal of the American Statistical Association 101*(476), 1352–1364.

National Institute for Health and Clinical Excellence (2008). National costing report: Physical activity and the environment. Technical report, NICE, London, UK.

National Institute for Health and Clinical Excellence (2009). Promoting physical activity, active play and sport for pre-school and school-age children and young people in family, pre-school, school and community settings. Technical report, NICE, London, UK.

Nyamundanda, G., L. Brennan, and I. Gormley (2010). Probabilistic principal component analysis for metabolomic data. *BMC Bioinformatics 11*(1), 571.

Nyamundanda, G., I. C. Gormley, and L. Brennan (2014). A dynamic probabilistic principal components model for the analysis of longitudinal metabolomics data. *Journal of the Royal Statistical Society, C*.

Oksanen, J., F. G. Blanchet, R. Kindt, P. Legendre, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, and H. Wagner (2013). *vegan: Community ecology package.* R package version 2.0-10.

Pearson, K. (1901). LIII. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2*(11), 559–572.

Petris, G. and L. Tardella (2013). HI: Simulation from distributions supported by nested hyperplanes. R package version 0.4.

Prentice-Dunn, H. and S. Prentice-Dunn (2011). Physical activity, sedentary behavior, and childhood obesity: A review of cross-sectional studies. *Psychology, Health & Medicine 17*(3), 255–273.

Pulsford, R. M., M. Cortina-Borja, C. Rich, F. E. Kinnafick, C. Dezateux, and L. J. Griffiths (2011). Actigraph accelerometer-defined boundaries for sedentary behaviour and physical activity intensities in 7 year old children. *PLOS ONE 6*(8), e21822.

R Development Core Team (2014). R: A language and environment for statistical computing.

Reilly, J. J., D. M. Jackson, C. Montgomery, L. A. Kelly, C. Slater, S. Grant, and J. Y. Paton (2004). Total energy expenditure and physical activity in young Scottish children: Mixed longitudinal study. *Lancet 363*(9404), 211–2.

Rich, C., M. Cortina-Borja, C. Dezateux, M. Geraci, F. Sera, L. Calderwood, H. Joshi, and L. J. Griffiths (2013). Predictors of non-response in a UK-wide cohort study of children's accelerometer-determined physical activity using postal methods. *BMJ Open 3*(3), e002290.

Rich, C., M. Geraci, L. J. Griffiths, F. Sera, C. Dezateux, and M. Cortina-Borja (2013). Quality control methods in accelerometer data processing: Defining minimum wear time. *PLoS ONE 8*(6), e67206.

Rich, C., M. Geraci, L. J. Griffiths, F. Sera, C. Dezateux, and M. Cortina-Borja (2014). Quality control methods in accelerometer data processing: Identifying extreme counts. *PLoS ONE 9*(1), e85134.

Robert, P. and Y. Escoufier (1976). A unifying tool for linear multivariate statistical methods: The RV-coefficient. *Applied Statistics 25*(3), 257–265.

Roweis, S. (1998). EM algorithms for PCA and SPCA. In M. I. Jordan, M. J. Kearns, and S. A. Solla (Eds.), *Advances in neural information processing systems 10: Proceedings of the 1997 conference*, Volume 10, pp. 626–632. Cambridge, MA: MIT Press.

Sera, F., M. Cortina-Borja, M. Geraci, L. J. Griffiths, C. Rich, and C. Dezateux (2011). Modelling accelerometer data from 7-year old British children using functional analysis of variance. *Journal of Epidemiology and Community Health 65 (Suppl 2)*, A26–A27.

Smith, K. and H. Joshi (2002). The Millennium Cohort Study. *Population Trends 107*, 30–34.

Staudenmayer, J., W. Zhu, and D. Catellier (2012). Statistical considerations in the analysis of accelerometry-based activity monitor data. *Medicine and Science in Sports and Exercise 44*(1S), S61–S67.

Tipping, M. E. and C. M. Bishop (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society, B 61*(3), 611–622.

Warburton, D. E., C. W. Nicol, and S. S. Bredin (2006). Health benefits of physical activity: The evidence. *Canadian Medical Association Journal 174*(6), 801–809.