

A probabilistic approach to the estimation of principal components with nonignorable missing data: Applications in accelerometer-based physical activity studies

Un approccio probabilistico per la stima di componenti principali con dati mancanti non-ignorabili: Applicazioni per l'analisi di attività fisica con dati di accelerometro

Marco Geraci e Alessio Farcomeni

Abstract Our paper is motivated by an accelerometer-based study of physical activity (PA) behaviours in a large cohort of UK school-aged children. Given the multiplicity of dimensions associated with PA, a comprehensive individual PA profile can include more than forty different outcomes which clearly poses challenges for the analysis. We explore the application of principal component analysis (PCA) to accelerometer measurements which are aggregated daily over several days of the week and are affected by nonresponse. The probabilistic approach to PCA with latent scores (Tipping and Bishop, 1999, *JRSS-B*) is extended to include nonignorable missing data. The extended likelihood is maximized through a Monte Carlo EM algorithm via adaptive rejection Metropolis sampling.

Abstract Il nostro articolo è motivato da uno studio sulla attività fisica (AF) in un'ampia coorte di bambini del Regno Unito in età scolare basato su dati di accelerometro. Dal momento che l'AF è caratterizzata da una molteplicità di dimensioni, un profilo individuale di AF può includere più di quaranta variabili che chiaramente rendono l'analisi difficoltosa. Esploriamo l'applicazione dell'analisi delle componenti principali (PCA) su dati di accelerometro aggregati per giorni della settimana ed affetti da mancata risposta. L'approccio probabilistico alla PCA con effetti latenti (Tipping and Bishop, 1999, *JRSS-B*) viene esteso al caso di dati mancanti non-ignorabili. La verosimiglianza estesa è massimizzata tramite un algoritmo Monte Carlo EM basato su campionamento 'adaptive rejection Metropolis'.

Key words: body mass index, EM algorithm, obesity, parallel computing, random effects

Marco Geraci

Centre for Paediatric Epidemiology and Biostatistics, Institute of Child Health, University College London, United Kingdom. e-mail: m.geraci@ucl.ac.uk

Alessio Farcomeni

Dipartimento di Sanità Pubblica e Malattie Infettive, Sapienza – Università di Roma, Italy. e-mail: alessio.farcomeni@uniroma1.it

1 Background

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 [11]. 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 [9]. It is also estimated that 54,000 premature deaths a year are linked to a sedentary lifestyle [2].

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. Physical activity and inactivity seem to be two dimensions over which adolescents aggregate into distinct clusters [1, 8] and these clusters can be identified and characterized by demographic, anthropometric and lifestyle factors [5].

The analysis of accelerometer measurements is often hindered by nonresponse and noncompliance. For example, missing data may arise when the monitor is not waterproof and this is taken off during aquatic activities. In general, if the missingness is suspected to be nonignorable, the missing data mechanism (MDM) needs to be modelled.

We develop a probabilistic model for principal component analysis (PCA) [10] to characterise physical activity and inactivity behaviours using accelerometer data. The model is extended to include nonignorable missing data. Estimation is carried out using a Monte Carlo EM algorithm via adaptive rejections Metropolis sampling (ARMS).

2 Data and Methods

We analyse data collected for an accelerometer-based study on PA determinants in children of the UK Millennium Cohort Study (MCS) [4]. Children were asked to wear an accelerometer for seven consecutive days, but to remove it during aquatic activities as the accelerometers are not waterproof. The dataset included measurements for 5682 children, of which only 1640 fully complied to the study protocol by providing data for every day of the week.

Accelerometers produce an output known as ‘acceleration counts’ which are dimensionless and thus requires calibration in order to be converted into physiologically more appropriate units (e.g., metabolic equivalent of task). PA is usually classified into categories of varying levels of intensity, with sedentary behaviour on one end of the spectrum, followed by light, moderate and vigorous activity.

Currently a comprehensive individual PA profile can include more than 40 different outcomes [3] which clearly poses challenges for the analysis. Table 1 illustrates an example of a PA profile for one child of the MCS 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 PA 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 with seven outcome variables for one child of the Millennium Cohort Study.

		<i>Mon</i>	<i>Tue</i>	<i>...</i>	<i>Sun</i>
<i>Total activity</i>					
(1)	counts ($\times 1000$)	287.7	564.7	...	305.7
(2)	steps ($\times 1000$)	7.4	13.3	...	6.5
<i>Proportion of time (%) spent</i>					
(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
<i>Sporadic (<10 minutes) moderate-to-vigorous bouts</i>					
(6)	total time (minutes)	37.5	92.75	...	38
(7)	frequency	91	169	...	96

Let $\mathbf{y}_i = (y_{i1}, \dots, y_{ip})^\top$ be the $p \times 1$ random vector collecting the variables of interest for the i th subject, $i = 1, \dots, n$. The probabilistic representation of PCA [10] is given by the model

$$\mathbf{y}_i = \boldsymbol{\mu} + \mathbf{W}\mathbf{u}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \dots, n, \quad (1)$$

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

Suppose that $\mathbf{y}_i$ contains s_i , $s_i < p$, missing values. Let us denote with $\mathbf{m}_i$, $i = 1, \dots, n$, the $p \times 1$ missing data indicator whose j th element m_{ij} equals 1 if y_{ij} is missing and 0 otherwise. 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 i th contribution to the complete data density of $(\mathbf{y}_i, \mathbf{u}_i, \mathbf{m}_i)$ is given by

$$f(\mathbf{y}_i, \mathbf{u}_i, \mathbf{m}_i | \boldsymbol{\theta}, \boldsymbol{\eta}) = f(\mathbf{y}_i | \mathbf{u}_i, \boldsymbol{\theta}) f(\mathbf{u}_i) f(\mathbf{m}_i | \mathbf{y}_i, \boldsymbol{\eta}), \quad (2)$$

where the likelihood based on model (1) now includes the additional factor $f(\mathbf{m}_i|\mathbf{y}_i, \eta)$, indexed by the parameter η . This is the MDM [7] which we assume to be independent from $\mathbf{u}_i$. The latter assumption can be relaxed, although at the expense of computational complexity (see Ibrahim *et al.*, 2001, for a discussion on the rationale of such assumption and alternative modelling strategies). In general, one cannot test the parametric form of the MDM itself from the data, although the results of the analysis can be used for sensitivity purposes [6]. Finally, care must be taken to avoid building too large models of the MDM that can lead to non-identifiability of the parameters.

Estimation of θ would in general require marginalizing the log-likelihood based on (2) over the unobserved data, which however leads to a rather intractable integral. Instead, we develop an EM algorithm where the E-step is approximated using a Monte Carlo technique via ARMS. 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, therefore greatly speeding up the computation.

3 Results

We developed a novel likelihood-based approach to the estimation of principal components with nonignorable missing data. The results obtained from the application of the proposed methods to the MCS data show that it is important to account for the MDM and support the hypothesis of informative nonresponse. Computation may be slow with large datasets. However, parallel computing substantially reduces computing time.

References

1. 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. *J. R. Soc. Promot. Health* 124(1), 29–33.
2. Department for Culture, Media and Sport (2002). Game plan: A strategy for delivering government’s sport and physical activity objectives. Technical report, Cabinet Office.
3. Esliger, D. W. and M. S. Tremblay (2007). Physical activity and inactivity profiling: The next generation. *Can. J. Public Health* 98 Suppl 2, S195–207.
4. Griffiths, L. J., C. Rich, M. Geraci, F. Sera, M. Cortina-Borja, T. Poulou, L. Platt, J. Johnson, and C. Dezaux (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.
5. 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. *J. Phys. Act. Health* 8(4), 457.
6. 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.

7. Little, R. and D. Rubin (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley.
8. 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. *Pediatr. Exerc. Sci.* 14(4), 401–417.
9. National Institute for Health and Clinical Excellence (2008). National costing report: Physical activity and the environment. Technical report.
10. Tipping, M. E. and C. M. Bishop (1999). Probabilistic principal component analysis. *J. Roy. Statist. Soc. Ser. B* 61(3), 611–622.
11. Warburton, D. E., C. W. Nicol, and S. S. Bredin (2006). Health benefits of physical activity: The evidence. *Can. Med. Assoc. J.* 174(6), 801–809.