Taylor & Francis
Taylor & Francis Group

Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: http://www.tandfonline.com/loi/cjas20

Functional principal component analysis in age–period–cohort analysis of body mass index data by gender and ethnicity

Jun Ye, Juan Xi & Richard L. Einsporn

To cite this article: Jun Ye, Juan Xi & Richard L. Einsporn (2017): Functional principal component analysis in age–period–cohort analysis of body mass index data by gender and ethnicity, Journal of Applied Statistics, DOI: 10.1080/02664763.2017.1322557

To link to this article: http://dx.doi.org/10.1080/02664763.2017.1322557

Published online: 04 May 2017.

JOURNAL OF APPLIED STATISTICS, 2017
https://doi.org/10.1080/02664763.2017.1322557

Taylor & Francis
Taylor & Francis Group

Functional principal component analysis in age–period–cohort analysis of body mass index data by gender and ethnicity

Jun Ye[a], Juan Xi[b] and Richard L. Einsporn[a]

[a]Department of Statistics, University of Akron, Akron, OH; [b]Department of Sociology, University of Akron, Akron, OH

ABSTRACT
In this paper, we propose a two-stage functional principal component analysis method in age–period–cohort (APC) analysis. The first stage of the method considers the age–period effect with the fitted values treated as an offset; and the second stage of the method considers the residual age–cohort effect conditional on the already estimated age-period effect. An APC version of the model in functional data analysis provides an improved fit to the data, especially when the data are sparse and irregularly spaced. We demonstrate the effectiveness of the proposed method using body mass index data stratified by gender and ethnicity.

ARTICLE HISTORY
Received 10 June 2016
Accepted 16 April 2017

KEYWORDS
B-splines; conditional expectation; longitudinal data; mixed-effects model; random effects

1. Introduction

1.1. Literature review

The dramatic increase in the prevalence of overweight and obesity in the US population during the past few decades has drawn much research attention [22]. The observed increasing trend can be considered as the result of temporary variation along three dimensions: the age trend, the secular changes, and the cohort variations. People at different ages have different risk for obesity. Middle-aged adults usually report the highest obesity rates. As baby boomers proceed along the age axis, the changing age structure of the US population might explain part of the observed overall obesity trend. The other well-discussed reason for the increased obesity prevalence is related to diet and lifestyle. The prevalence of fast food and processed food and a diet high in sugar and salt and low in fiber are blamed for contributing to the obesity trend [28]. The birth cohort variation of obesity has also been reported. The Silent Generation (born in 1925–1945) and Generation X (born in 1965–1980) have been found to have higher prevalence of obesity than the baby boomers [26]. These three temporal trends are intertwined with each other and together they shape the observed overall obesity trends. To disentangle them, the age-period-cohort (APC) model has been widely used and discussed by a number of recent methodological innovations. An APC model

CONTACT Jun Ye ✉ jye1@uakron.edu

contains the effects of age groups, periods of observation and birth cohorts. The decomposition of the three factors usually provides a particularly clear summary of longitudinal data; and can well delineate temporal trends and cohort patterns. See [9,11,14,24] for recent methodological innovations to the APC model.

APC analysis is a popular analytic approach in sociological studies allowing for a better understanding of age, period and cohort effects. However, the main issue of concern in the APC analysis is the non-collinearity problem. Due to the perfect linear dependency among factors of age, period and cohort, i.e. period = age + cohort, the APC model including all three factors suffers from the collinearity. Given any of two factors, the third factor can be exactly computed, and all three factors cannot be simultaneously estimated in a linear model [14]. Recently, there have been many statistical approaches to APC models for dealing with the non-collinearity problem. However, different approaches based on different subjective judgements often lead to different estimates [3]. As the analysis of APC problem is not data specific but model specific [24], there is no consensus in the literature as to which method is optimal [3].

Longitudinal data analysis often involves in irregularly spaced and infrequent measurements, resulting in an inherent difficulty in traditional parametric statistical analysis. A flexible nonparametric data analysis approach has advantages for such data. Functional data analysis (FDA) is a data-driven statistical technique that has been widely used in modern quantitative research [23]. The main idea of FDA is smoothing, which allows flexible structure of the effects for the age, period and cohort factors [3]. There are many different approaches of smoothing in the FDA literature; see [3,11,14,24]. The advantages of smoothing are that a smoothing model does not suffer from the non-collinearity problem, and provides more accurate curve estimation for the nonlinear trend changes in the effects.

A principal component analysis (PCA) is concerned with explaining the data structure through a few linear combination of the random variables. The general objective of PCA is data reduction and interpretation; see [15] for more details. There are lots of PCA methods proposed in the APC analysis in the recent years, e.g. Yang *et al.* [29] and Yang and Land [30,31]and Fukuda [4–6]. However, these methods only consider the principal component scores as fixed effects without applying additional FDA technologies to the APC analysis procedure. The main novelty of our work is that we introduce the mixed-effects functional PCA method for the case of sparse and unbalanced data in the APC analysis, where the data are considered as functional and the principal component scores are considered as random effects. For the BMI data by gender and ethnicity, the numbers of points in period and cohort are too small to satisfy the large sample criteria required by the maximum likelihood estimation of variance components. In addition, the observations are irregularly spaced. Hence the errors in variance components could produce extra uncertainty in the estimations. To address these problems, it is useful to apply a method of FDA to produce more accurate estimates.

Functional data have board applicability in many fields. In contrast with traditional statistics, the data in FDA are treated as random functions, e.g. curves. As a particular case of FDA, functional PCA is currently under intense methodological research. Functional PCA refers to a particular method of PCA that is applied to functions instead of vectors, where the functions are different from vectors by the smoothness. In the recent literature, there has been increased interest in functional PCA with mixed effects. James *et al.* [13] proposed a reduced rank mixed-effects model by B-splines smoothing for the sparse data.

Yao *et al.* [32] proposed a conditional expectation method in functional PCA in estimating PC scores for the irregular longitudinal data. The proposed functional PCA approach is flexible, and allows for varying patterns of observations with regard to the measurements of the response functions.

In this paper, we consider the mixed-effects model of PCA as discussed by James *et al.* [13], Ye *et al.* [33] and Zhou *et al.* [34]. The proposed model is cast into a mixed model framework where the random effects approach in the principal component scores is explored. The mixed-effects model in functional PCA leads to predictions of random effects for the principal component scores. By James *et al.* [13], this kind of model has the following advantages: first, it estimates the trajectories using all observed data when there are insufficient data from each individual trajectories; secondly, the method automatically assigns correct weight to each individual trajectories; thirdly, the method allows for individual variation where the principal patterns of variation about the mean curve are referred to as the functional curves.

We advocate a functional PCA by incorporating smoothing splines in the PCA. Functional PCA attempts to characterize the dominant modes of variation of a sample of random trajectories around an overall mean function. As the measurements over age, period and cohort are sparse and irregularly spaced for the individuals, the functional PCA method is well-suited here. The proposed functional PCA provides data-driven estimates of smoothing parameters by a mixed-effects model. To better deal with the linear dependency among associated effects of age, period and cohort, we further propose to use a two-stage functional PCA method for the conditional three-factor APC model [8]. This method includes age-period analysis in the first stage and age–cohort analysis in the second stage, and can well describe the age, period and cohort effects of the data.

Currently, there are no advanced methods that have emphasized the overall trends of age, period and cohort simultaneously by gender and ethnicity [17,18,24,25]. One possible reason could be that the numbers of observations in some gender and ethnicity combinations are not large enough to cover the whole range of the age, period and cohort [18]. Also, the results from parametric methods in APC are not consistent and hence one can have little faith in their validity [17]. This research paper is motivated by the fact that the BMI data by gender and ethnicity are sparse and irregularly spaced with missing observations. Hence, it is necessary to consider them as functional data. Our paper innovates by applying functional PCA with mixed-effects modeling to the analysis procedure, which reflects a deeper data analysis procedure under age, period and cohort.

To the best of our knowledge, the proposed method is the first attempt to use the functional PCA method to deal with the trends of age, period and cohort in APC models. The major contributions of the paper are (1) The use of two-stage functional PCA in the three-factor analysis is innovate. It is noted that the matrix formats are not the same in the different stages, which is age × period in the first stage and age × cohort in the second stage. Hence the proposed method can be considered as a combination of two two-factor analysis problems, which can solve the non-collinearity problem in the APC model. Due to the complexities of the numerical computation in the mixed-effects model, a multi-stage or hierarchical model is usually employed for adequate estimates of the error variations in the random effects of the model [30,31]. Hence, the proposed two-stage approach is more likely appropriate because it is straightforward. (2) The proposed method considers B-splines smoothing in the data analysis, which is more flexible for our data. Moreover, the use of FDA can overcome the non-collinearity problem caused by linear dependency among factors of age, period and cohort. (3) The proposed method considers period and cohort as the random effects for implementing data smoothing, which has shown to have an advantage of dealing with the sparse and irregularly spaced longitudinal data.

1.2. Introduction of the study

The National Health Interview Survey (NHIS) provides us national representative BMI information for US adults overtime. In this study, we focus on NHIS data collected from 1997 to 2013. Because the purpose of the study is to test our proposed APC model, we selected a random subset of 6000 individuals from the original NHIS data which includes millions of observations when we pooled data from 1997 to 2013. After deleting cases with invalid weight and height information, our analytic sample contains 5946 individuals. Body mass index (BMI) is a measure of relative size based on the mass and height of an individual. If the weight is in kilograms and the height in meters, the BMI is measured by the ratio of weight to squared height, i.e. kg/m^2. In this paper, we mainly investigate the trajectories of BMI in the study. To better understand the temporal trends and cohorts, we perform the APC analysis for BMI. The study includes 5946 individuals with age ranging from 0 to 67 years old. Interviews conducted during period 1997 and 2013. Usually the researchers group the age, period and cohort properties into time intervals of different lengths [31]. Because meaningful cohort is often considered to have a duration longer than a single year, it will be feasible to group the cohort into a multi-year period. In the BMI data, the cohort is from 0 to 14, i.e. it is grouped into 5-year intervals by rounding (period − age −1940)/5 to the nearest integer. In the study, we mainly focus on the analysis of the data stratified by gender and ethnicity. Over the 5946 individuals, there are 126 female Asians, 116 male Asians; 483 female Blacks, 328 male Blacks; 588 female Hispanics, 434 male Hispanics; and 2134 female Whites, 1737 male Whites.

Longitudinal data stratified by gender and ethnicity are usually sparse within subgroups. One of the major merits of our proposed model is its capacity of analyzing sparse data. Although pooling groups together will reduce the sparseness of the data, there are reasons that we are cautious about pooling ethnic groups together. First, previous research has established significant differences in the prevalence of overweight and obesity across ethnic groups [21]. Asians were found repeatedly to have a lower average BMI than other ethnic groups [18]. Black females had the highest level of prevalence of obesity [21] and those from Hispanic origin had the fastest rate of BMI increase over the past two decades [18]. Secondly, weight status is closely linked to diet, culture and lifestyles. Difference ethnic groups were influenced by different culture and ethnic life styles. Due to the large scale of immigration from Asia and Latin America over the past few decades, the increasing supply of ethnic goods for Asian and Hispanics might impact the periodical trends for these two ethnic groups particularly. Hence, we are reluctant to pool ethnic groups together and assume similar trends across ethnic groups. Although NHIS uses nationally representative samples, with pooled data, the dominant ethnic group will have greater influence on the detected results and the pooled results will not be representative to ethnic groups of smaller size.

We develop functional PCA for the analysis of individual trajectories from sparse and irregular observations, and aim at a flexible nonparametric FDA approach. It is necessary to consider the observations as functional data when the number of observations is not large enough to cover the whole range of the age, period and cohort.

Table 1 and 2 show the numbers of individuals, the related unique observations and missing observations under the matrices of age × period (i.e. 68 × 17) and age × cohort (i.e. 68 × 15) by gender and ethnicity. It is noted that sum of the number of unique observations and the number of missing observations is 1156 under age × period and is 1020

Table 1. The numbers of individuals, the related unique observations and missing observations under the matrix of age × period (i.e., 68 × 17) by gender and ethnicity.

	Individuals			Unique observations			Missing observations		
	Females	Males	All	Females	Males	All	Females	Males	All
All	3331	2615	5946	1061	111	1133	95	172	23
Asians	126	116	242	115	282	207	1041	1045	949
Blacks	483	328	811	383	346	564	773	874	592
Hispanics	588	434	1022	431	868	613	725	810	543
Whites	2134	1737	3871	950	984	1101	206	288	55

Table 2. The numbers of individuals, the related unique observations and missing observations under the matrix of age × cohort (i.e., 68 × 15) by gender and ethnicity.

	Individuals			Unique observations			Missing observations		
	Females	Males	All	Females	Males	All	Females	Males	All
All	3331	2615	5946	273	270	273	747	750	747
Asians	126	116	242	98	109	141	922	911	879
Blacks	483	328	811	209	186	243	811	834	777
Hispanics	588	434	1022	210	189	240	810	831	780
Whites	2134	1737	3871	270	268	273	750	752	747

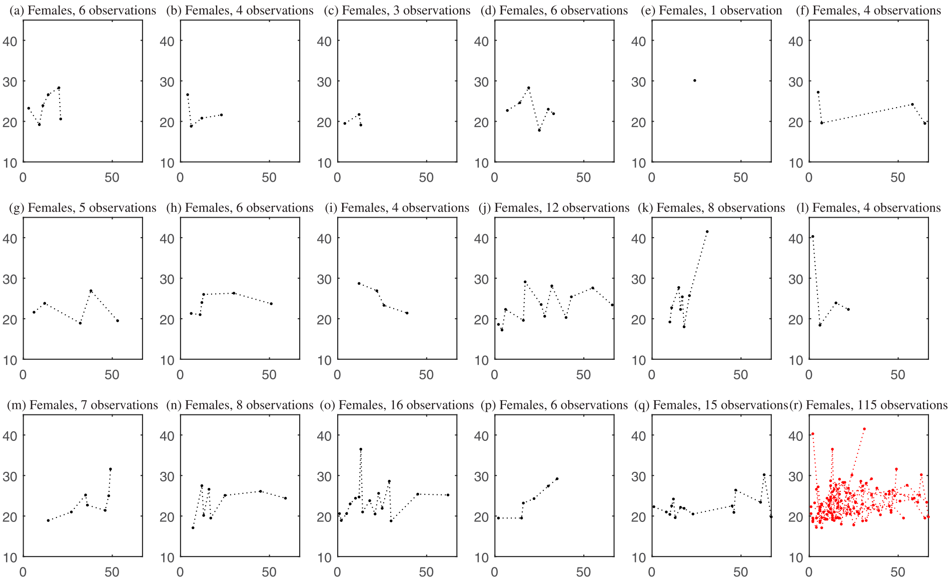

Figure 1. The observed BMI measurements of females in ages (large dots) and trajectories (dot lines) under different periods for Asians. Panels (a)–(q) are the observed BMI measurements in ages and trajectories from 1997 to 2013. Panel (r) is the observed BMI measurements in ages and trajectories for all 17-year periods. In each panel, *y*-axis is the BMI measurement and *x*-axis is the age.

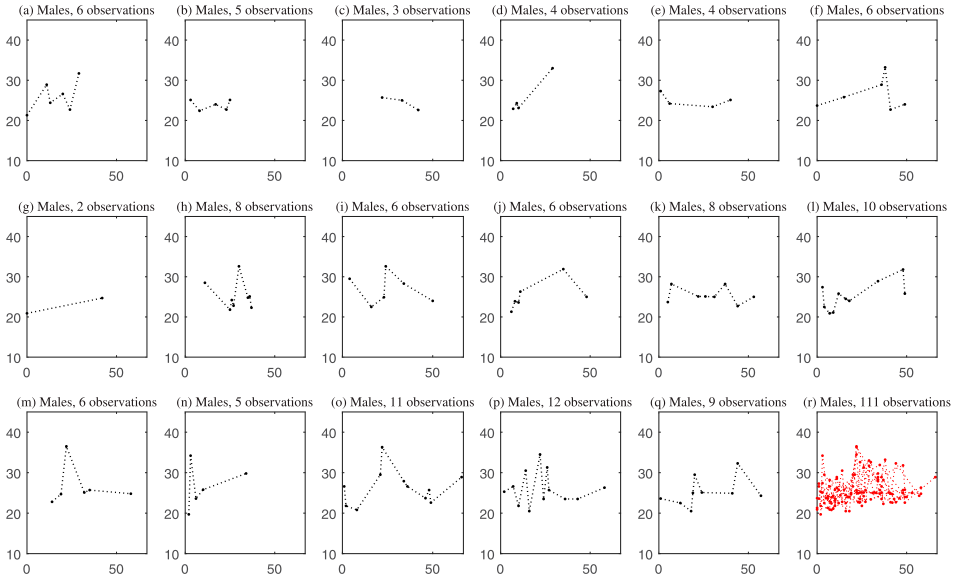

Figure 2. The observed BMI measurements of males in ages (large dots) and trajectories (dot lines) under different periods for Asians. Panels (a)–(q) are the observed BMI measurements in ages and trajectories from 1997 to 2013. Panel (r) is the observed BMI measurements in ages and trajectories for all 17-year periods. In each panel, *y*-axis is the BMI measurement and *x*-axis is the age.

under age × cohort. As the number of individuals decreases, the data set will become more sparse with more missing observations. As examples, Figure 1 and 2 show the observed BMI measurements of females and males in ages (large dots) and trajectories (dot lines) under different periods for Asians. It is seen that the ages are sparse and irregularly spaced under different periods for Asians. The data set for any other ethnicity shows similar patterns. When the data are measured on a fine grid of equally spaced points, the problem can be solved by applying the standard PCA. However, if the data are sparse with measurement at irregularly spaced points as given in Tables 1 and 2 and Figures 1 and 2, we need to impose functional PCA in the data analysis procedure.

1.3. Organization of the paper

The rest of the paper is organized as follows: In Section 2, we introduce basic concepts and definitions of the proposed two-stage functional PCA method. Application of our method to BMI data is in Section 3. Section 4 discusses the proposed method and gives the conclusion.

2. Method

In the proposed model, the age variable is parameterized as fixed effects, while the period and cohort variables can be parameterized as random effects. Since the range of the age categories is fixed and can be regarded as unique, we specify the age effects as fixed [24]. On the other hand, the time period and cohort categories are available for any specific analysis, typically are only samples from the population, we specify the period and cohort effects as random [24]. The proposed two-stage functional PCA provides estimates for a three-factor APC model. In the first stage, we estimate the age-period model; in the second stage, we consider the residual age–cohort effect conditional on the already estimated age-period effect [8].

Data smoothing for functional data attracts substantial interests recently for modeling sample trajectories. There are many data smoothing methods in FDA. We choose B-splines smoothing because it is well known for providing good approximation to smooth functions and its application in nonparametric smoothing is board [2,13]. The model uses a set of B-splines basis functions to represent the smoothed trajectories. Other choices of smoothing such as truncated power basis smoothing can in principal be used, but B-splines smoothing is preferable in the current context because it is in particular convenient and numerical stable [12,34].

2.1. Summary of the method

We mainly address APC analysis by a new two-stage functional PCA method. In the proposed method, we first describe the fixed age effect by a B-splines function, then the remaining term is described as two-stage functional PCA for both period and cohort. In the two-stage functional PCA, we consider the APC decomposition using principal components. The description of the eigenfunctions leads to more interpretable results. The random fluctuations in periods and cohorts are reliably described by the B-splines eigenfunctions. In the first stage, we perform functional PCA for the age–period effect. In the second stage, we consider the residual age–cohort effect conditional on the already estimated age–period effect. Hence the mixed effects for both period and cohort are included in a model with the fixed age effect. We capture all effects by B-splines curves in the model.

The proposed functional PCA model with B-splines has its advantage that it is well conditioned in the APC analysis [9]. Since the range of the age categories is fixed and can be regarded as unique, we specify the age effect as fixed [24]. On the other hand, the time period and cohort categories are available for any specific analysis, typically are only samples from the population, we specify the period and cohort effects as random under fixed trends [24]. Since both the fixed effect of age and mixed effects of period and cohort are assumed to be nonlinear by the B-splines, the curve fitting approach resolves the non-collinearity problem in the APC analysis [16].

2.2. FDA in the APC analysis

The main advantage of nonparametric over parametric models is their flexibility. The curve fitting approach in FDA involves smoothing curves with sparse and unbalanced data, which is well suited for our BMI data. In functional PCA, as long as we include enough of a number of spline basis, the placement of knots is not critical for the performance of the estimations. The B-spline basis functions are flexible enough to capture the patterns of the data. Please see James *et al.* [13], Ye *et al.* [33] and Zhou *et al.* [34] for more detailed discussions. To simplify the data analysis in the proposed method, we define the B-spline basis on equally spaced knots. Given the nature of the functional data in our analysis, 6–12 knots is often sufficient. In the FDA, we choose a cubic B-spline basis $\{\phi_l(t) : 1 \leq l \leq L\}$

with equally spaced knots, so that $\beta(t) = \sum_{l=1}^{L} \beta_l \phi_l(t)$. In the mixed-effects model framework, the fixed effect models the mean curve of the trend and the random effect allows for variations around the trend. The spline smoothing is performed for both the fixed and random effects. In our APC analysis, any random fluctuations in periods and cohorts will be reliably described and interpreted [1].

2.3. Functional PCA in the APC analysis

The main tool in FDA is functional PCA, where the observed trajectories are decomposed into a mean trend function and eigenfunctions, and the predictions of the functional principal component scores can serve as the random effects in the model. The functional PCA allows us to achieve the following three major goals [7]: Summarizing the data by a few functional principal components by dimension reduction of the functional data; estimating the functional principal components from sparse and unbalanced data; further analysis based on functional principal components scores.

Let N denote the number of study individuals. For the ith individual, let $Y_i(t) = \{Y_i(t_{ij}), j = 1, \ldots, n_i\}$ be the trajectory, t_{ij} be the observation times within the time interval $\mathcal{T}$, and n_i be the total number of such observation times. We consider a generalized functional mixed model, and assume that longitudinal observations $Y_i(t)$ are realizations of the canonical exponential family [19]. with a probability density or mass function

$$f(Y_{ij}|\theta_{ij}, \phi) = \exp\left[\frac{1}{a(\phi)}\{Y_{ij}\theta_{ij} - b(\theta_{ij})\} + c(Y_{ij}, \phi)\right], \qquad (1)$$

where θ_{ij} is the canonical parameter and ϕ is a dispersion parameter. Denote μ_{ij} as the mean of Y_{ij}, then μ_{ij} is the first derivative of $b(\cdot)$ at θ_{ij}, i.e. $\mu_{ij} = b^{(1)}(\theta_{ij})$. The inverse function of $b^{(1)}(\cdot)$, denoted as $g(\cdot)$, is the canonical link function, see [19] for the details of the model. Under the assumption that BMI data are Gaussian trajectories, we consider the longitudinal process $Y_i(t) = X_i(t) + \varepsilon_i(t)$, where the independent random error $\varepsilon_i(t) \sim \text{Normal}(0, \sigma_\varepsilon^2)$, and X_i yields a standard Karhunen–Loève expansion

$$X_i(t) = \mu(t) + \psi(t)^{\mathrm{T}}\xi_i, \quad \text{for } t \in \mathcal{T}, \qquad (2)$$

where the mean function $\mu(t) = \mathrm{E}\{X_i(t)\}$ represents the overall mean, $\psi = (\psi_1, \ldots, \psi_p)^{\mathrm{T}}$ is a vector of orthonormal functions also known as the eigenfunctions, the random vector $\xi_i = (\xi_{i1}, \ldots, \xi_{ip})^{\mathrm{T}} \sim \text{Normal}(0, D_\xi)$ are the principal component scores, $D_\xi = \text{diag}(d_1, \ldots, d_p)$ and $d_1 \geq d_2 \geq \cdots \geq d_p > 0$ are the eigenvalues. The number of principal components p will be chosen by a data-driven method. In theory, there can be infinite number of principal components, but p is often assumed to be finite for practical considerations. With the method of functional PCA, it is particularly important to identify the number of principal components. Some criteria have been proposed to determine the number of principal components. James *et al.* [13] discussed two natural approaches. The first approach is to calculate the proportion of variance explained by each principal component; and the second approach involves calculating the likelihood for the model as the number of principal components varies. The accuracy check found that the first approach worked well on a simulated data set [13]. Hence, we choose the first approach in our data analysis and determine the number of the principal component by the proportion of variance explained by the principal components.

In functional PCA method, we approximate the unknown functions $\mu(t)$ and $\psi(t)$'s by B-splines [13,34]. Let $\mathscr{B}(t) = \{\mathscr{B}_1(t), \dots, \mathscr{B}_q(t)\}^T$ be a q-dimensional B-spline basis defined on equally spaced knots in $\mathcal{T}$, θ_μ be a $q \times 1$ vector and $\Theta_\psi = (\theta_{\psi 1}, \dots, \theta_{\psi p})$ be a $q \times p$ matrix of spline coefficients, then the unknown functions are represented as $\mu(t) = \mathscr{B}(t)^{\mathrm{T}}\theta_\mu$ and $\psi^{\mathrm{T}}(t) = \mathscr{B}(t)^{\mathrm{T}}\Theta_\psi$. The general recommendation for choosing q in the literature is choosing a relatively large number $q \gg p$. As the number of functional principal components is relatively small in our APC applications, the number of basis functions q is usually selected to be a moderate number in the range of 6−12. The original B-spline basis functions are not orthonormal, therefore, we employ the procedure prescribed by Zhou *et al.* [34] to orthogonalize them so that $\int \mathscr{B}(t)\mathscr{B}(t)^{\mathrm{T}}dt = I_q$, where I_q is a $q \times q$ identity matrix. Under this construction, the orthonormal constraints on $\psi(t)$ translate to constraints on the coefficients, i.e. $\Theta_\psi^T \Theta_\psi = I_p$. Then the model takes the form

$$X_i(t) = \mathscr{B}(t)^{\mathrm{T}}\theta_\mu + \mathscr{B}(t)^{\mathrm{T}}\Theta_\psi \xi_i, \quad \text{subject to } \Theta_\psi^T \Theta_\psi = I_p. \tag{3}$$

The proposed functional PCA model is data adaptive, which does not require prespecified functional forms for longitudinal trajectories [7,32]. Under the definition of the exponential family, the longitudinal observations could be either continuous or discrete types, such as Gaussian, binomial or Poisson outcomes. By introducing a latent Gaussian process model for any types of observations, we establish a connection to the generalized FDA model.

2.4. Two-stage analysis

We consider a two-stage analysis in functional PCA, which provides estimates for the three-factor APC model [8]. We perform the age–period association model at first, and then consider the residual as the age–cohort effect conditional on the already estimated age–period effect, which is considered as an offset. Following the general guidelines in [8], the proposed mixed-effects model is

$$Y_i = X_{i_{\text{age}}} + X_{i_{\text{period}}} + \varepsilon_{i_{\text{residual1}}} = X_{i_{\text{age}}} + X_{i_{\text{period}}} + X_{i_{\text{cohort}}} + \varepsilon_{i_{\text{residual2}}}, \tag{4}$$

where $X_{i_{\text{age}}}, X_{i_{\text{period}}}, X_{i_{\text{cohort}}}$ are defined as

$$X_{i_{\text{age}}}(t_{i_{\text{age}}}) = \mathscr{B}(t_{i_{\text{age}}})^{\mathrm{T}}\theta_{\mu_{\text{age}}}, \tag{5}$$

$$X_{i_{\text{period}}}(t_{i_{\text{period}}}) = \mathscr{B}(t_{i_{\text{period}}})^{\mathrm{T}}\theta_{\mu_{\text{period}}} + \mathscr{B}(t_{i_{\text{period}}})^{\mathrm{T}}\Theta_{\psi_{\text{period}}}\xi_{i_{\text{period}}}, \tag{6}$$

$$X_{i_{\text{cohort}}}(t_{i_{\text{cohort}}}) = \mathscr{B}(t_{i_{\text{cohort}}})^{\mathrm{T}}\theta_{\mu_{\text{cohort}}} + \mathscr{B}(t_{i_{\text{cohort}}})^{\mathrm{T}}\Theta_{\psi_{\text{cohort}}}\xi_{i_{\text{cohort}}}. \tag{7}$$

In the first stage of functional PCA, we consider both age and period trends as the fixed effects and the age–period effect as the random effect. The age–period association model is

$$Y_{i_{\text{stage1}}} = X_{i_{\text{age}}}(t_{i_{\text{age}}}) + X_{i_{\text{period}}}(t_{i_{\text{period}}}) + \varepsilon_{i_{\text{residual1}}}. \tag{8}$$

In the second stage of functional PCA, we consider the residual effect conditional on the already estimated age–period effect, i.e., $Y_{i_{\text{stage2}}} = \varepsilon_{i_{\text{residual1}}}$. We apply the functional PCA to the conditional age–cohort effect, leading to the model in the second stage as

$$Y_{i_{\text{stage2}}} = X_{i_{\text{cohort}}}(t_{i_{\text{cohort}}}) + \varepsilon_{i_{\text{residual2}}}. \tag{9}$$

Under this approach, the age–cohort-dependent principal components are modeled as random walk time series, conditional on the age–period effect [8].

2.5. Implementation of the method

The study is to examine the effects of BMI in diverse gender groups including Asians, Blacks, Hispanics and Whites. The measurements of age, period and cohort are sparse and irregularly spaced and may differ widely across the individuals. As BMI is a continuous measurement, the FDA enables prediction of individual smooth trajectories for the measurements. Hence the functional PCA method is feasible to handle the special longitudinal data. We implement functional PCA at two stages to solve the non-collinearity problem in the APC model.

The algorithms for performing functional PCA include the expectation–maximization (EM) algorithm [13] and the conditional expectation algorithm [32]. In the EM algorithm, the random effects, which are the principal component scores in the model, are treated as missing values, and parameter estimations are based on expectation steps and maximization steps alternatively. In functional PCA, the mixed-effects model can be considered as a reduced rank mixed-effects framework [33]. The reduced rank fitting procedure considers a rank constraint on the principal component scores and attempts to estimate the principal component curves by the EM algorithm. In functional PCA, the random rank model is a kind of mixed-effects model which focuses on a small number of leading principal component by B-splines. Another algorithm in functional PCA is the conditional expectation algorithm proposed by Yao *et al.* [32]. This method represents the continuous trajectories through the Karhunen–Loève expansion, determining the eigenfunction from the data. The conditional expectation algorithm is straightforward and works well in the presence of sparse and irregular longitudinal data under the Gaussian assumption. In the proposed method, the calculations of functional PCA take advantages of the two algorithms. We first perform the conditional expectation algorithm [32] to get the eigenvalues and eigenfunctions, then further use the idea of reduced rank model to smooth the trend and eigenfunctions by B-splines, where the principal components are subject to the orthogonality constraint. We determine the number of the principal component by the proportion of variance explained by the principal components.

To get tractable answers, we have made the assumption that the dependent variable is Gaussian among the observations. This assumption is both technically and practically reasonable in the application of BMI data analysis. Under this assumption, we consider a longitudinal process with a standard Karhunen–Loève expansion, which is the PCA in a continuous domain. However, if the data set is not Gaussian due to its nature, we can consider it as latent Gaussian and easily transform it by a link function under the exponential family.

3. Results

In the APC analysis of the BMI data described in Section 1.2, we consider the observed data as random curves. The functional PCA method attempts to characterize the random

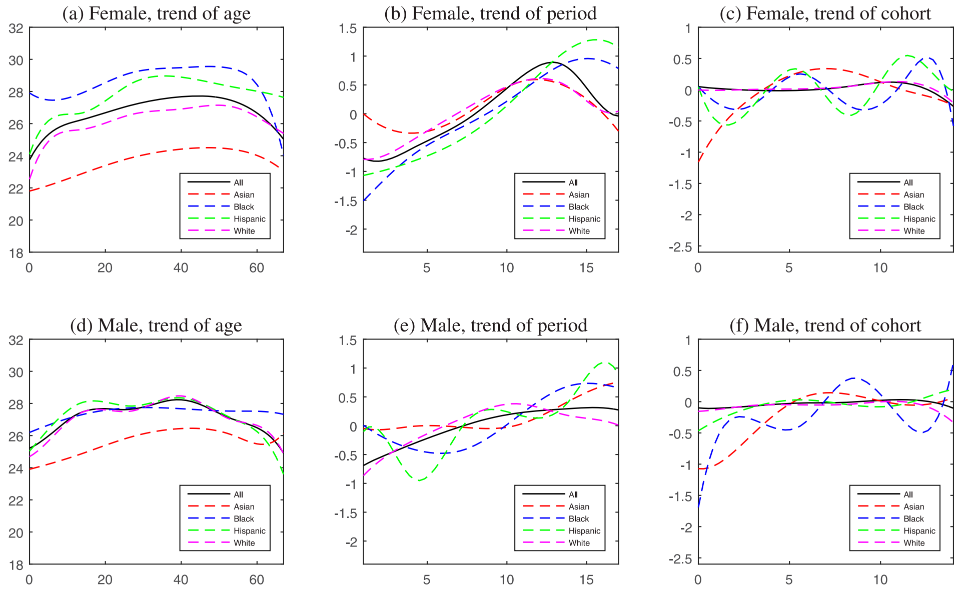

Figure 3. Different trends of the fixed-effect curves by gender and ethnicity. Panels (a) and (d) are the fixed age effect curves for females and males by ethnicity, where the age ranges from 0 to 67 years old; Panels (b) and (e) are the fixed period effect curves for females and males by ethnicity, where the period ranges from 1 to 17, corresponding to years 1997–2013; Panels (c) and (f) are the fixed cohort effect curves for females and males by ethnicity, where the cohort ranges from 0 to 14, corresponding to rounded values of (period — age —1940)/5.

trajectories around the overall mean trend functions. We represent the trajectories directly through the standard Karhunen–Loève expansion, determining the eigenfunctions from the data. The smooth estimates of the trajectories describe the trends of age, period and cohort at discrete points by gender and ethnicity. We case our approaches into smoothed mixed-effects models. The age effect is described by a fixed age trend curve from 0 to 67 years old (Figure 3). The period effect is described by mixed effects in the age–period association model at the first stage. The curves are decomposed as the sum of a fixed period trend curve from 1 to 17 (i.e. years 1997–2013) and random deviations from the trend period curve. The deviations are subsequently summarized by a few smoothed eigenfunctions extracted from the period trend subtracted data. The cohort effect is described by mixed effects in the age–cohort conditional association model at the second stage. The curves are decomposed as the sum of a fixed cohort trend curve from 0 to 14 and random deviations from the trend cohort curve. The deviations are subsequently summarized by a few smoothed eigenfunctions extracted from the cohort trend subtracted data.

Figure 3 reveals the different trends of the fixed-effect curves by gender and ethnicity. Panels (a) and (d) in Figure 3 are the fixed age effect curves for females and males by ethnicity. There is a clear indication of an age-specific trend. The curvature trend has a change point occurred at 46 for all females and at 40 for all males. The age effects are quite substantial, especially in the females. The overall means of Asians and Whites are below the average of all females; and the overall means of Blacks and Hispanics are above the average of all females (Figure 4(a)). The peaks are reached at ages 48, 48, 37 and 51 for Asians,

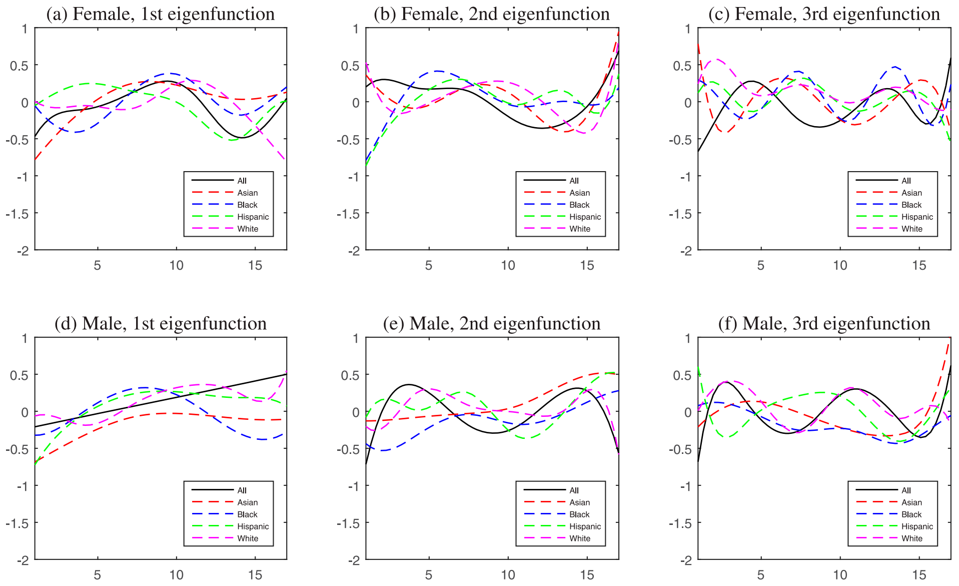

Figure 4. First three eigenfunctions over period effect by gender and ethnicity. Panels (a)–(c) are the results for females; and panels (d)–(f) are the results for males. The period ranges from 1 to 17, corresponding to years 1997 to 2013;

Blacks, Hispanics and Whites, respectively, among the females. For males, the shapes of the curves are a little bit different from those in females. The overall means of Blacks, Hispanics and Whites are close to the average of all males; and only the mean of Asians is below the average of all males (Figure 4(d)). The peaks are reached at ages 44, 32, 40 and 40 for Asians, Blacks, Hispanics and Whites, respectively.

Panels (b) and (e) in Figure 3 are the fixed period effect curves for females and males by ethnicity. Both females and males have increasing trends. However, the trends in females are more substantial. The trend of all females has a change point at around 13 (i.e. year 2009); and the trend of all males has a change point at around 16 (i.e. year 2012). In females, the trend curves of different ethnicity are more stable and the shapes are similar (Figure 3(b)). The peaks are reached at periods 12, 15, 16 and 12 (i.e. years 2008, 2011, 2012 and 2008) for Asians, Blacks, Hispanics and Whites for the females. Among males, the trend curves of different ethnicity are quite different (Figure 3(e)). The peaks are reached at periods 17, 15, 16 and 10 (i.e., years 2013, 2011, 2012 and 2006) for Asians, Blacks, Hispanics and Whites.

Panels (c) and (f) in Figure 3 are the fixed cohort effect curves for males and females by ethnicity. It is noted that there are no clear substantial trends for all females and males. However, for Hispanic females and Black females and males, there are periodicity patterns. Curvature patterns are observed for Asian females and males.

Tables 3 and 4 describe the proportion of variance associated with each of the first three principal components by gender and ethnicity. It is noted that for all gender and ethnicity groups, the first three principal components explain more than 85% of the variance in BMI. Especially, compared with the second and third components, the first principal component explains more than 47% of the variance in each ethnic group.

Table 3. In the first stage of functional PCA, the percentage variances of the first three principal components for the related individuals by gender and ethnicity.

PC Index	Females	PC1	PC2	PC3	Males	PC1	PC2	PC3
All	1.0000	0.4722	0.3696	0.1172	1.0000	0.7868	0.2105	0.0027
Asians	1.0000	0.9307	0.0647	0.0046	0.9997	0.9745	0.0217	0.0035
Blacks	0.9981	0.4900	0.3407	0.0983	0.9996	0.9324	0.0473	0.0199
Hispanics	0.9999	0.6085	0.2767	0.0940	0.9668	0.7617	0.1490	0.0561
Whites	1.0000	0.5829	0.1818	0.1289	0.9869	0.6616	0.2521	0.0732

Table 4. In the second stage of functional PCA, the percentage variances of the first three principal components for the related individuals by gender and ethnicity.

PC Index	Females	PC1	PC2	PC3	Males	PC1	PC2	PC3
All	0.9590	0.4722	0.3696	0.1172	1.0000	1.0000	< 0.0001	< 0.0001
Asians	1.0000	0.9307	0.0647	0.0046	0.9992	0.7233	0.2586	0.0173
Blacks	0.9290	0.4900	0.3407	0.0983	0.9805	0.4751	0.3467	0.1587
Hispanics	0.9792	0.6085	0.2767	0.0940	0.9919	0.7771	0.1800	0.0348
Whites	0.8936	0.5829	0.1818	0.1289	0.8507	0.5058	0.2121	0.1328

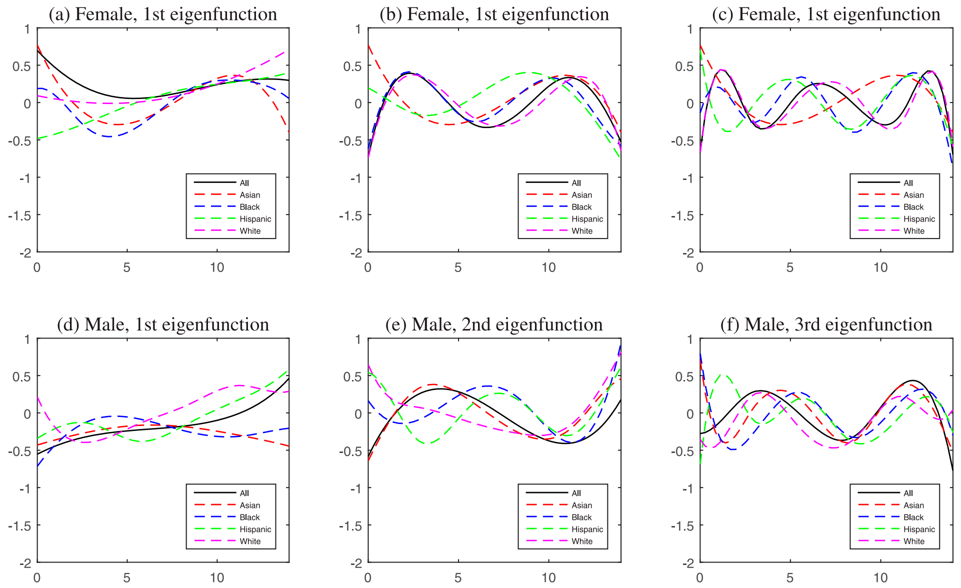

Figure 5. First three eigenfunctions over cohort effect by gender and ethnicity. Panels (a)–(c) are the results for females; and panels (d)–(f) are the results for males. The cohort ranges from 0 to 14, corresponding to rounded values of (period — age —1940)/5.

Figures 4 and 5 reveal the first three eigenfunctions in period and cohort by gender and ethnicity. The eigenfunctions correspond to the effects of different level shifting from the overall trend curve; they reflect variations about the trend curve over the period or cohort effect. Since the first principal component is the most important, the first eigenfunction essentially yields a summary statistics that is comparable with the fixed trend in the period effect or the cohort effect. The second and third eigenfunctions show more periodicity in the period and cohort. Compared with the first eigenfunction, they explains less variance in the BMI data and can be viewed as correction factors from the first principal component. From Figures 4 and 5, it is seen that the first eigenfunctions for females are different from those for males, reflecting the different variations over the period or cohort effect.

4. Conclusion, discussion and future work

4.1. Conclusion

In this paper, we have investigated a new approach to APC analysis and applied this approach to the BMI data by gender and ethnicity. We propose the use of a two-stage functional PCA method to describe the variability in longitudinal data in order to obtain more accurate estimates. The proposed new method adds a great degree of flexibility in the APC model using a nonparametric model structure. Incorporating the functional PCA method in the APC analysis improves the statistical and numerical stability of the estimations. The principal patterns of variations about the trend curves are described by the different eigenfunctions. The use of random effects adds flexibility of the estimates during the period or cohort when there are variabilities among the individuals, which reduces the random trajectories to a set of functional principal component scores. Our results show that the functional PCA method works well for the BMI data. The method is especially suitable for data with sparse and irregularly spaced measurements over age, period and cohort. The two-stage approach in functional PCA resolves the non-collinearity problem in the APC analysis and shows its capability to reveal the three-way structure in the age, period and cohort effects.

4.2. Discussion and future work

It is noted that the random effects in the proposed model are new style random effects for implementing data smoothing in FDA [10]. Compared with the old style random effects, the new style random effects do not meet the traditional definition of random effects [27]. We understand that the new style random effects are simply tools for estimating ensembles of fixed but unknown quantities [10]. In the classic linear mixed-effects models, the old style random effects are integrated out and the fixed effects are estimated using marginal likelihood [20]. However, the functional mixed-effects models with new style random effects mainly focus on nonlinear spline smoothing, which are different from the classic linear mixed models. The new style random effects are not integrated out, but instead are conditioned upon in an additive multi-step procedure. Thus, these conditional random effects behave like the fixed effects, but further provide smoothed estimates for specific individuals [20].

In the APC models, there is always a non-collinearity problem in parameter estimations of age, period and cohort. The proposed two-stage function PCA method has two advantages in the APC analysis: the first is to deal with the non-collinearity problem by smoothing; the second is to impute the missing data from predicted trajectories. Because the purpose of the study is to understand the relationship of age, period and cohort, the two-stage method reasonably approximates the relationship by the method of conditional expectations.

In APC analysis, the observed data are always measured by age and period initially; and then the data under cohort are calculated by the relationship of age, period and cohort. Although there is a perfect linear relationship among the three factors, i.e. period = age + cohort, the researchers group the age, period and cohort properties into time intervals of different lengths in practice. Since meaningful cohort is often considered to have a longer duration, it is usually grouped from age and period by a rounded integer during the analysis procedure. For example, the cohort is grouped by rounding (period − age − 1940) / 5 to the nearest integer in our paper. As the data under cohort are always calculated from the data under age and period, it makes more sense to consider the estimation of period as an earlier stage than the estimation of cohort. Hence, the researchers usually consider an APC model in the APC analysis.

Theoretically, when the data are regularly spaced over age, period and cohort and with a perfect linear relationship, the estimated results could be virtually the same if we revise the order of period and cohort in the two stages, i.e. changing APC model to age–cohort–period (ACP) model. Practically, the estimated results may have some kinds of differences in the APC model and ACP model depending on how the cohort is grouped from age and period and how many degrees of smoothness are performed for the sparse and irregularly spaced data.

Since a two-stage model might result in biased estimates of the conditional effects, the joint model approach can get better estimates because it aims to model the effects of age, period and cohort simultaneously. In the joint model approach, some parameters of the model need to be considered as missing values and the calculation of the integration will be approximated by a numerical approach. Hence, the Monte Carlo method and EM algorithm are typically employed to solve the problem in the joint modeling [13,33,34]. Because the joint model approach is more complex than the two-stage approach outlined in the paper, it is unknown whether any bias or undercoverage in the two-stage methods is large enough to warrant this extra modelling complexity. In our future work, we will compare the bias and efficiency of on two-stage approach and joint model approach under different assumptions in simulated data sets.

The bootstrap method can be used to produce pointwise confidence intervals for the overall mean function and the principal components. In the bootstrap, we resample the real data with replacement, fit the model to the bootstrap samples using the same analysis procedure as for the real data, and estimate the standard deviations of the estimators using their replicates pointwisely. However, the estimators are less informative because the data are too sparse in the resampling. Given the nature of the sparse and irregularly spaced measurements and typically low signal-to-noise ratio in the data, some statistical smoothing methods need to be carefully chosen during the analysis procedure. How to best resample the sparse data will be considered in our future work.

Acknowledgements

We thank the editor and the referees of an earlier version of this paper, who gave helpful suggestions on clarifying and explaining our ideas.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

[1] A. Bell and K. Jones, *Bayesian informative priors with Yang and Lands hierarchical age-period-cohort model*, Qual. Quantity 49 (2015), pp. 255–266.

[2] C. de Boor, *A Practical Guide to Splines*, Springer, New York, 1978.

[3] W. Fu, *A smoothing cohort model in age–period–cohort analysis with applications to homicide arrest rates and lung cancer mortality rates*, Sociol. Methods Res. 36 (2008), pp. 327–361.

[4] K. Fukuda, *Age–period–cohort decompositions using principal components and partial least squares*, J. Stat. Comput. Simul. 81 (2011), pp. 1871–1878.

[5] K. Fukuda, *A simple method for age-period-cohort decomposition of firm survival data*, Appl. Math. Comput. 219 (2012), pp. 741–747.

[6] K. Fukuda, *Principal component based generalized least squares approach for panel data*, J. Stat. Comput. Simul. 86 (2016), pp. 874–890.

[7] P. Hall, H.-G. Müller, and F. Yao, *Modelling sparse generalized longitudinal observations with latent Gaussian processes*, J. R. Statist. Soc., Ser. B 70 (2008), pp. 703–723.

[8] P. Hatzopoulos and S. Haberman, *A dynamic parameterization modeling for the age-period-cohort mortality*, Insur: Math. Econ. 49 (2011), pp. 155–174.

[9] C. Heuer, *Modeling of time trends and interactions in vital rates using restricted regression splines*, Biometrics. 53 (1997), pp. 161–177.

[10] J.S. Hodges, *Richly Parameterized Linear Models: Additive, Time Series, and Spatial Models Using Random Effects*, Chapman & Hall/CRC, New York, 2013.

[11] T. Holford, *Approaches to fitting age-period-cohort models with unrequal intervals*, Stat. Med. 25 (2006), pp. 977–933.

[12] J. Huang and L. Liu, *Polynomial spline estimation and inference of proportional hazards regression models with flexible relative risk form*, Biometrics. 62 (2006), pp. 793–802.

[13] G.M. James, T.J. Hastie, and C.A. Sugar, *Principal component models for sparse functional data*, Biometrika. 83 (2000), pp. 587–602.

[14] B. Jiang and K. Carriere, *Age–period–cohort models using smoothing splines: a generalized additive model approach*, Stat. Med. 33 (2014), pp. 595–606.

[15] R.A. Johnson and D.W. Wichern, *Applied Multivariate Statistical Analysis*, Pearson Prentice-Hall, Upper Saddle River, NJ, 2007.

[16] L. Knorr-Held and E. Rainer, *Projections of lung cancer mortality in West germany: A case study in Bayesian prediction*, Biostatistics 2 (2001), pp. 109–129.

[17] J. Komlos and M. Brabec, *The trend of mean BMI values of US adults, birth cohorts 1882–1986 indicates that the obesity epidemic began earlier that Hitherto thought*, Amer. J. Human Biol. 22 (2010), pp. 631–638.

[18] P.M. Krueger, K. Coleman-Minahan, and R.N. Rooks, *Race/ethnicity, nativity and trends in BMI among US adults*, Obesity 22 (2014), pp. 1739–1746.

[19] P. McCullagh and J.A. Nelder, *Generalized Linear Models*, Chapman and Hall, London, 1989.

[20] J.S. Morris, *Functional Regression*, Annu. Rev. Stat. Appl. 2 (2015), pp. 321–359.

[21] C.L. Ogden and M.D. Carrol, *Prevalence of Obesity among Children and Adolescents: United States, Trends 1963-1965 through 2007-2008*, Centers for Disease Control and Prevention, National Center for Health Statistics, Hyattsville, MD, 2010.

[22] C.L. Ogden, M.D. Carroll, B.K. Kit, and K.M. Fegal, *Prevalence of obesity and trends in body mass index among US children and adolescents, 1999–2010*, J. Amer. Med. Assoc. 307 (2012), pp. 483–490.

[23] J.O. Ramsay and B.W. Silverman, *Functional Data Analysis*, 2nd ed., Springer, New York, 2005.

[24] E. Reither, R. Masters, Y. Yang, D. Powers, H. Zheng, and K.C. Land, *Should age–period–cohort studies return to the methodologies of the 1970s*, Social Sci. Med. 128 (2015), pp. 356–565.

[25] E. Reither, R.M. Hauser, and Y. Yang, *Do birth cohorts matter? Age–period–cohort analyses of the obesity epidemic in the United States*, Social Sci. Med. 69 (2009), pp. 1439–1448.

[26] W.R. Robibson, R.L. Utz, K.M. Keyes, C.L. Martin, and Y. Yang, *Birth cohort effects on abdominal obesity in the United States: The Silent Generation, Baby Boomers and Generation X*, Int. J. Obes. 37 (2013), pp. 1129–1134.

[27] H. Scheffé, *The Analysis of Variance*, Wiley, New York, 1959.

[28] B.A. Swinburn, I. Caterson, J.C. Seidell, and W.P. James, *Diet, nutrition and the prevention of excess weight gain and obesity*, Public. Health. Nutr. 7 (2004), pp. 123–146.

[29] Y. Yang, W.J. Fu, and K.C. Land, *A methodological comparison of age–period–cohort models: The intrinsic estimator and conventional generalized linear models*, Sociol. Methodol. 34 (2004), pp. 75–110.

[30] Y. Yang and K.C. Land, *A mixed models approach to the age–period–cohort analysis of repeated cross-section surveys: Trends in verbal test scores*, Sociol. Methodol. 36 (2006), pp. 75–97.

[31] Y. Yang and K.C. Land, *Age-period-cohort analysis of repeated cross-section surveys: Fixed or random effects*? Sociol. Methods Res. 36 (2008), pp. 297–326.

[32] F. Yao, H.-G. Müller, and J.-L. Wang, *Functional data analysis for sparse longitudinal data*, J. Am. Stat. Assoc. 100 (2005), pp. 577–590.

[33] J. Ye, Y. Li, and Y. Guan, *Joint modeling of longitudinal drug using pattern and time to first relapse in cocaine dependence treatment data*, Ann. Appl. Stat. 9 (2015), pp. 1621–1642.

[34] L. Zhou, J. Huang, J.G. Martinez, A. Maity, V. Baladandayuthapani, and R.J. Carroll, *Reduced rank mixed effects models for spatially correlated hierarchical functional data*, J. Am. Stat. Assoc. 105 (2010), pp. 390–400.