scientific reports

Check for updates

OPEN

Bayesian semiparametric inference in longitudinal metabolomics data

Abhra Sarkar[1]✉, Ornella Cominetti[2]✉, Ivan Montoliu[2,5], Joanne Hosking[3], Jonathan Pinkney[3], Francois-Pierre Martin[2] & David B. Dunson[4]

The article is motivated by an application to the EarlyBird cohort study aiming to explore how anthropometrics and clinical and metabolic processes are associated with obesity and glucose control during childhood. There is interest in inferring the relationship between dynamically changing and high-dimensional metabolites and a longitudinal response. Important aspects of the analysis include the selection of the important set of metabolites and the accommodation of missing data in both response and covariate values. With this motivation, we propose a flexible but parsimonious Bayesian semiparametric joint model for the outcome and the covariate generating processes, making novel use of nonparametric mean processes, latent factor models, and different classes of continuous shrinkage priors. The proposed approach efficiently addresses daunting dimensionality challenges, simplifies imputation tasks, and automates the selection of important predictors. Implementation via an efficient Markov chain Monte Carlo algorithm appropriately accounts for uncertainty in various aspects of the analysis. Simulation experiments illustrate the efficacy of the proposed methodology. The application to the EarlyBird cohort study illustrates its practical utility in enabling statistical integration of different molecular processes involved in glucose production and metabolism. From this study, we were able to show that glucose levels from 5 to 16 years of age are associated with different circulating levels of metabolites in the blood serum and can be fitted over time for a wide range of shapes of trajectories. The metabolites contributing the most to explaining glucose trajectories tend to be involved in different central energy metabolomic pathways. The methodology provides a tool to generate new hypotheses related to obesity and glucose control during childhood and adolescence.

Metabolomics refers to comprehensive studies of amino acids, lipids, organic acids, or nucleotides, collectively known as metabolites, in biological systems. Metabolites' levels change in response to genetic or environmental changes. Metabolomics can thus provide detailed information about the biochemical mechanisms happening in an organism, potentially leading to discoveries of important biomarkers that can be used to diagnose, monitor, or predict the risk of diseases[1–8].

Generated by nuclear magnetic resonance (NMR) and/or mass spectrometry (MS), metabolomics datasets commonly involve tens of metabolites. Methods for analyzing static metabolomics data have been well developed[9,10]. However, longitudinal studies of the evolution of the course of time of the metabolites are becoming increasingly common[11–14]. Identifying the important metabolites responsible for the onset or progression of certain diseases is a challenging task, especially when dealing with complex datasets that include missing values. Efficient statistical methods to address these challenges within a coherent framework are however lacking. This need for sophisticated inference methods in high-dimensional longitudinal metabolomics datasets with missing values is the main motivation behind this article. The methods presented here, however, are broadly applicable to longitudinal studies beyond metabolomics, so the exposition of the statistical approach is kept fairly general.

The earlybird study: overview, inferential goals and analytical challenges

The rising prevalence of pre-diabetes and type 2 diabetes is a growing and alarming problem, associated with several short-term and long-term metabolic and cardiovascular complications[15–17]. However, the understanding of the underlying mechanisms that link the regulation of glucose and insulin in the early years of life is still incomplete.

The EarlyBird cohort study[18,19] is a non-interventional prospective longitudinal study of healthy UK children, designed to explore how anthropometrics and clinical and metabolic processes are associated with

[1]Department of Statistics and Data Sciences, University of Texas at Austin, Austin 78712-1823, USA. [2]Nestlé Research, Lausanne 1015, Switzerland. [3]University of Plymouth, Peninsula Schools of Medicine and Dentistry, Plymouth PL6 8BT, UK. [4]Department of Statistical Science, Duke University, Durham 27708-0251, USA. [5]Present address: Merck Biotech Development Center, Corsier-sur-Vevey 1809, Switzerland. ✉email: abhra.sarkar@utexas.edu; ornella.cominetti@rd.nestle.com

glucose control during childhood and adolescence. The full cohort comprises 307 children, 170 of which are boys (the sub-cohort considered in this study due to availability of metabolomics data comprises 129 subjects, 92 boys, and 37 girls), who were followed up with medical examination on an annual basis from 5 to 16 years of age (12 time points). The collected data included anthropometrics, glucose, and insulin measures. In addition, a metabolic profiling approach was applied to the serum samples collected from the children at each time point, using proton nuclear magnetic resonance (1H-NMR) spectroscopy. This method allowed the collection of quantitative information on the serum content in lipoprotein-bound fatty acyl groups found in triglycerides, phospholipids and cholesteryl esters, and major low molecular weight molecules present in blood, such as amino acids and other major organic acids. Based on internal databases, several 1H-NMR signals are assigned and representative peaks are integrated to provide quantitative information on the different biochemical compounds. 1H-NMR signals that could not be assigned based on experimental datasets and internal reference databases are coded as U.X, where X corresponds to the 1H-NMR chemical shift where the signal was detected. The final dataset contains 82 1H-NMR-derived measurements (metabolites) for each time point, 4 of which were highly correlated and were removed to generate a second analysis set. The Materials and Methods section provides additional details.

There is evidence in both adults and children that glucose levels high within the normal range are indicative of future diabetes. One-third of children showing transient hyperglycemia in the absence of any serious illness can be expected to develop diabetes within one year[20,21]. Therefore, we consider serum glucose to be the principal indicator of disease propensity. We also consider fasting blood glucose at the age of 16 years as closer to adult concentration values. Likewise, impaired fasting glycemia (IFG) identified at age 16 is considered to be a strong predictor of type 2 diabetes later in young adulthood. IFG is defined by the American Diabetes Association criteria with level from 5.6 mmol/L (100mg/dL) to 6.9 mmol/L (125 mg/dL).

In the present study, we employ novel longitudinal models to assess the association between fasting glucose concentration (the response y) and individual clinical variables, metabolites and select anthropometric measurements (the covariates x). Different metabolomic pathways have been postulated to elucidate the roles metabolites play in glucose production in the liver. The Cahill or glucose-alanine cycle[22,23], for example, refers to the metabolic pathway of the transport of carbons and amino groups from the muscle to the liver, whereas the Cori or lactic acid cycle[24] corresponds to the set of reactions that transports lactate from the muscle to the liver, where it is converted to glucose in the absence of oxygen and is then metabolized back to lactate to later re-enter the liver. The Krebs or citric acid cycle[25,26], on the other hand, is important in synthesizing glucose into other key biochemical products. Assessing the relevance of different metabolites in influencing glucose concentration in the EarlyBird study helps quantitatively elucidate these relationships in glucose metabolism in children and adolescents.

Significantly compounding the analytical challenges, glucose concentration, clinical variables, and metabolites all contain missing values with different missingness patterns (Figs. 1 and 2). The metabolites have missing values when the blood sample is not present, either because the subject did not attend the annual visit or because the sample was of poor quality after storage or not of enough volume for the NMR analysis. Clinical data are also missing for similar reasons. Specifically, the clinical variables measuring physical activity and respiratory quotient (variables 5 and 9 in Fig. 2) required additional visits and additional effort from the participants. In the case of the first variable, the children had to wear a device to determine their physical activity. The respiratory quotient measures the basal metabolic rate and was obtained by putting a face-mask on the children for 30 minutes. These intrusive measurements showed poorer compliance than the rest of the measurements obtained

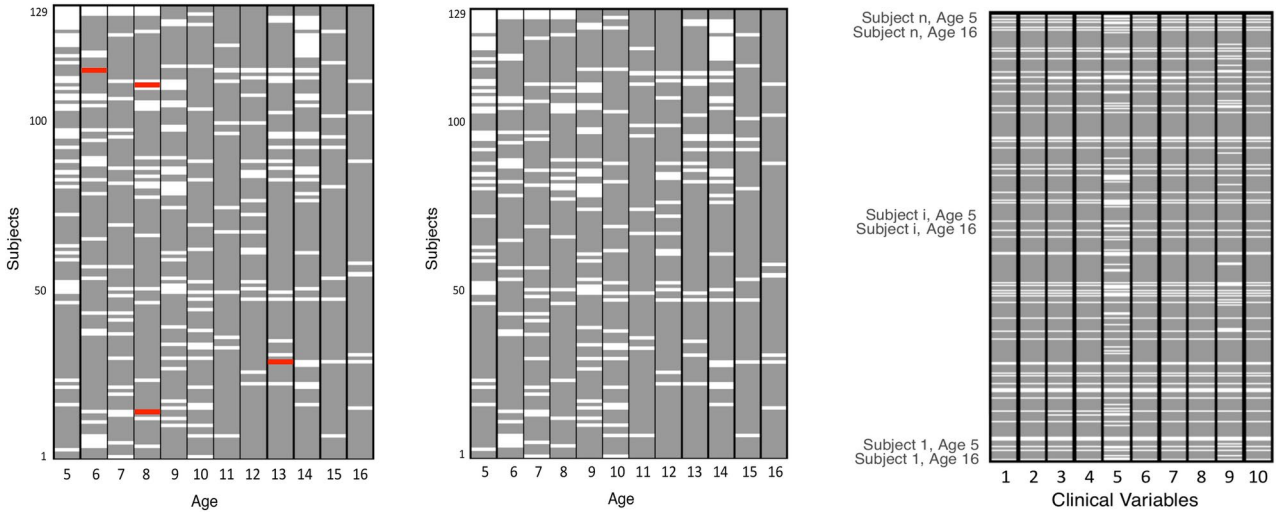

Figure 1. Missingness patterns of samples in the subcohort of the EarlyBird study. The grey cells represent observed values, and the white cells represent missing values. From left to right - missingness patterns in fasting glucose levels, in metabolite samples, and in clinical variables. The red cells on the left panel show additional missing values present only in fasting glucose levels.

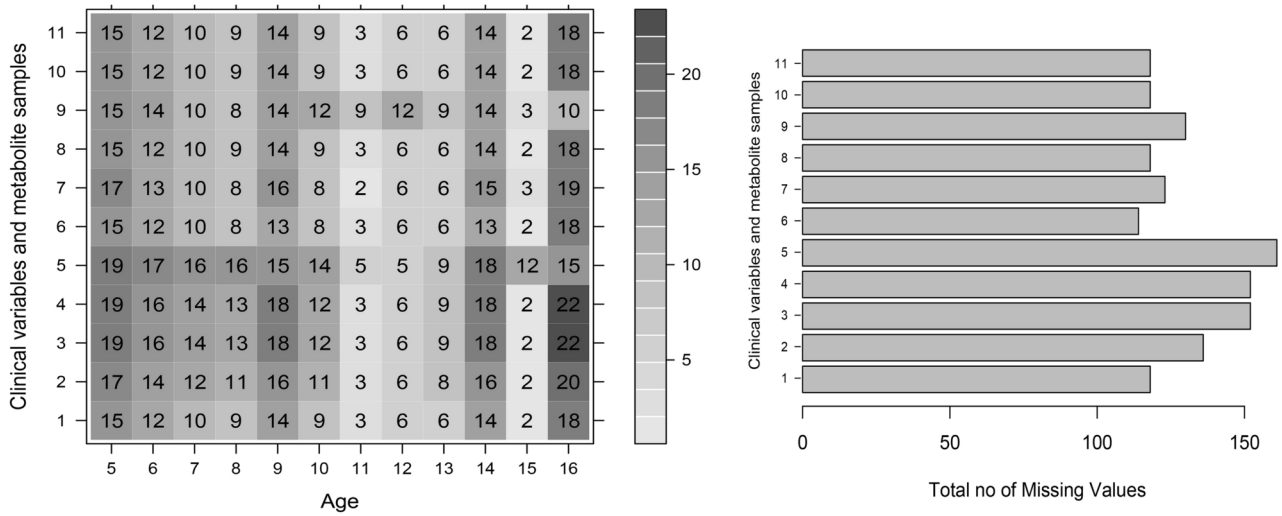

Figure 2. Number of missing values in the EarlyBird study in clinical variables (variables 1–10 on the y-axes) and metabolite samples (variable 11 on the y-axes) across different ages (left panel) and in total (right panel).

during their annual visit and were missing when the participants did not attend the additional visits. Glucose concentration had some additional missing values, shown in red in Fig. 1, attributable to human recording errors. The missingness mechanisms may be assumed to not depend on the true unobserved values of the missing data points and hence ignorable in nature (See Background and Section S.1 in Supplementary Information).

In addition to missingness, the high correlation between some of the metabolites is also an issue. Multicollinearity is evident from plots of variance inflation factors (Supplementary Figs. S.11 and S.13). Furthermore, the trajectories of the predictors are also widely different, with variability changing over time (Fig. 3).

The challenges presented by this complex dataset therefore include the confounding effect of growth in the metabolic signal, the presence of differently patterned missing values in the predictors and the response, the high dimensionality of the predictors, and their widely variable trajectories.

Our inferential tasks include the imputation of the missing values in y and x, which can be used in a variety of downstream analyses, inference on the relationships between y and x, and the selection of important covariates.

Background

The literature on longitudinal data and missing values is extensive. See, for example, books[27–34], and review papers[35–40], and the references therein.

The Bayesian paradigm provides a useful framework for handling missing data[29,41,42]. Specifying an appropriate joint probability model for the observed data, missing data, missingness mechanism, and the associated model parameters, Bayesian inferential machinery can naturally accommodate problems with missing data. Uncertainty in imputing the missing values is taken into account, and their finite sample estimates can also be readily obtained from samples drawn from the posterior.

The missingness mechanism can be ignored when the missing values are missing at random (MAR), i.e., the missingness does not depend on the missing data conditional on the observed data[43]. Bayesian inference then naturally relies on working with a joint model $p(y, \mathrm{x})$. It is often convenient to factorize $p(y, \mathrm{x})$ as $p(\mathrm{x})p(y|\mathrm{x})$ and then focus separately on $p(\mathrm{x})$ and $p(y|\mathrm{x})$. Jointly imputing the missing x values using a model $p(\mathrm{x})$ that properly accommodates their dependencies is especially beneficial when the components are strongly correlated, as is the case in the EarlyBird cohort. It is also natural to exploit the relationship encoded in $p(y \mid \mathrm{x})$ to impute the missing y's. This regression model can also play a role in imputing the missing x, although it is typically not very informative in that context. Flexible and innovative modeling strategies for $p(\mathrm{x})$ and $p(y|\mathrm{x})$ are crucial in further simplifying the imputation tasks and making the inferential exercises more robust.

We first discuss the challenges in building flexible and useful marginal models $p(\mathrm{x})$. For time-invariant multivariate covariates with missing values, it may be practically convenient to work with $p(x_j|x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p)$ for each j[44,45]. However, such sequential regression models may not correspond to a valid joint probability model for x. A related strategy, without this limitation, factors the joint distribution as $p(x_1, \ldots, x_p) = p(x_p|x_1, \ldots, x_{p-1}) \ldots p(x_2|x_1)\ p(x_1)$ and then models each one-dimensional conditional distribution separately[46,47]. With an increase in the dimension of the covariates, specifying a separate model for each component quickly becomes a difficult task. Additional complications arise when the covariates also evolve temporally, as in the EarlyBird study. The longitudinal trajectories of the different covariates, which look widely different, now have to be additionally modeled. The correlation structure among the x_j's may also be changing with time. The development of flexible and automated models for longitudinally evolving high dimensional covariates, accommodating widely varying individual trajectories without requiring to specify and fine-tune a separate model for each covariate while also allowing easy missing data imputation, is thus extremely challenging. Addressing this problem is an important focus of this article.

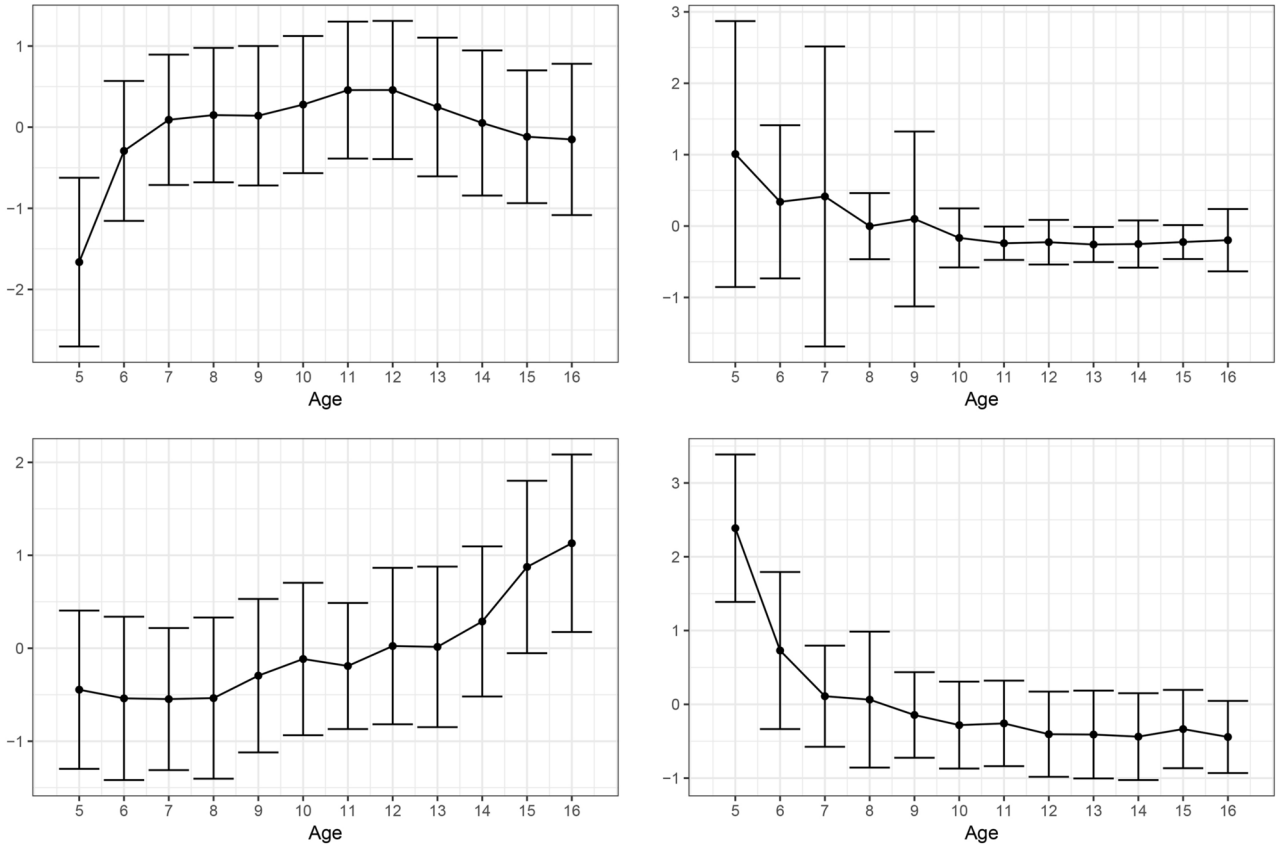

Figure 3. Mean longitudinal trajectories of four different standardized covariates. Clockwise from top left: unsaturated fatty acids, 3-hydroxybutyrate, creatinine, and acetate. The whiskers around the mean show one standard deviation unit in each direction.

Building a regression model $p(y|\mathrm{x})$ poses additional challenges. For a small number of covariates, traditional linear mixed models can be considered[48–50]. Unconstrained analysis involving all covariates, however, leads to complicated models with inflated variance, loss of predictive power, and difficulty in interpretation. The identification of important predictors is also of practical and scientific importance. The literature on variable selection in static, complete data settings is enormous. In recent years, it has become common to rely on optimizing a goodness-of-fit loss function having a penalty added to favor parsimony. Famous methods of this type include LASSO[51], adaptive LASSO[52], SCAD[53], and elastic net[54] among others. Adaptations to missing-data problems are not straightforward as they involve computing and optimizing penalized log-likelihood functions based only on the observed data. However, the EM algorithm can potentially be used[55–57].

Popular Bayesian strategies for variable selection include placing two-component 'spike and slab' priors on the regression coefficients[58–60]. Such priors place a point mass or spike at zero characterizing the redundant covariates and a continuous component or slab representing the signals. Adapting such approaches to missing data problems is conceptually straightforward by placing a probability model on the high-dimensional covariates and conducting posterior computation under the resulting joint model using MCMC[61]. However, the computational burden can be quite daunting, as even computation without missing predictors, or a model on $p(\mathrm{x})$, is pretty challenging. To reduce the computational burden in the absence of missing data, it is popular to rely on continuous shrinkage priors that are concentrated at zero with heavy tails[62–65]. These methods can induce variable selection via thresholding[63] or use of appropriate loss functions[66,67]. Adapting these techniques to build parsimonious regression models for longitudinal datasets involving high-dimensional covariates with ignorable missing values is another important goal of this article.

Toward these goals, we developed a detailed joint model to analyze the EarlyBird data set. The model was carefully designed to conform to a more general modeling framework, described in Section S.1 in the Supplementary Information, which simplifies the imputation tasks in both response and covariate values in longitudinal data with ignorable missingness. Specifically, we used nonparametric mean processes to capture widely varying covariate trajectories. We used latent factor formulations of the residual process to accommodate time-varying correlation structures, while meeting the related dimensionality challenges. Importantly, the implied conditional independence relationships also greatly simplified the imputation tasks. Taking a more structured approach, we model the response variable using a parametric regression model that accommodates the effects of the covariates and those of the aging process. Different classes of shrinkage priors stabilize estimation and help automate selection of the latent factors and important predictors via posterior variable selection summaries.

Materials and methods

Study population

The EarlyBird cohort study incorporates a 1995/1996 birth cohort recruited in 2000/2001 when the children were 5 years old (307 children, 170 boys). Several clinical and anthropometric variables were measured on an annual basis from the age of 5 to the age of 16. Untargeted metabolomics was performed in a subset of the full cohort (129 subjects, 92 boys), for a total number of 82 metabolites. The study was conducted in accordance with the ethics guidelines of the Declaration of Helsinki II; ethics approval was granted by the Plymouth Local Research Ethics Committee (1999), and parents gave written informed consent and children verbal informed assent.

Anthropometric variables

The 4 anthropometric variables are described in Table 1. BMI was derived from direct measurement of height (Leicester Height Measure; Child Growth Foundation, London, U.K.) and weight (Tanita Solar 1632 electronic scales), performed in blind duplicate and averaged. BMI SD scores were calculated from the British 1990 standards.

Clinical variables

The 10 clinical variables recorded in the EarlyBird study are described in Table 1. Peripheral blood was collected annually into EDTA tubes after an overnight fast and stored at -80° C. Insulin resistance (IR) and beta cell function were determined each year from fasting glucose (Cobas Integra 700 analyzer; Roche Diagnostics) and insulin (DPC IMMULITE) (cross-reactivity with proinsulin, 1%) using the homeostasis model assessment (HOMA-IR and HOMA-B, respectively).

Serum metabolomics

To measure the metabolites, 400μL of blood serum were mixed with 200μL of deuterated phosphate buffer solution 0.6 M KH2PO4, containing 1mM of sodium 3-(trimethylsilyl)-[2,2,3,3-2H4]-1-propionate (TSP, chemical shift reference δH = 0.0ppm). 550μL of the mixture were transferred into 5mm NMR tubes. 1H-NMR metabolic profiles of serum samples were acquired with a Bruker Avance III 600 MHz spectrometer equipped with a 5mm cryoprobe at 310K (Bruker Biospin, Rheinstetten, Germany) and processed using TOPSPIN (version 2.1, Bruker Biospin, Rheinstetten, Germany) software package as reported previously. Standard 1H-NMR one-dimensional pulse sequence with water suppression, Carr-Purcell-Meiboom-Gill (CPMG) spin-echo sequence with water suppression, and diffusion-edited sequences were acquired using 32 scans with 98K data points. The spectral data (from δ0.2 to δ10) were imported into Matlab software with a resolution of 22K data-points (version R2013b, the Mathworks Inc, Natwick MA) and normalized to the total area after solvent peak removal. Poor quality or highly diluted spectra were discarded from the subsequent analysis.

1H-NMR spectrum of human blood plasma enables the monitoring of signals related to lipoprotein bound fatty acyl groups found in triglycerides, phospholipids, and cholesteryl esters, together with peaks from the glyceryl moiety of triglycerides and the choline head group of phosphatidylcholine. This data also covers quantitative profiling of major low molecular weight molecules present in blood. Based on internal database, representative signals of metabolites assignable on 1H CPMG NMR spectra were integrated, including asparagine, leucine, isoleucine, valine, 2-ketobutyric acid, 3-methyl-2-oxovaleric acid, alpha-ketoisovaleric acid, (R)-3-hydroxybutyric acid, lactic acid, alanine, arginine, lysine, acetic acid, N-acetyl glycoproteins, O-acetyl glycoproteins, acetoacetic acid, glutamic acid, glutamine, citric acid, dimethylglycine, creatine, citrulline, trimethylamine, trimethylamine

Name	Description	Type
Sex	Gender (female or male).	Dichotomous
Weight cat	Weight category: grouping of children according to their average baseline weight: 0 Underweight (centile <= 2); 1 Normal weight (2< centile < 91); 2 Overweight (91<= centile < 98) and 3 Obese (centile >= 98).	Ordinal
ch gest	Gestational age of child measured in weeks.	Continuous
ch bwt sds	Birth weight SD scores.	Continuous
ch bmisd	Standardized BMI SD scores.	Continuous
Insulin	Plasma level of insulin measured through a standard biochemistry assay.	Continuous
HOMA2-B	Improved homeostatic model assessment of beta cell function, computed with the program HOMA Calculator v2.2.2 using the fasting plasma glucose and fasting plasma insulin levels.	Continuous
HOMA2-IR	Improved homeostatic model assessment of insulin resistance, computed with the program HOMA Calculator v2.2.2 using the fasting plasma glucose and fasting plasma insulin levels.	Continuous
ch mvpa	Moderate-and-vigorous physical activity measured through 7-d actigraph accelerometry.	Continuous
ch skf	Sum of skin-fold thickness at five sites on the left-hand side of the body (triceps, biceps, sub-scapular, supra-iliac and umbilical) by skin-fold calipers. Mean of two measurements.	Continuous
ch wcsd	Waist circumference, assessed by metal circumference measure. Mean of two measurements.	Continuous
ch wtsds	Weight. Mean of three repeated measurements taken, rounded to the nearest 100 g.	Continuous
RQ	Respiratory Quotient, corresponding to the ratio of carbon dioxide exhaled to oxygen inhaled as the body expends its energy reserves.	Continuous
ch bmi centile	BMI centile.	Continuous

Table 1. Anthropometric and clinical variables recorded in the EarlyBird cohort.

N-oxide, taurine, proline, methanol, glycine, serine, creatinine, histidine, tyrosine, formic acid, phenylalanine, threonine, and glucose. In addition, in diffusion edited spectra, signals associated to different lipid classes were integrated, including phospholipids containing choline, VLDL subclasses, unsaturated and polyunsaturated fatty acids. The signals are expressed in arbitrary unit corresponding to a peak area normalized to total metabolic profiles, which is representative of relative change in metabolite concentration in the serum.

Statistical analysis

Let y_{it} denote the response for subject i at time t, and x_{ijt} the associated j^{th} covariate, $i = 1, \ldots, n; t = 1, \ldots, T; j = 1, \ldots, p$.

Modeling the covariates Simple parametric models are insufficiently flexible for accommodating the wide variety of shapes of the longitudinal covariate trajectories (Fig. 3). Fine-tuning such models individually for each separate predictor to adapt them to different shapes is practically infeasible in high-dimensional applications like ours. Ideally, we would want to build flexible automated models which can accommodate widely varying shapes without any supervision. To this end, we model the covariate-generating process as

$$\begin{aligned} &\boldsymbol{\mu}_{x,t} = \boldsymbol{\mu}_{x,t-1} + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathrm{MVN}_p(0, \boldsymbol{\Delta}_\epsilon), \\ &\mathrm{x}_{it} = \boldsymbol{\mu}_{x,t} + \mathrm{b}_{x,i} + \boldsymbol{\xi}_{it}, \quad \mathrm{b}_{x,i} \sim \mathrm{MVN}_p(0, \boldsymbol{\Delta}_{x,b}), \quad \boldsymbol{\xi}_{it} \sim \mathrm{MVN}_p(0, \boldsymbol{\Sigma}_{x,t}). \end{aligned}$$

Here $\mathrm{MVN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a p-variate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The Markovian but otherwise unstructured mean process $\boldsymbol{\mu}_{x,t}^{\mathrm{T}} = [\mu_{x,1t}, \ldots, \mu_{x,pt}]$, that characterizes the temporal evolution of $[x_1, \ldots, x_p]^{\mathrm{T}}$, is crucial in accommodating widely varying trajectories of different x_j's in an automated way. The associated error process $\boldsymbol{\epsilon}_t$ has covariance $\boldsymbol{\Delta}_\epsilon = \mathrm{diag}\{\sigma^2_{\epsilon,1}, \ldots, \sigma^2_{\epsilon,p}\}$. The random vector $\mathrm{b}^{\mathrm{T}}_{x,i} = [b_{x,i1}, \ldots, b_{x,ip}]$, with covariance $\boldsymbol{\Delta}_{x,b} = \mathrm{diag}\{\sigma^2_{x,b,1}, \ldots, \sigma^2_{x,b,p}\}$, collects individual-specific random effects. We assign conjugate inverse-Gamma priors on the variance parameters as

$$\sigma^2_{x,b,j} \sim \text{Inv-Ga}(a_{x,b,\sigma}, b_{x,b,\sigma}), \qquad \sigma^2_{\epsilon,j} \sim \text{Inv-Ga}(a_{\epsilon,\sigma}, b_{\epsilon,\sigma}).$$

Here $\text{Inv-Ga}(a, b)$ denotes an inverse-Gamma distribution with shape parameter a and scale parameter b.

Exploratory analysis indicated some of the covariates to be highly correlated with changing correlation patterns over time. Time indexed covariance matrices for the $\boldsymbol{\xi}_{it}$, namely $\boldsymbol{\Sigma}_{x,t}$, greatly improve model flexibility but their large dimensions also present significant modeling challenges. To address this issue, we consider factor analytic representations as

$$\boldsymbol{\xi}_{it} = \boldsymbol{\Lambda}_t \boldsymbol{\eta}_{it} + \mathrm{u}_{it}, \quad \boldsymbol{\eta}_{it} \sim \mathrm{MVN}_p(0, \mathrm{I}), \quad \mathrm{u}_{it} \sim \mathrm{MVN}_p(0, \boldsymbol{\Delta}_{u,t}),$$

where $\boldsymbol{\Lambda}_t = ((\lambda_{tjh}))_{j=1,h=1}^{p,q_t} = [\boldsymbol{\lambda}_{1t}, \ldots, \boldsymbol{\lambda}_{pt}]^{\mathrm{T}}$ are $p \times q_t$ loading matrices, $\boldsymbol{\eta}^{\mathrm{T}}_{it} = [\eta_{it1}, \ldots, \eta_{itq_t}]$ are latent factors and u_{it} are associated idiosyncratic errors with covariance matrix $\boldsymbol{\Delta}_{u,t} = \mathrm{diag}(\boldsymbol{\sigma}^2_{u,t}) = \mathrm{diag}(\sigma^2_{u,1t}, \ldots, \sigma^2_{u,pt})$. Marginalizing out the latent factors, we have $\boldsymbol{\Sigma}_{x,t} = \boldsymbol{\Lambda}_t \boldsymbol{\Lambda}^{\mathrm{T}}_t + \boldsymbol{\Delta}_{u,t}$. Since any positive definite matrix $\boldsymbol{\Sigma}^{p\times p}$ admits a low rank and diagonal matrix decomposition $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}} + \boldsymbol{\Delta}$ for some $\boldsymbol{\Lambda}^{p\times q}$ and $\boldsymbol{\Delta} = \mathrm{diag}[\sigma^2_1, \ldots, \sigma^2_p]$ for some $0 \le q \le p$, in theory, the latent factor model is completely flexible. $\boldsymbol{\Sigma}$ involves $p(p+1)/2$ elements, whereas the number of parameters in a latent factor specification with q columns is $p(q+1)$. Often $q \ll p$ produces very good approximations of $\boldsymbol{\Sigma}$ while achieving a significant reduction in the number of parameters. In practice, data-driven and automated selection of q can be achieved using sparsity-inducing priors as described below. The overall strategy is particularly relevant in our application with a high-dimensional $\boldsymbol{\Sigma}_{x,t}$ at every t.

Separate latent factors $\boldsymbol{\eta}_{it}$ for each time point t is still too flexible for high-dimensional settings like ours, especially since entire samples of covariates can be missing (Fig. 1) in which case the $\boldsymbol{\eta}_{it}$ are informed entirely by the regression of y_{it} on x_{it}. Taking a middle path between a restrictive diagonal covariance matrix and a fully flexible model, we allow the latent factors $\boldsymbol{\eta}_i$ to be shared across time points, further greatly reducing model complexity. Integrating out both $\boldsymbol{\eta}_i$ and $\mathrm{b}_{x,i}$ thereby induce flexible variance and cross-covariance structures $\mathrm{cov}(\mathrm{x}_{it}) = \boldsymbol{\Lambda}_t \boldsymbol{\Lambda}^{\mathrm{T}}_t + \boldsymbol{\Delta}_{u,t} + \boldsymbol{\Delta}_{x,b}$ for all t and $\mathrm{cov}(\mathrm{x}_{it}, \mathrm{x}_{it'}) = \boldsymbol{\Lambda}_t \boldsymbol{\Lambda}^{\mathrm{T}}_{t'} + \boldsymbol{\Delta}_{x,b}$ for all $t \neq t'$.

Precluding the necessity to pre-specify the number of latent factors, we next allow the loading matrices to have a-priori a potentially infinite number of columns. Sparsity-inducing priors, that favor more shrinkage as the column index increases, can then be used to shrink the redundant columns toward zero. We do this via the multiplicative gamma process shrinkage priors[68] (MGPS) that allow easy posterior computation. For $t = 1, \ldots, T$ and $h = 1, \ldots, \infty$, we assign priors as follows

$$\begin{aligned} &\lambda_{tjh} \sim \text{Normal}(0, \phi^{-1}_{\lambda,tjh} \tau^{-1}_{\lambda,th}), \qquad \phi_{\lambda,tjh} \sim \mathrm{Ga}(\nu_\lambda/2, \nu_\lambda/2), \\ &\tau_{\lambda,th} \sim \textstyle\prod_{\ell=1}^{h} \delta_{t\ell}, \qquad \delta_{\lambda,t\ell} \sim \mathrm{Ga}(a_{\lambda,\ell}, 1), \qquad \sigma^2_{u,jt} \sim \text{Inv-Ga}(a_{u,\sigma}, b_{u,\sigma}). \end{aligned}$$

Here $\text{Normal}(\mu, \sigma^2)$ denotes a Normal distribution with mean μ and variance σ^2, and $\mathrm{Ga}(\alpha, \beta)$ denotes a Gamma distribution with shape parameter α and rate parameter β. The parameters $\{\phi_{\lambda,tjh}\}_{j=1}^{p}$ control the local shrinkage of the elements in the h^{th} column of $\boldsymbol{\Lambda}_t$, whereas $\tau_{\lambda,th}$ controls the global shrinkage. When $a_{\lambda,h} > 1$

for $h = 2, \ldots, \infty$, the sequence $\{\tau_{\lambda,th}\}_{h=1}^{\infty}$ is stochastically increasing and thus favors more shrinkage as the column index h increases. The shrinkage prior also helps alleviate rotational non-identifiability by assigning the strongest effects to the foremost factors.

Modeling the response Conditional on the covariates x_{it}, we model the response generating process as

$$y_{it} = \mu_{y,t} + b_{y,i} + \mathrm{x}_{it}^{\mathrm{T}}\boldsymbol{\beta} + v_{it}, \quad b_{y,i} \sim \mathrm{Normal}(0, \sigma_{y,b}^2), \quad v_{it} \sim \mathrm{Normal}(0, \sigma_v^2).$$

The mean process $\mu_{y,t}$ captures the temporal evolution of y and is modeled as

$$\mu_{y,t} = \mathrm{p}_{s,t}^{\mathrm{T}}\boldsymbol{\alpha},$$

where $\mathrm{p}_{s,t}^{\mathrm{T}} = [1, f(t,1), \ldots, f(t,s)]$, $f(t,\ell)$ being a normalized version of t^{ℓ}, and $\boldsymbol{\alpha}^{\mathrm{T}} = [\alpha_0, \alpha_1, \ldots, \alpha_s]$ denotes the associated coefficients. Sparsity-inducing priors, that favor more shrinkage as the degree of the polynomial increases, are used to favor simpler lower-degree relationships. We do this by adapting the MGPS priors as

$$\alpha_k \sim \mathrm{Normal}(0, \tau_{\alpha,k}^{-1}), \quad \tau_{\alpha,k} \sim \textstyle\prod_{\ell=1}^{k} \delta_{\alpha,\ell}, \qquad \delta_{\alpha,\ell} \sim \mathrm{Ga}(a_{\alpha,\ell}, 1).$$

As before, when $a_{\alpha,\ell} > 1$ for $\ell = 2, \ldots, s$, the sequence $\{\tau_{\alpha,k}\}_{k=1}^{s}$ is stochastically increasing, thereby favoring more shrinkage of the higher order coefficients towards zero.

Spline-based semiparametric regression models[69] could also be used to flexibly model $\mu_{y,t}$. The polynomial regression model with MGPS shrinkage priors on the coefficients, however, allows us to straightforwardly assess departures from simpler parametric alternatives, including a first-degree linear model.

The variables $b_{y,i}$, with variance $\sigma_{y,b}^2$, denote individual specific random effects in y_{it}. The effect of the predictor x_{it} on y_{it} is captured via linear regression with coefficients $\boldsymbol{\beta}^{\mathrm{T}} = [\beta_1, \ldots, \beta_p]$. To stabilize inference and favor the selection of important predictors in the presence of high-dimensional covariates with many possibly insignificant components, we use sparsity-inducing continuous priors on the regression coefficients. Unlike the columns of factor loading matrices and elements of $\boldsymbol{\alpha}$, there is no natural prior ordering of the elements of $\boldsymbol{\beta}$ making MGPS priors an inappropriate choice here.

The starting point of our search for an appropriate prior for $\boldsymbol{\beta}$ is a 'spike and slab' prior $\beta_j \sim \pi\delta_0 + (1-\pi)\mathrm{Normal}(0, \sigma_\beta^2), j = 1, \ldots, p$. Such priors, however, often have poor performance in high-dimensional settings, especially when the parameter vector is highly sparse. Choosing $\pi = 1/2$, for example, leads to an exponentially small prior probability of 2^{-p} assigned to the null model. Although this issue can be mitigated by assigning a hierarchical beta prior on π[70], posterior sampling in high-dimensional settings will still require a stochastic search over an enormous space, leading to slow mixing and convergence[71]. Continuous shrinkage priors such as the Bayesian LASSO[62], the horseshoe[64] and the Dirichlet-Laplace (DL)[65] mimic the spike and slab strategy by having a peak at zero to capture sparsity and heavy tails to capture significance, but improve computational issues by allowing efficient posterior sampling through hierarchical auxiliary variable constructions. Importantly, however, unlike the Bayesian LASSO and the horseshoe that only mimic the marginal behavior of point mass mixture priors, the DL prior also mimics the joint behavior of hierarchical spike and slab type mixture priors, thereby having better control of the joint sparsity of the parameter vector.

A DL prior on the regression coefficients can be specified as

$$\beta_j \sim \mathrm{DE}(\phi_{\beta,j}\tau_\beta), \quad (\phi_{\beta,1}, \ldots, \phi_{\beta,p}) \sim \mathrm{Dir}(a_\beta, \ldots, a_\beta), \quad \tau_\beta \sim \mathrm{Ga}(pa_\beta, 1/2).$$

Here $\mathrm{DE}(\sigma)$ denotes a double exponential distribution with location 0 and scale σ, and $\mathrm{Dir}(\alpha_1, \ldots, \alpha_p)$ denotes a Dirichlet distribution with concentration parameter $(\alpha_1, \ldots, \alpha_p)$. For $a_\beta < 1$, with $\phi_{\beta,j}$ integrated out, the marginal distribution of β_j given τ_β has a singularity at zero. The parameter τ_β determines how the tails of the marginal distribution decay as $|\beta_j|$ increases. The hyper-prior on τ_β allows uncertainty in this parameter and learning from the data. The DL prior can be equivalently represented as

$$\begin{aligned} &\beta_j \sim \mathrm{Normal}(0, \psi_{\beta,j}\phi_{\beta,j}^2\tau_\beta^2), \quad \psi_{\beta,j} \sim \mathrm{Exp}(1/2), \\ &(\phi_{\beta,1}, \ldots, \phi_{\beta,p}) \sim \mathrm{Dir}(a_\beta, \ldots, a_\beta), \quad \tau_\beta \sim \mathrm{Ga}(pa_\beta, 1/2), \end{aligned}$$

which facilitates straightforward posterior computation via an efficient block Gibbs sampler.

Finally, we assign conjugate priors on the variance parameters as

$$\sigma_{y,b}^2 \sim \text{Inv-Ga}(a_{y,b,\sigma}, b_{y,b,\sigma}), \qquad \sigma_v^2 \sim \text{Inv-Ga}(a_{v,\sigma}, b_{v,\sigma}).$$

In many applications, especially in epidemiological studies with children and young adults as subjects, the natural growing or aging process might be expected to have some effect on the outcome of primary interest y. The mean process $\mu_{y,t}$ separates this effect from that of the covariates. Unlike $\boldsymbol{\mu}_{x,t}$ which captures the temporal

evolution of x_{it} but is of no independent interest to us and hence is left unstructured, understanding $\mu_{y,t}$, the natural evolution of y as the subjects age, is often an important goal in epidemiological studies. To this end, we took a more structured approach towards modeling $\mu_{y,t}$ and used a polynomial function of time that is easy to interpret and, if needed, related tests of hypotheses can also be easily performed.

Posterior inference and missing data imputation For posterior inference, we rely on samples drawn from the posterior using a Markov chain Monte Carlo (MCMC) sampler, missing data imputation being a naturally integrated part of the sampler. The model described above is designed carefully to conform to a more generic framework for longitudinal data with ignorable missingness, described in Section S.1 in the Supplementary Information, which simplifies the imputation steps of the MCMC sampler for both missing response and covariate values.

The design of the sampler exploits the conjugacy of the priors and the conditional independence relationships encoded in different layers of the hierarchy. The latent factor formulation of the model for the covariates plays a particularly important role in inducing conditional independence between different components of x_{it} for each i at each t, simplifying the sampling of their missing values. We also evaluate post-processing model and variable selection criteria based on samples drawn from the posterior. See Section S.2 in the Supplementary Information for additional details.

Results

This section summarizes the results of our proposed approach applied to the EarlyBird dataset. To our knowledge, this is the first time we can report a comprehensive metabolic contribution of specific metabolic processes to overall blood glucose variations in a longitudinal and continuous manner. Such findings highlight the importance of specific metabolites in amino acid, ketone body, glycolysis, and fatty acid metabolism, in describing the variations of blood glucose throughout childhood.

We performed two sets of analyses. First, with all 96 time-varying predictors included, then, removing the last 4 metabolites that were the most strongly collinear. Some degree of multicollinearity was still present in the second set of 92 predictors (Supplementary Fig. S.13). Our proposed shrinkage prior based approach is robust to the presence of multicollinearity and the results obtained in the two cases were very similar. A competitor method, described later in this section and referred to as the 'lme' method, can not handle multicollinearity well. The results produced by this method were generally unstable, significantly more so in the first scenario. For space constraints, we summarize here only the results of the second set of analyses which were more favorable to the lme method.

Figure 4 illustrates the excellent performance of our model in capturing widely varying individual and average trajectories of the time-varying covariates. In Fig. 4 and similar others, the ages 5, 6, ..., 16 are represented by the time points 1, 2, ..., 12. Additional figures presented in the Supplementary Information (boxplots of observed and fitted values of the time-varying predictors across different time points in Fig. S.9 and plots of empirical correlations between time-varying predictors and the corresponding model estimates in Fig. S.13) show the effectiveness of our latent factors based approach in capturing the widely different and time-varying variance-covariance structures of the high dimensional time-varying predictors.

Figure 5 illustrates the results of the regression model, relating fasting glucose concentrations to age, 4 baseline anthropometric variables, 10 time-varying clinical variables and 78 time-varying metabolites. The average fasting glucose concentration levels vary nonlinearly with age (Fig. 5a and b). Among the key contributors or influencers of glucose trajectories, we observed several positive contributors to glucose variations, including glucose derived variables (HOMA-B, HOMA-IR), BMI z scores, 1H-NMR derived quantitative signals from glucose, lactate, and alanine (time-varying predictors 3, 4, 10, 27, 28, 66, 76 and 80 in Fig. 5c). In addition, we identified some

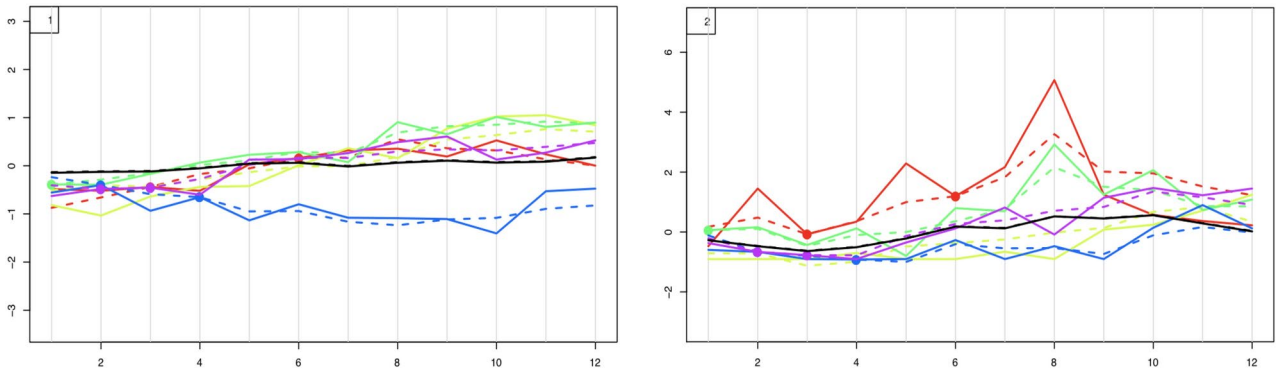

Figure 4. Results for the EarlyBird study: Observed (solid lines) and fitted (dotted lines) trajectories for the first 2 time-varying predictors for 5 randomly selected subjects super-imposed over time-specific sample means across all subjects (solid black line) and the corresponding fitted values (dotted black line). The bullets represent the mean imputed missing values, assumed to be equal to the unknown true values for plotting purposes.

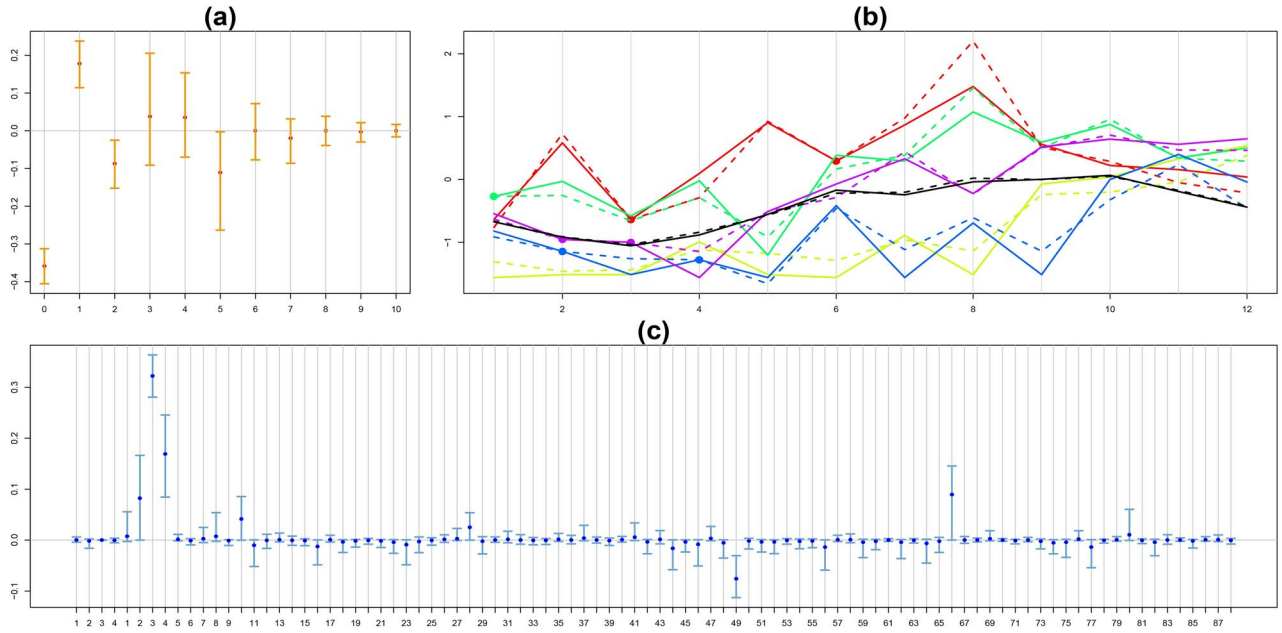

Figure 5. Results for the EarlyBird study: (**a**) Estimated posterior means of components of $\boldsymbol{\alpha}$ and their 90% credible intervals. (**b**) Observed (solid lines) and fitted (dotted lines) trajectories of y for 5 randomly selected subjects super-imposed over time-specific sample means across all subjects (solid black line) and the corresponding fitted values (dotted black line). The bullets represent mean imputed missing values, assumed to equal the unknown true values for plotting purposes. (**c**) Estimated posterior means of components of $\boldsymbol{\beta}$ and their 90% credible intervals. The first 4 components correspond to anthropometric baseline predictors. The remaining 88 components correspond to the time-varying predictors, comprising 10 clinical variables and 78 metabolites.

other metabolites characterized with a negative contribution to glucose variation, including 1H-NMR derived quantitative signals from citrate, ketone body 3-D-hydroxybutyrate, leucine, asparagine, and lipoprotein bound fatty acyl groups (time-varying predictors 11, 16, 23, 46, 49, 56, 64 and 77 in Fig. 5c). These results confirm the expected behavior between glucose, insulin resistance (HOMA-IR), and beta cell function (HOMA-B), as well as cross-platform measurement relationships (serum glucose by enzymatic assay and 1H-NMR spectroscopy).

The model thus allowed us to link childhood blood glucose variations with very different circulating levels of metabolites in the blood that correspond to the different central energy metabolic cascades. This is also reflected by the increasing contribution of lactate and alanine (time-varying predictors 27 and 28 in Fig. 5c) which corresponds to a higher contribution of the Cori and Cahyll cycles for glucose production, respectively. It is worth noticing how these molecular variations occur in a period of high metabolic flexibility during which the body of children switches from a high fat-protein basal metabolism towards a more carbohydrate dependent metabolic state. For instance, the model strongly highlights how decreasing circulating levels of the ketone body 3-D-hydroxybutyrate and citrate (a key intermediate in Krebs cycle) (time-varying predictors 23, 46, 49 and 77 in Fig. 5c) decrease concomitantly to the decrease in the circulation of some fatty acids over time. This corresponds to a trend towards lower fluxes of fatty acids to the liver for energy production. The amino acid asparagine production is highly connected to another Krebs cycle intermediate, oxaloacetate, and therefore the model may be describing these additional biological relationships.

In our discussion of the results so far, we have considered covariates with coefficients having marginal posterior credible intervals not including zero to be potentially important predictors of glucose concentrations. Taking a liberal approach, we also considered the covariates associated with highly variable coefficients, taking non-negligible values at least in some MCMC iterations, to also be potentially important. Conservative post-processing variable selection guidelines, described in Section S.2.4, lead to a model with 12 variables instead (Fig. 6). One such model with predictors chosen to correspond to coefficients with the largest absolute posterior means include (normalized versions of) $[1, t, t^2, t^5]$ and the clinical variables metabolites HOMA-B, HOMA-IR, BMI, decrease concomitantly to the decrease in the circulation of some fatty acids, lipids (mainly HDL, fatty acid CH3 moieties), an unknown metabolite (U.2.21), citrate, asparagine and 1H-NMR derived quantitative signals from glucose (time-varying predictors 3, 4, 10, 11, 44, 49, 56 and 66).

We also fitted a simpler sub-model for the covariates, referred to as the BSP-Diag model henceforth, $\mathrm{x}_{it} = \boldsymbol{\mu}_{x,t} + \mathrm{b}_{x,i} + \mathrm{u}_{it}$, $\mathrm{b}_{x,i} \sim \mathrm{MVN}_p(0, \boldsymbol{\Delta}_{x,b})$, $\mathrm{u}_{it} \sim \mathrm{MVN}_p(0, \boldsymbol{\Delta}_{u,t})$, excluding latent factors and assuming diagonal covariance matrices $\boldsymbol{\Sigma}_{x,t} = \boldsymbol{\Delta}_{u,t}$ for all t and hence independence of the covariate components.

Table 2 reports the estimated deviance information criterion (DIC)[72] and log pseudo marginal likelihood (LPML)[73] for the two Bayesian methods - the proposed latent factor based method and the diagonal covariance

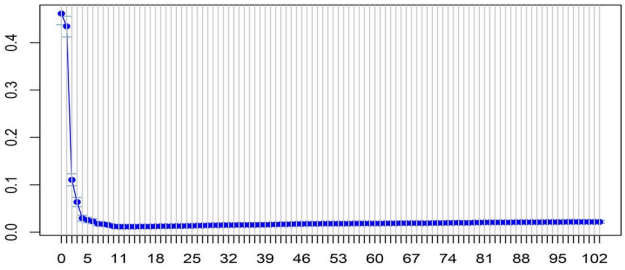

Figure 6. Post-processing variable selection results for the EarlyBird study: Model size vs the corresponding excess error ψ_λ. See Section S.2.4 in the Supplementary Information for additional details.

	Method	
Selection criteria	BSP-LF	BSP-Diag
DIC	142,823.1	286,136.1
LPML	-59.3189	-96.1627

Table 2. Deviance information criterion (DIC) and log pseudo marginal likelihood (LPML) estimates for the proposed Bayesian semiparametric latent factor based model (BSP-LF), and a related sub-model with diagonal covariance matrices for the covariates (BSP-Diag).

model described above. Compared to its diagonal covariance matrix restriction, the proposed latent factor based method clearly provides a much better fit to the EarlyBird data set.

We also implemented a standard lme approach, where we first imputed the missing values using cross means, implemented using the longitudinalData package in R, and then fitted a linear mixed model to the complete dataset, implemented using the lme4 package in R. To adhere to space constraints, we curtail the results produced by the BSP-Diag method and discuss here only the results produced by the lme method in greater detail.

The results for the lme method are summarized in Fig. 7. Most regression coefficients had high estimated variance (Fig. 7, note the much larger y-axis scales compared to Fig. 5). The covariates important according to the lme method (coefficients with confidence intervals not including zero) were anthropometric variables sex and weight (baseline predictors 1, 2) and clinical variables and metabolites BMI SD, insulin derived measures, HOMA-B, HOMA-IR, physical activity, skin thickness, waist circumference, isoleucine, N-acetylcysteine, glutamate, polyunsaturated fatty acids and phenylalanines (time varying predictors 1, 2, 3, 4, 5, 6, 8, 10, 17, 38, 47, 50, 61). However, the lme method did not perform well in realistic simulation settings (Section S.3, Figs. S.2, S.3, S.7, and S.8 in the Supplementary Information), thus undermining the reliability of these results.

Discussion

In this article, we considered the problem of estimating the longitudinal evolution of an outcome of interest in the presence of high-dimensional predictors comprising baseline covariates as well as longitudinally evolving ones when both the response and the time-varying covariates included missing values. We developed flexible statistical frameworks for inference in such problems when the missingness is ignorable in nature. Nonparametric mean processes captured widely varying covariate trajectories. Flexible, yet parsimonious, latent factor formulations of high-dimensional time-varying covariance matrices helped meet daunting dimensionality challenges, while also greatly simplifying the imputation tasks. Different classes of shrinkage priors automated the selection of latent factors and significantly important predictors, effectively guarding against model overfitting. An efficient Markov chain Monte Carlo algorithm accounted for uncertainty in various aspects of the analysis, including the imputation tasks.

Our assumption of ignorable missingness could be justified by the design of the EarlyBird study but may not be realistic in other scenarios, including studies involving mass-spectrometry-based metabolic data. Ongoing directions of research include extension of the proposed methodology to accommodate time-varying covariate effects, more flexible covariance patterns, more general mixed effects regression models, unequally spaced time points, non-ignorable missing values, nonlinear regression relationships etc.

Some of the metabolites and clinical variables identified as the best time-varying predictors of fasting glucose were known and expected, such as glucose-derived variables (bmi, HOMA-IR and HOMA-B) or glucose-metabolite byproduct (lactate). Additionally, metabolites like citrate, known for their impact on energy metabolism, and leucine, influencing insulin secretion, were also among the identified predictors. However, others, such as asparagine, may have indirect effects and could be context-dependent, worth exploring in more detail.

The sparsity-inducing priors in our work, originally designed for high-dimensional $n \ll p$ settings, are scalable to hundreds of predictors. We thus anticipate our method to also perform well under more extreme values of n and p. However, prior studies have not examined these methods in complex, longitudinal settings with missing data, as we do here. Testing their limits would thus require large-scale simulations with extreme

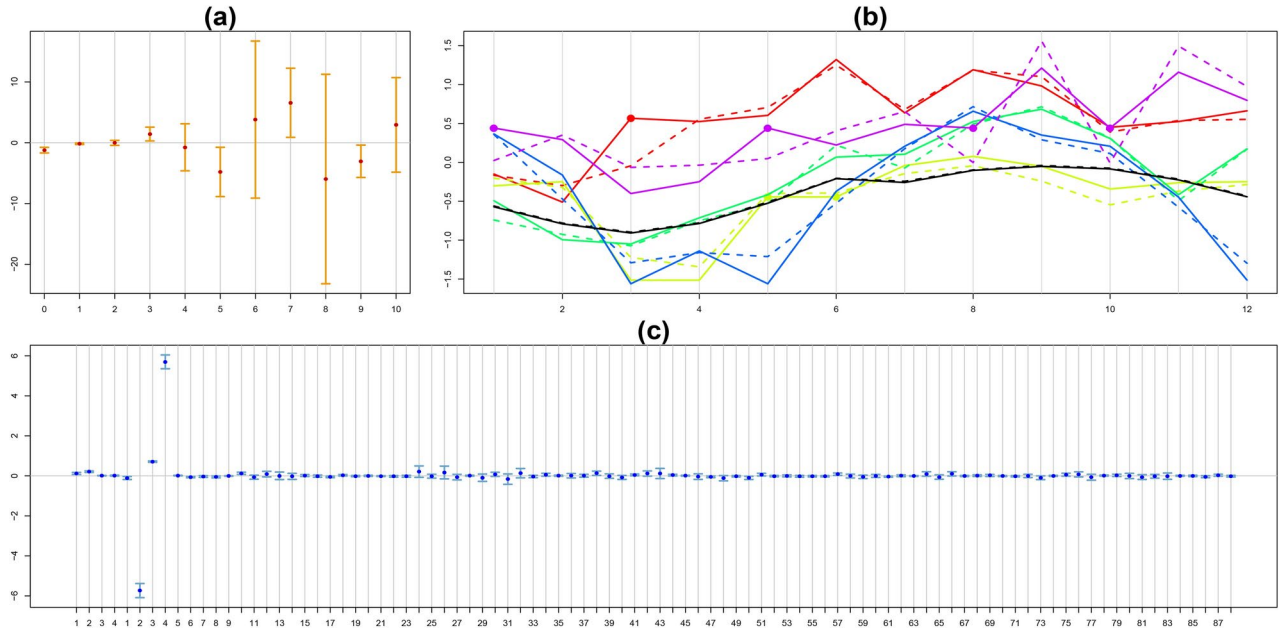

Figure 7. Results for the EarlyBird study obtained by the lme method: (**a**) Estimated values of α and their 90% confidence intervals. (**b**) Observed (solid lines) and fitted (dotted lines) trajectories of y for 5 randomly selected subjects super-imposed over time-specific sample means across all subjects (solid black line) and the corresponding fitted values (dotted black line). The bullets represent mean imputed missing values, assumed to equal the unknown true values for plotting purposes. (**c**) Estimated values of β and their 90% confidence intervals. The first 4 components correspond to anthropometric baseline predictors. The remaining 88 components correspond to the time-varying predictors, comprising 10 clinical variables and 78 metabolites.

n and p values, which, lacking a real-world problem at this scale, lies beyond the current scope but could be pursued in future work.

Additional information

Supplementary Information accompanying this paper presents a more generic modeling framework for longitudinal data with ignorable missingness, which simplifies the imputation tasks for both missing response and covariate values, details of the MCMC algorithm used to sample from the posterior, post-processing model selection and variable selection procedures, a simulation study evaluating the proposed method in synthetic settings, and a few additional figures summarizing the results of the EarlyBird application and the simulation studies.

Data Availability

Data may be available upon request to F.P.M. and J.P., subject, in particular, to ethical and privacy considerations. We have included a synthetic data set, $Simulated\_Data.RData$, which was simulated by closely mimicking the EarlyBird data set analyzed in the paper. Code is also shared.

Received: 17 July 2024; Accepted: 9 December 2024
Published online: 28 December 2024

References

1. Kaddurah-Daouk, R. & Krishnan, K. R. R. Metabolomics: A global biochemical approach to the study of central nervous system diseases. *Neuropsychopharmacology* **34**, 173 (2009).
2. Griffiths, W. J. et al. Targeted metabolomics for biomarker discovery. *Angew. Chem. Int. Ed.* **49**, 5426–5445 (2010).
3. Griffin, J. L., Atherton, H., Shockcor, J. & Atzori, L. Metabolomics as a tool for cardiac research. *Nat. Rev. Cardiol.* **8**, 630–643 (2011).
4. Patti, G. J., Yanes, O. & Siuzdak, G. Metabolomics: The apogee of the omic triology. *Nat. Rev. Mol. Cell Biol.* **13**, 263 (2012).
5. Nin, N., Izquierdo-García, J. & Lorente, J. The metabolomic approach to the diagnosis of critical illness. *Ann. Update Intensive Care Emerg. Med.* 43–52 (2012).
6. Savorani, F., Rasmussen, M. A., Mikkelsen, M. S. & Engelsen, S. B. A primer to nutritional metabolomics by NMR spectroscopy and chemometrics. *Food Res. Int.* **54**, 1131–1145 (2013).
7. Xia, J., Broadhurst, D. I., Wilson, M. & Wishart, D. S. Translational biomarker discovery in clinical metabolomics: An introductory tutorial. *Metabolomics* **9**, 280–299 (2013).
8. Gonzalez-Covarrubias Vanessa, D. B.-P. L., & Martínez-Martínez, E. The potential of metabolomics in biomedical applications. *Metabolites***12**, 2 (2022).
9. Bartel, J., Krumsiek, J. & Theis, F. J. Statistical methods for the analysis of high-throughput metabolomics data. *Comput. Struct. Biotechnol. J.* **4**, 1–9 (2013).

10. Ren, S., Hinzman, A. A., Kang, E. L., Szczesniak, R. D. & Lu, L. J. Computational and statistical analysis of metabolomics data. *Metabolomics* **11**, 1492–1513 (2015).
11. Jansen, J. J., Hoefsloot, H. C., Boelens, H. F., Van Der Greef, J. & Smilde, A. K. Analysis of longitudinal metabolomics data. *Bioinformatics* **20**, 2438–2446 (2004).
12. Rubingh, C. M. et al. Analyzing longitudinal microbial metabolomics data. *J. Proteome Res.* **8**, 4319–4327 (2009).
13. Smilde, A. K. et al. Dynamic metabolomic data analysis: A tutorial review. *Metabolomics* **6**, 3–17 (2010).
14. Mäkinen, V.-P. et al. Longitudinal metabolomics of increasing body-mass index and waist-hip ratio reveals two dynamic patterns of obesity pandemic. *Int. J. Obes.* **47**, 453–462 (2023).
15. Rosenbloom, A. L., Joe, J. R., Young, R. S. & Winter, W. E. Emerging epidemic of type 2 diabetes in youth. *Diabetes Care* **22**, 345 (1999).
16. Marcovecchio, M. L. & Chiarelli, F. Obesity and growth during childhood and puberty. *World Rev. Nutr. Diet.* **106**, 135–141 (2013).
17. Cominetti, O., Collino, S. & Martin, F.-P. Monitoring metabolism across childhood: Biomarkers for nutritional health and disease risk management. *Agro Food Ind. Hi Tech.* **25**, 14–18 (2014).
18. Voss, L. D. et al. Preventable factors in childhood that lead to insulin resistance, diabetes mellitus and the metabolic syndrome: The Earlybird diabetes study I. *J. Pediatr. Endocrinol. Metab.* **16**, 1211–1224 (2003).
19. Lauria, M. et al. Consensus clustering of temporal profiles for the identification of metabolic markers of pre-diabetes in childhood (EarlyBird 73). *Sci. Rep.* **8**, 1–16 (2018).
20. Herskowitz-Dumont, R., Wolfsdorf, J. I., Jackson, R. A. & Eisenbarth, G. S. Distinction between transient hyperglycemia and early insulin-dependent diabetes mellitus in childhood: A prospective study of incidence and prognostic factors. *J. Pediatr.* **123**, 347–354 (1993).
21. Hosking, J. et al. Divergence between hba1c and fasting glucose through childhood: implications for diagnosis of impaired fasting glucose (earlybird 52). *Pediatr. Diabetes* **15**, 214–219 (2014).
22. Kaneko, J. J., Harvey, J. W. & Bruss, M. L. *Clinical Biochemistry of Domestic Animals* (Academic Press, 2008).
23. Kohlmeier, M. *Nutrient Metabolism - Structures, Functions and Genes* (Academic Press, 2015).
24. Nuttall, F. Q., Ngo, A. & Gannon, M. C. Regulation of hepatic glucose production and the role of gluconeogenesis in humans: Is the rate of gluconeogenesis constant?. *Diabetes Metab. Res. Rev.* **24**, 438–458 (2008).
25. Lowenstein, J. M. *Methods in Enzymology, Volume 13: Citric Acid Cycle* (Boston: Academic Press) (1969).
26. Krebs, H. A. & Weitzman, P. D. *Krebs' citric acid cycle: Half a century and still turning* (Biochemical Society, London, 1987).
27. Diggle, P. *Analysis of Longitudinal Data* (Oxford University Press,) (2002).
28. Molenberghs, G. & Kenward, M. *Missing data in clinical studies* (John Wiley & Sons, 2007).
29. Daniels, M. J., & Hogan, J. W. *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis* (CRC Press, 2008).
30. Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. *Longitudinal Data Analysis* (CRC Press, 2008).
31. Verbeke, G., & Molenberghs, G. *Linear Mixed Models for Longitudinal Data* (Springer Science & Business Media,) (2009).
32. Enders, C. K. *Applied Missing Data Analysis* (Guilford Press, 2010).
33. van Buuren, S. *Flexible Imputation of Missing Data* (CRC press, 2018).
34. Little, R. J. A. & Rubin, D. B. *Statistical Analysis with Missing Data* (John Wiley & Sons, 2020).
35. Little, R. J. A. Missing-data adjustments in large surveys. *J. Bus. Econ. Stat.* **6**, 287–296 (1988).
36. Little, R. J. A. Regression with missing X's: a review. *J. Am. Stat. Assoc.* **87**, 1227–1237 (1992).
37. Schafer, J. L. & Graham, J. W. Missing data: Our view of the state of the art. *Psychol. Methods* **7**, 147–177 (2002).
38. Ibrahim, J. G. & Molenberghs, G. Missing data methods in longitudinal studies: A review. *TEST* **18**, 1–43 (2009).
39. Ibrahim, J. G., Chu, H. & Chen, M.-H. Missing data in clinical studies: Issues and methods. *J. Clin. Oncol.* **30**, 3297–3303 (2012).
40. Carpenter, J. R. & Smuk, M. Missing data: A statistical framework for practice. *Biom. J.* **63**, 915–947 (2021).
41. Little, R. J. A. Calibrated Bayes, for statistics in general, and missing data in particular. *Stat. Sci.* **26**, 162–174 (2011).
42. Luo, S., Lawson, A. B., He, B., Elm, J. J. & Tilley, B. C. Bayesian multiple imputation for missing multivariate longitudinal data from a Parkinson's disease clinical trial. *Stat. Methods Med. Res.* **25**, 821–837 (2016).
43. Rubin, D. B. Inference and missing data. *Biometrika* **63**, 581–592 (1976).
44. Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J. & Solenberger, P. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Surv. Methodol.* **27**, 85–96 (2001).
45. Van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M. & Rubin, D. B. Fully conditional specification in multivariate imputation. *J. Stat. Comput. Simul.* **76**, 1049–1064 (2006).
46. Ibrahim, J. G., Lipsitz, S. R. & Chen, M.-H. Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *J. R. Stat. Soc. B* **61**, 173–190 (1999).
47. Ibrahim, J. G., Chen, M.-H. & Lipsitz, S. R. Bayesian methods for generalized linear models with covariates missing at random. *Can. J. Stat.* **30**, 55–78 (2002).
48. Laird, N. M. & Ware, J. H. Random-effects models for longitudinal data. *Biometrics* 963–974 (1982).
49. Verbeke, G., Molenberghs, G. & Verbeke, G. *Linear mixed models for longitudinal data* (Springer, 1997).
50. Rosa, G., Gianola, D. & Padovani, C. Bayesian longitudinal data analysis with mixed models and thick-tailed distributions using MCMC. *J. Appl. Stat.* **31**, 855–873 (2004).
51. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* **58**, 267–288 (1996).
52. Zou, H. The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **101**, 1418–1429 (2006).
53. Fan, J. & Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **96**, 1348–1360 (2001).
54. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. Roy. Stat. Soc. B* **67**, 301–320 (2005).
55. Ibrahim, J. G., Zhu, H. & Tang, N. Model selection criteria for missing-data problems using the EM algorithm. *J. Am. Stat. Assoc.* **103**, 1648–1658 (2008).
56. Garcia, R. I., Ibrahim, J. G. & Zhu, H. Variable selection for regression models with missing data. *Stat. Sin.* **20**, 149 (2010).
57. Jiang, J., Nguyen, T. & Rao, J. S. The E-MS algorithm: model selection with incomplete data. *J. Am. Stat. Assoc.* **110**, 1136–1147 (2015).
58. George, E. I. & McCulloch, R. E. Variable selection via Gibbs sampling. *J. Am. Stat. Assoc.* **88**, 881–889 (1993).
59. George, E. I. & McCulloch, R. E. Approaches for Bayesian variable selection. *Stat. Sin.* **7**, 339–373 (1997).
60. Ishwaran, H. & Rao, J. S. Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Stat.* **33**, 730–773 (2005).
61. Yang, X., Belin, T. R. & Boscardin, W. J. Imputation and variable selection in linear regression models with missing covariates. *Biometrics* **61**, 498–506 (2005).
62. Park, T. & Casella, G. The Bayesian LASSO. *J. Am. Stat. Assoc.* **103**, 681–686 (2008).
63. Carvalho, C. M., Polson, N. G. & Scott, J. G. Handling sparsity via the horseshoe. *In AISTATS* **5**, 73–80 (2009).
64. Carvalho, C. M., Polson, N. G. & Scott, J. G. The horseshoe estimator for sparse signals. *Biometrika* **97**, 465–480 (2010).
65. Bhattacharya, A., Pati, D., Pillai, N. S. & Dunson, D. B. Dirichlet-Laplace priors for optimal shrinkage. *J. Am. Stat. Assoc.* **110**, 1479–1490 (2015).
66. Bondell, H. D. & Reich, B. J. Consistent high-dimensional Bayesian variable selection via penalized credible regions. *J. Am. Stat. Assoc.* **107**, 1610–1624 (2012).

67. Hahn, P. R. & Carvalho, C. M. Decoupling shrinkage and selection in Bayesian linear models: a posterior summary perspective. *J. Am. Stat. Assoc.* **110**, 435–448 (2015).
68. Bhattacharya, A. & Dunson, D. B. Sparse Bayesian infinite factor models. *Biometrika* **98**, 291 (2011).
69. Ruppert, D., Wand, M. P. & Carroll, R. J. *Semiparametric Regression* (Cambridge University Press) (2003).
70. Scott, J. G. & Berger, J. O. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Stat.* **38**, 2587–2619 (2010).
71. Polson, N. G. & Scott, J. G. Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Stat.* **9**, 501–538 (2010).
72. Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Van Der Linde, A. Bayesian measures of model complexity and fit. *J. R. Stat. Soc. B* **64**, 583–639 (2002).
73. Geisser, S. & Eddy, W. F. A predictive approach to model selection. *J. Am. Stat. Assoc.* **74**, 153–160 (1979).

Acknowledgements

We thank Hélène Ruffieux (University of Cambridge and previously Nestlé Institute of Health Sciences) and Irina Irincheeva (Bristol Myers Squibb and previously Nestlé Institute of Health Sciences) for their helpful discussions and Martin Kussmann (Kussmann Biotech GmbH and previously Nestlé Institute of Health Sciences) and Loïc Dayon (Nestlé Research) for their support. Finally, we acknowledge the life and work of our former colleague Terence Wilkin (1945-2017), Professor of Endocrinology and Metabolism, whose vision and original thinking led to the creation of the EarlyBird study and the establishment of the collaboration that made possible the metabolomics study reported here. A.S. and D.B.D. thank Nestlé for partially funding a postdoctoral position for A.S. under D.B.D.'s supervision at Duke University through a University Research Agreement during the early stages of the work. D.B.D. was also supported in part by Merck & Co., Inc., through its support for the Merck Biostatistics and Research Decision Sciences (BARDS) Academic Collaboration.

Author contributions

J.H., J.P, and F.P.M. conceived and designed the study and interpreted the results. J.H. and J.P. collected the data. A.S., O.C., and D.B.D. developed the analytical approach. A.S., O.C., and I.M. analyzed the data. A.S. and O.C. wrote the paper. All authors read and approved the final manuscript.

Declarations

Competing interests

The authors declare that the Earlybird study received funding from Bright Future Trust, The Kirby Laing Foundation, Peninsula Medical Foundation, Diabetes UK, the EarlyBird Diabetes Trust, and Nestlé Research. Nestlé Research was involved in metabonomics data generation and data interpretation. The other funders were not involved in the study design, collection, analysis, interpretation of data, the writing of this article, or the decision to submit it for publication. O.C. is a full-time employee at the Societé des Produits Nestlé. F.P.M. and I.M. were full-time employees at the Societé des Produits Nestlé for the majority of the project duration. J.H. and J.P. are employees of Plymouth University Peninsula School of Medicine and Dentistry. J.H. has received funding from Nestlé. Nestlé also partially funded a postdoctoral position for A.S. under the supervision of D.B.D. at Duke University through a University Research Agreement during the early stages of the work.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-82718-8.

Correspondence and requests for materials should be addressed to A.S. or O.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.