Submitted to the Annals of Applied Statistics

FACTOR-ASSISTED LEARNING OF ULTRAHIGH-DIMENSIONAL COVARIATES WITH DISTRIBUTED FUNCTIONAL AND SCALAR MIXTURES WITH APPLICATIONS TO THE AVON LONGITUDINAL STUDY OF PARENTS AND CHILDREN

BY SHOUDAO WEN [1], LI LIU [2], JIN LIU [3], YI LI [4] AND HUAZHEN LIN [*1,a]

[1]*Center of Statistical Research and School of Statistics, New Cornerstone Science Laboratory, Southwestern University of Finance and Economics, Chengdu, China,*
[a]*linhz@swufe.edu.cn*

[2]*School of Mathematics and Statistics, Wuhan University, Wuhan, China*

[3]*School of Data Science, The Chinese University of Hong Kong, Shenzhen, China*

[4]*Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA*

Atherosclerosis is a chronic, multifaceted disease that affects multiple arterial systems. Its progression is primarily driven by low-density lipoprotein (LDL) cholesterol accumulation, which promotes localized arterial lesion formation. These lesions can lead to severe complications, including ischemic heart disease (IHD) and stroke. Both genetic factors, particularly single nucleotide polymorphisms (SNPs), and age-related changes in body composition significantly influence LDL levels, generating extensive ultrahigh-dimensional covariates from functional and scalar mixtures (UDFSM), which may be stored at different sites due to the massive amount of data and the different data representations. To analyze the impact of genetic and physiological variables on LDL levels, we first separately extract features from ultrahigh-dimensional functional and scalar covariates in an unsupervised manner. Then, we propose a novel regression model that incorporates these features, which may be correlated due to the underlying correlations in the ultrahigh-dimensional covariates comprising both functional and scalar mixtures. Our methodology employs a factor regression model with an additive multiple-index component to sufficiently and effectively capture latent feature-response variable relationships. We enhance model interpretability and account for covariate correlations by imposing column sparsity and low-rank structures on the regression coefficients matrix, thereby incorporating structural information to improve efficiency and robustness. This distribution-agnostic approach to the response variable ensures greater flexibility and versatility. For model fitting, we develop a sieve likelihood-based framework that leverages the problem's inherent structure to provide efficient and robust estimates. We apply our method to the Avon Longitudinal Study of Parents and Children (ALSPAC) dataset, achieving high prediction accuracy for LDL levels and identifying significant SNPs and anthropometric measures affecting LDL. We specifically examine how various anthropometric measures influence LDL levels over ages. We further extend our analysis to identify key parental and individual characteristics that influence adult body mass index (BMI).

1. Introduction. Atherosclerosis is a chronic and multifaceted disease affecting multiple arterial systems, often resulting in severe clinical outcomes, including ischemic heart disease (IHD) and stroke. According to the Global Burden of Disease 2021 report (Institute for

* **Corresponding Author**

Keywords and phrases: Supervised factor model, ultrahigh-dimensional covariates of distributed functional and scalar mixtures, kernel density estimation, group penalty, low rank.

Health Metrics and Evaluation (IHME), 2024), IHD and stroke ranked first and third among global causes of mortality, with age-standardized death rates of 108.7 and 87.4 per 100,000 population, respectively. Projections indicate that by 2050, these conditions will become the leading contributors to global disease burden. Therefore, effective and timely management of atherosclerosis risk factors is crucial for prevention and treatment. From 1990 to 2020, advances in atherosclerosis prevention and treatment have substantially improved life expectancy. Low-density lipoprotein (LDL) cholesterol, the primary cholesterol-transporting lipoprotein in human plasma, plays a central role in the progression of atherosclerosis. According to the unifying hypothesis of atherosclerosis etiology and pathogenesis (Schwartz et al., 1991), LDL cholesterol accumulation at lesion-prone arterial sites represents a critical early event in disease progression. As Packard et al. (2000) emphasized, enhancing our understanding of the biological mechanisms governing LDL metabolism and regulation could provide crucial insights for developing targeted therapeutic strategies against atherosclerotic disease.

Genetic determinants play a pivotal role in regulating LDL levels. Large-scale genome-wide association studies (GWAS) have identified numerous genetic loci, particularly single nucleotide polymorphisms (SNPs), that significantly influence LDL levels (Sandhu et al., 2008; Willer et al., 2013). Mendelian randomization studies have further established causal relationships between specific SNPs and LDL cholesterol (Ference et al., 2012; Jansen et al., 2014), validating these SNPs as reliable genetic markers for LDL. These findings highlight the importance of SNP data as reliable genetic markers for LDL. Based on these scientific findings, we utilize SNP data to develop a predictive model for LDL levels, providing a foundation for prevention and therapeutic strategies.

Anthropometric measures, including height, body mass index (BMI), and body fat percentage, are intrinsically linked to lipid metabolism and significantly influence LDL regulation. Understanding these relationships is crucial for developing effective LDL management strategies. Moreover, since anthropometric measures evolve with age, single time-point measurements may fail to capture critical temporal patterns, thereby limiting our understanding of LDL regulation. To address this limitation, our predictive model incorporates functional data on these anthropometric measures, enabling us to capture age-related trends and examine their impact on LDL levels across different life stages, ultimately enhancing prediction accuracy.

The Avon Longitudinal Study of Parents and Children (ALSPAC, https://www.bristol.ac.uk/alspac/), initiated in 1991, is a comprehensive longitudinal study investigating the effects of environmental factors, lifestyle, and genetics on health and development. This ongoing study has tracked approximately 14,000 pregnant women and their children, accumulating extensive data on pregnancy outcomes, birth characteristics, child development, and family environmental factors. ALSPAC represents an invaluable resource for analyzing health trajectories across the life course, providing detailed longitudinal measurements of physiological parameters—including weight, height, and blood pressure—alongside comprehensive genetic information such as SNP data. This combination yields abundant data with ultrahigh-dimensional risk predictors and allows data to be recorded with high frequency from diverse scientific fields, providing large volumes of ultrahigh-dimensional covariates of functional and scalar mixtures (UDFSM). The challenges in building the analysis framework for the ALSPAC dataset arise from the following aspects. First, ultrahigh-dimensional covariates of functional and scalar may be stored at different sites due to the massive amount of data and the different data representations. Second, longitudinal measurements for anthropometrics are functional while the SNPs are scalar covariates. Modeling the mixing types of functional and scalar covariates can be further complicated by the high dimensionality of both covariates, where functional covariates themselves are infinite-dimensional. Third, irregular longitudinal measurements make functional covariates sparse, presenting additional challenges

in a high-dimensional scenario. Therefore, to account for the distributed storage of the data and the different types of data, as well as to modulate the association between LDL and high-dimensional functional and scalar covariates, it is important to first separately and sufficiently extract a low-dimensional representation and then associate the extracted representation with the response. The paper attempts to build a highly predictive and intrinsically interpretable featured factor regression model to explicitly express the relationship.

It is well-known that achieving high prediction accuracy and model interpretability may conflict with each other. For example, a deep neural network may provide accurate predictions but with low interpretability, while some statistical regression models are explainable but have less predictability. In particular, a large amount of statistical literature has been developed to handle high-dimensional scalar covariates under the sparse assumption, including variable selection (Tibshirani, 1996; Fan and Li, 2001; Zou, 2006) and sure independence screening (Fan and Lv, 2008; Ma et al., 2017). However, their performance highly depends on the sparsity assumption and the well-known restricted eigenvalue condition, where the former requires that only a limited number of the covariates are associated with the response, and the latter requires that the effects of important covariates are well separated from those of the null covariates. These assumptions are violated due to the inevitable correlation among high-dimensional covariates (Hall et al., 2005; Fan et al., 2020). To alleviate the limitation of the sparsity assumption and make full use of the correlation among high-dimensional scalar covariates, factor regularized methods are proposed (Wang, 2012; Jiang et al., 2019; Fan et al., 2020).

In contrast to scalar variables, functional data involving infinite-dimensional processes are more complex. With advances in data generation, various methods have been developed to analyze functional data, including functional linear regression (Yao et al., 2005b; Cai and Hall, 2006), generalized functional linear regression (Reiss and Ogden, 2010), functional additive regression (Müller et al., 2013), functional adaptive models (James and Silverman, 2005), functional index regression (Chen et al., 2011), and semiparametric mixed normal transformation models (Zhong et al., 2021). Some methods have also been proposed to simultaneously model functional and scalar predictors (Lu et al., 2014; Kong et al., 2016; Wong et al., 2019). However, these works mainly focus on finite multivariate functional predictors.

Recently, Xue and Yao (2021) considered high-dimensional functional regression models by selecting important functional covariates under a linear model structure, and Fan et al. (2015) proposed a functional additive regression to flexibly model the nonlinear relationship between a response and high-dimensional functional predictors. However, the performance of these works still relies heavily on the sparsity assumption and the restricted eigenvalue condition. Moreover, these methods focus on the cumulative information of functional covariates and often require complete information for predictor functions, which may be infeasible in practice. Additionally, recovering whole random curves using parametric or nonparametric techniques can be challenging when the original observations are sparse or observed at irregular time points (Yao et al., 2005b; Li and Hsing, 2010). Furthermore, even when the entire curve is observed, applying functional regression directly to the whole functions may not be optimal with much noise, especially in high-dimensional scenarios.

To overcome these challenges, functional regression based on functional principal component analysis (FR-FPCA) has been developed (Zhu et al., 2014; Wong et al., 2019; Liu et al., 2021). FR-FPCA utilizes standard functional principal component analysis or spline approximation techniques to extract scores and performs regression on these scores. However, FR-FPCA focuses on the scores extracted from functional covariates, potentially ignoring important information regarding the relationship between the response and the covariates. Specifically, existing FR-FPCA methods for functional covariates or factor models for scalar covariates conduct regression on the functional principal scores $\boldsymbol{\zeta}_i$ or factors $\mathbf{F}_i$ using models such as $Y_i = g(\boldsymbol{\zeta}_i^q) + \varepsilon_i$ or $Y_i = g(\mathbf{F}_i^q) + \varepsilon_i$, where $\boldsymbol{\zeta}_i^q$ and $\mathbf{F}_i^q$ are the first q factors of

$\boldsymbol{\zeta}_i$ and $\mathbf{F}_i$, respectively, and $g(\cdot)$ represents various known or unknown functions. Although the first q factors are important for the functional and scalar covariates, they may not capture the full relationship between the response and the covariates, potentially overlooking crucial information.

To fix these problems, we propose a functional factor regression model (FFRM) by considering a multiple-index model $Y_i = \psi(\boldsymbol{\Omega}\mathbf{f}_i) + \varepsilon_i$, where $\mathbf{f}_i = (\boldsymbol{\zeta}_i^{\mathbf{T}}, \mathbf{F}_i^{\mathbf{T}})^{\mathbf{T}}$ includes all sufficient information from high-dimensional functional and scalar covariates, respectively. Here, $\boldsymbol{\Omega} \in \mathbb{R}^{d\times q}$ is the coefficient matrix, with both d and q allowed to diverge to infinity to capture sufficient information. The introduction of $\boldsymbol{\Omega}$ provides an opportunity to detect significant factors or directions by distinguishing between zero and nonzero columns of $\boldsymbol{\Omega}$. By excluding the zero columns of $\boldsymbol{\Omega}$, we can identify the important features of $\{\mathbf{X}_i(t), \mathbf{Z}_i\}$ that are relevant to the relationship between the response and UDFSM. To balance the prediction accuracy, stability, and interpretability, we consider the additive multiple-index structure for $\psi(\cdot)$, which serves as a universal approximator for any function when d is sufficiently large (Pinkus, 1999). Furthermore, to incorporate the correlation between $\boldsymbol{\zeta}_i$ and $\mathbf{F}_i$ arising from the dependence of functional and scalar covariates, we impose a low-rank structure on $\boldsymbol{\Omega}$, allowing for an explicit expression of the dependence between $\mathbf{X}_i(t)$ and $\mathbf{Z}_i$, as fully explained later.

To enhance the efficiency and flexibility of the estimators, we propose to estimate all component functions and parameters based on a penalized likelihood, even when the distribution of ε_i is unknown. As presented by our simulation studies, the proposed likelihood-based method outperforms the estimator based on the least square error (LSE), even when ε_i follows a normal distribution (see Figure 6(b,c)). The superior performance in Figure 6(b,c) can be attributed to the use of structural information, specifically the estimated density function $\widehat{f}(\cdot)$. This observation is consistent with findings in the literature (Zhou et al., 2019; Lin et al., 2021). In addition, we develop an iterative procedure that updates each parameter or function using existing packages, making computation and programming simple. We establish the selection and estimation consistency, as well as the asymptotic normality of the proposed estimators.

We apply the proposed method to analyze the data from the second-generation ALSPAC dataset. As demonstrated in Table 1, our method achieves superior prediction accuracy for LDL levels compared to existing approaches. The proposed FFRM identifies 424 SNPs significantly associated with LDL levels, many of which corroborate findings from previous research. Furthermore, we discover several anthropometric measures that substantially influence adult LDL levels. Notably, our analysis reveal considerable variation in how these characteristics affect LDL over ages. We observe that diverse anthropometric measures exhibit synchronized patterns in their influence on LDL levels, with both maximum and minimum effects occurring at similar ages.

We extend our analysis to examine the relationship between parental and individual characteristics and adult BMI. Furthermore, we apply the proposed method to analyze six different responses in ALSPAC, and Figure 5 demonstrates that our method provides higher prediction accuracy compared to existing methods for each response. In particular, the two ALSPAC analyses and simulation studies confirm that the proposed FFRM method outperforms shallow neural network (SNN) with sufficient width in terms of prediction accuracy.

The remainder of the paper is organized as follows. Section 2 introduces the FFRM. The estimation method is provided in Section 3. Section 4 applies the proposed method to analyze the ALSPAC data and presents the scientific findings and interpretations. In Section 5, we supplement additional analyses on the ALSPAC study. Furthermore, we evaluate the performance of the proposed estimation procedure through simulation studies in Sections

6. In Section 7, we provide a brief discussion of further research. The details on the theoretical result, along with some simulation and real data analysis results, are included in the Supplementary Materials (Suppl).

2. Model. Let Y_i denote the response variable, $\mathbf{X}_i(t) = \{X_{i1}(t), \cdots, X_{ip}(t)\}^{\mathbf{T}}$ represent the functional variables measured at various time points $(t_{i1}, \cdots, t_{in_i})$ prior to the measurement of Y_i, where n_i is the number of observations for the i-th individual (Zhang and Wang, 2016), and $\mathbf{Z}_i = (Z_{i1}, \cdots, Z_{im})^{\mathbf{T}}$ denote the scalar covariates. Observations $\{Y_i, \mathbf{X}_i(t), \mathbf{Z}_i\}$ $(i = 1, \cdots, n)$, are assumed to be independent and identically distributed. Particularly, in the ALSPAC data, for the i-th young adult, Y_i represents the LDL cholesterol level at age 24, $\mathbf{X}_i(t)$ are the anthropometrics measured at different ages prior to 24, and $\mathbf{Z}_i$ refer to the genotypes of the SNPs. For simplicity, we assume that the mean of $X_{ij}(t)$ and Z_{ij} has been subtracted, which means that $E\{X_{ij}(t)\} = 0$ and $E(Z_{ij}) = 0$ for any j and t.

Following Bai and Ng (2013), we assume that $\mathbf{Z}_i$ are correlated with the shared latent factors $\mathbf{F}_i$ and consider the following model

$$\mathbf{Z}_i = \mathbf{\Lambda}\mathbf{F}_i + \mathbf{e}_i, \tag{1}$$

where $\mathbf{F}_i$ is a q_1-dimensional vector with $q_1 \ll m$, $\mathbf{\Lambda}$ is a loading matrix, and $\mathbf{e}_i$ represents random errors that are independent of $\mathbf{F}_i$ with $E(\mathbf{e}_i) = \mathbf{0}$ and $\text{var}(\mathbf{e}_i) = \sigma_1^2\mathbf{I}_m$, where $\mathbf{I}_m$ denotes the m-dimensional identity matrix. We suppose Conditions (I1) and (I2) in Suppl. S6 to ensure model (1) is identifiable by following Bai and Ng (2013).

Similarly, we assume that $\mathbf{X}_i(t)$ are correlated due to sharing a vector of latent processes $\mathbf{h}_i(t)$ and can be modeled as

$$\mathbf{X}_i(t) = \mathbf{B}\mathbf{h}_i(t) + \mathbf{u}_i(t), \tag{2}$$

where $\mathbf{h}_i(t) = \{h_{i1}(t), \cdots, h_{iq_2}(t)\}^{\mathbf{T}}$ is a q_2-dimensional latent processes with $q_2 \ll p$, $\mathbf{B} = (\mathbf{b}_1, \cdots, \mathbf{b}_p)^{\mathbf{T}} = (b_{jk})_{p\times q_2}$ is a loading matrix, and $\mathbf{u}_i(t)$ represent random errors independent of $\mathbf{h}_i(t)$ with $E\{\mathbf{u}_i(t)\} = \mathbf{0}$ and $\text{cov}\{\mathbf{u}_i(t), \mathbf{u}_i(s)\} = \sigma_2^2 1_{\{t=s\}}\mathbf{I}_p$. By applying the Karhunen-Loève expansion (Ash and Gardner, 1975) to $h_{ij}(t)$, we suppose

$$h_{ij}(t) = \sum_{k=1}^{K} \xi_{ijk}\phi_{jk}(t), \tag{3}$$

where $\phi_{jk}(\cdot)$ is the k-th orthonormal eigenfunction for factor process j; ξ_{ijk} is the score with $E(\xi_{ijk}) = 0$ and $\text{cov}(\xi_{ijk}, \xi_{ijk'}) = \rho_{jk}1_{\{k=k'\}}$. Model (3) has been extensively studied in the literature of FPCA when K is fixed (Yao et al., 2005a; Zhou et al., 2018). To improve flexibility, we allow $K \to \infty$ and varies with n (Hall and Hosseini-Nasab, 2006).

Denote $\boldsymbol{\zeta}_i = (\boldsymbol{\zeta}_{i1}^{\mathbf{T}}, \cdots, \boldsymbol{\zeta}_{iq_2}^{\mathbf{T}})^{\mathbf{T}}$ with $\boldsymbol{\zeta}_{ij} = (\xi_{ij1}, \cdots, \xi_{ijK})^{\mathbf{T}}$ and $\mathbf{\Phi}(\cdot) = \text{diag}\{\mathbf{\Phi}_1(\cdot), \cdots, \mathbf{\Phi}_{q_2}(\cdot)\}$ with block j being $\mathbf{\Phi}_j(\cdot) = \{\phi_{j1}(\cdot), \cdots, \phi_{jK}(\cdot)\}^{\mathbf{T}}$. By the expression in (3), model (2) can be written as

$$\mathbf{X}_i(t) = \mathbf{B}\mathbf{\Phi}^{\mathbf{T}}(t)\boldsymbol{\zeta}_i + \mathbf{u}_i(t). \tag{4}$$

To ensure the identifiability of model (4), we impose Conditions (I3) and (I4) in Suppl. S6, as shown in Proposition 1 of the same Suppl. S6. Models (1) and (4) capture the heterogeneity of realizations $\{\mathbf{X}_i(t), \mathbf{Z}_i\}$ $(i = 1, \cdots, n)$, where the features within each realization $\{\mathbf{X}_i(t), \mathbf{Z}_i\}$ are fully determined by the latent factors $\mathbf{f}_i = (\boldsymbol{\zeta}_i^{\mathbf{T}}, \mathbf{F}_i^{\mathbf{T}})^{\mathbf{T}}$, provided that q_1, q_2, and K are sufficiently large. Features are extracted from $\{\mathbf{X}_i(t), \mathbf{Z}_i\}$ in an unsupervised manner based on models (1) and (4), which may introduce redundancy for modeling the relationship between Y_i and $\{\mathbf{X}_i(t), \mathbf{Z}_i\}$. To address this, we introduce $\mathbf{\Omega}$ to refine the features, emphasizing the relationship between Y_i and $\mathbf{f}_i$, leading to the following model for Y_i

$$Y_i = \psi(\mathbf{\Omega}\mathbf{f}_i) + \varepsilon_i, \tag{5}$$

to detect significant factors or directions by identifying zero collums of $\mathbf{\Omega}$. To achieve a balance between prediction accuracy, stability, and interpretability, we propose the following additive multiple-index model

$$Y_i = \sum_{j=1}^{d} \psi_j(\boldsymbol{\alpha}_j^{\mathbf{T}} \boldsymbol{\zeta}_i + \boldsymbol{\beta}_j^{\mathbf{T}} \mathbf{F}_i) + \varepsilon_i, \tag{6}$$

with $\mathbf{\Omega} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_d)^{\mathbf{T}}$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_d)^{\mathbf{T}}$ and divergent d, where both $\psi_j(\cdot)$ and error distribution f are unknown. In model (6), we suppose that $E\{\psi_j(\cdot)\} = 0$ for identifiability and the density function f satisfies the smooth Condition (C1) in Suppl. S4. In fact, additive multiple-index model is a universal approximator provided that d is sufficiently large (Pinkus, 1999). If we fix each component function $\psi_j(\cdot)$ to be a prespecified activation function, it reduces to a SNN with inputs $\{\mathbf{X}_i(t), \mathbf{Z}_i\}$, inferred factors $\mathbf{f}_i$, features $\mathbf{s}_i = \mathbf{\Omega}\mathbf{f}_i$, activation function ψ_j and output Y_i. Our simulation studies in Section 6 show that the data-driven activation function is helpful in improving prediction accuracy.

Furthermore, there is typically a correlation between the functional and scalar covariates, leading to the correlation between $\boldsymbol{\zeta}_i$ and $\mathbf{F}_i$. To incorporate the correlation between them, we impose a low-rank structure on their effects $\mathbf{\Omega} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$. That is, we assume $\mathbf{\Omega} = \mathbf{UV}$ with $\mathbf{U} = (\mathbf{U}_1, \cdots, \mathbf{U}_d)^{\mathbf{T}} \in \mathbb{R}^{d \times r}$, $\mathbf{V} = (\mathbf{V}_{[1]}, \cdots, \mathbf{V}_{[q]}) = (\mathbf{V}_1, \mathbf{V}_2) \in \mathbb{R}^{r \times q}$ and rank $r < \min(d, q)$, where $\mathbf{V}_{[k]}$ is the k-th column of the matrix $\mathbf{V}$, $q = q_1 + Kq_2$ denotes the dimension of features extracted from functional and scalar covariates, and $\mathbf{V}_1$ and $\mathbf{V}_2$ correspond to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively. With $\mathbf{\Omega} = (\boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathbf{UV}$, we can write $\boldsymbol{\alpha}_j = \mathbf{V}_1^{\mathbf{T}} \mathbf{U}_j$ and $\boldsymbol{\beta}_j = \mathbf{V}_2^{\mathbf{T}} \mathbf{U}_j$. It is clear that the terms $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_j$ are correlated with each other due to the shared component $\mathbf{U}_j$. This, in turn, captures the dependence between $\mathbf{X}_i(t)$ and $\mathbf{Z}_i$. The low-rank structure serves two main purposes. First, it addresses the challenge of modeling the correlation between functional and scalar covariates. Second, this low-rank structure reduces the dimensionality of the parameters in model (6), thereby improving the model's predictive performance and stability, as demonstrated in Table 1 for predicting LDL and Figure 5 for other responses.

Remark 1. Since the structure of $\text{var}\{\mathbf{u}_i(t)\}$ does not affect the identifiability and estimation of model (4), we assume $\text{var}\{\mathbf{u}_i(t)\} = \sigma_2^2 \mathbf{I}_p$ for simplicity. The diagonal structure can be relaxed to allow weak correlations among $\mathbf{u}_i(t)$. For example, there exists a constant $C > 0$ such that $\sum_{j'=1}^{p} \|E\{u_{ij}(t) u_{ij'}(t)\}\|_1 \leqslant C$ for each j and uniformly over t.

3. Estimation Procedure.

3.1. *Estimation of latent factors and scores.* Following Bai and Ng (2013), the estimation of $(\mathbf{\Lambda}, \mathbf{F})$ is a principal component problem, where $\mathbf{F} = (\mathbf{F}_1, \cdots, \mathbf{F}_n)^{\mathbf{T}}$. We hence directly estimate $\mathbf{F}$ by $\widehat{\mathbf{F}} = n^{1/2} \mathbf{E}_v(\mathbf{ZZ}^{\mathbf{T}}; q_1)$, where $\mathbf{Z} = (\mathbf{Z}_1, \cdots, \mathbf{Z}_n)^{\mathbf{T}}$ and $\mathbf{E}_v(A; c)$ is a matrix composed by the first c eigenvectors of matrix A.

The estimation of the model (4) involves three terms: $\mathbf{B}$, $\mathbf{\Phi}(\cdot)$ and $\boldsymbol{\zeta}_i$. However, it is not straightforward to use existing methods, such as those proposed by Bai and Ng (2013), to estimate $\boldsymbol{\zeta}_i$ due to the individual-dependent observation time. To avoid the need for iterative procedures, which often fail to converge without good initial values, Wen et al. (2025a) proposed a direct estimator for $\mathbf{B}$, $\mathbf{\Phi}(\cdot)$ and $\boldsymbol{\zeta}_i$. To calculate $\mathbf{\Phi}(\cdot)$, let $\mathbf{M}_1(\cdot) = \{M_{11}(\cdot), \cdots, M_{1,\tau_{1n}}(\cdot)\}^{\mathbf{T}}$ be a vector of B-spline basis functions on $[0, 1]$, then we have $\phi_{jk}(\cdot) \approx \mathbf{\Theta}_{jk}^{\mathbf{T}} \mathbf{M}_1(\cdot)$. After some computations (Wen et al., 2025a) on (4), $\mathbf{B}$ and $\boldsymbol{\zeta}_{[k]} = (\boldsymbol{\zeta}_{1k}, \cdots, \boldsymbol{\zeta}_{nk})^{\mathbf{T}}$ are estimated by the following closed-form expressions,

$$\widehat{\mathbf{B}} = p^{1/2} \mathbf{E}_v \left[n^{-1} \left\{ \sum_{i=1}^{n} n_i^{-1} \sum_{l=1}^{n_i} \mathbf{X}_i(t_{il}) \mathbf{X}_i^{\mathbf{T}}(t_{il}) \right\}; q_2 \right],$$

$$\widehat{\boldsymbol{\zeta}}_{[k]} = \tau_{1n}^{-1/2} \mathbf{W}_k \times \mathbf{E}_v(\mathbf{W}_k^{\mathbf{T}} \mathbf{W}_k; K) \quad (k = 1, \cdots, q_2),$$

where $\mathbf{W}_k = (\mathbf{w}_{1k}, \cdots, \mathbf{w}_{nk})^{\mathbf{T}}$ with $\mathbf{w}_{ik} = \{\sum_{l=1}^{n_i} \mathbf{M}_1(t_{il})\mathbf{M}_1^{\mathbf{T}}(t_{il})\}^{-1}\{\sum_{l=1}^{n_i} p^{-1} \sum_{j=1}^{p} \mathbf{M}_1(t_{il})\widehat{b}_{jk}X_{ij}(t_{il})\}$.

Remark 2. We require $n_i \geqslant \tau_{1n}$ to ensure invertibility of $\sum_{l=1}^{n_i} \mathbf{M}_1(t_{il})\mathbf{M}_1^{\mathbf{T}}(t_{il})$. Particularly, for a second-order Hölder continuous functions, we can take $\tau_{1n} = O(n^{1/5})$ to achieve the optimal convergence rate. Then $n_i = O(n^{1/5})$ is enough to ensure the invertibility of $\sum_{l=1}^{n_i} \mathbf{M}_1(t_{il})\mathbf{M}_1^{\mathbf{T}}(t_{il})$.

3.2. *Estimation of loadings and component functions.* Denote $\mathbf{\Psi}(\cdot) = \{\psi_1(\cdot), \cdots, \psi_d(\cdot)\}^{\mathbf{T}}$ and let $\mathbf{M}_2(\cdot) = \{M_{21}(\cdot), \cdots, M_{2,\tau_{2n}}(\cdot)\}^{\mathbf{T}}$ be a vector of B-spline basis functions, we have $\psi_j(\cdot) \approx \mathbf{a}_j^{\mathbf{T}}\mathbf{M}_2(\cdot)$. Here, we use the different basis functions for $\phi_{jk}(\cdot)$ and $\psi_j(\cdot)$ to ensure the convergence rate of $\widehat{\mathbf{\Omega}}$ and $\widehat{\mathbf{\Psi}}$, see details in Section S7. Define the sieve space as $\mathscr{F} = \prod_{j=1}^{d}\{\mathbf{a}_j^{\mathbf{T}}\mathbf{M}_2(u) : \mathbf{a}_j \in \mathbb{R}^{\tau_{2n}}, u \in [u_l,\ u_u]\}$. The features $\mathbf{f}_i$ are extracted from $\{\mathbf{X}_i(t), \mathbf{Z}_i\}$ in an unsupervised manner based on models (1) and (4), which may introduce redundancy in modeling the relationship between Y_i and $\{\mathbf{X}_i(t), \mathbf{Z}_i\}$. To effectively capture the relevant information from $\{\mathbf{X}_i(t), \mathbf{Z}_i\}$ for the response Y_i, we identify the important latent factors of $\mathbf{f}_i$ that contribute to Y_i using the sparsity assumption. The sparsity assumption also allows us to select larger dimensions (q_1, q_2, K), so that as much information as possible from $\{\mathbf{X}_i(t), \mathbf{Z}_i\}$ is retained. The identification of important latent factors of $\mathbf{f}_i$ is equivalent to detecting the zero columns in the coefficient matrix $\mathbf{\Omega} = (\mathbf{\Omega}_{[1]}, \cdots, \mathbf{\Omega}_{[q]})$ with $q = q_1 + Kq_2$. Given the low-rank structure $\mathbf{\Omega} = \mathbf{UV}$, it suffices to identify the zero columns of the matrix $\mathbf{V}$. To accomplish this, we apply a group penalty to the columns of matrix $\mathbf{V}$ and estimate $\boldsymbol{\gamma} = (\mathbf{U}, \mathbf{V}, \mathbf{\Psi})$ as follows

$$\widehat{\boldsymbol{\gamma}} = \underset{\mathbf{U},\mathbf{V},\mathbf{\Psi}\in\mathscr{F}}{\arg\max} \left\{ \ell_n(\boldsymbol{\gamma}; \widehat{\mathbf{f}}, \rho) - \lambda \sum_{k=1}^{q} w_k \|\mathbf{V}_{[k]}\|_2 \right\}, \tag{7}$$

where $\ell_n(\boldsymbol{\gamma}; \mathbf{f}, \rho) = n^{-1}\sum_{i=1}^{n} \ell(\boldsymbol{\gamma}; Y_i, \mathbf{f}_i, \rho) = n^{-1}\sum_{i=1}^{n} \rho\{Y_i - \sum_{j=1}^{d} \psi_j(\mathbf{U}_j^{\mathbf{T}}\mathbf{V}\mathbf{f}_i)\}$ and $\mathbf{f} = \{\mathbf{f}_1, \cdots, \mathbf{f}_n\}$, λ is a tuning parameter, w_k is a known weight such as that for adaptive LASSO (Zou, 2006), $\|\mathbf{w}\|_2$ is the L_2-norm of vector $\mathbf{w}$, and $\rho(\cdot)$ is a given loss function; for example, the least squares loss $\rho(t) = -t^2$, which is the most commonly used, is optimal when the error ε_i is normally distributed but inefficient when the normality is violated. Some adjustments have been suggested, for example, the kernel M-smoother, median smoothing, locally weighted regression, and the local least absolute method. These methods are mathematically convenient, robust, and easily implemented. However, they are suboptimal. If the density function f is known, an ideal choice of the loss function $\rho(\cdot)$ is $\log f(\cdot)$ (Stone, 1975), which is the likelihood function. In this paper, to enhance efficiency, we propose a maximum likelihood estimator for $\boldsymbol{\gamma}$ even when $f(\cdot)$ is unknown. The key idea is to replace the unknown density function $f(\cdot)$ by an estimated density function, $\widehat{f}(\cdot)$. The density function can be estimated by the Nadaraya-Watson kernel: $\widehat{f}(y) = n^{-1}\sum_{i=1}^{n} \mathcal{K}_h\{Y_i - \sum_{j=1}^{d} \psi_j(\mathbf{U}_j^{\mathbf{T}}\mathbf{V}\mathbf{f}_i) - y\}$, where $\mathcal{K}_h(\cdot) = \mathcal{K}(\cdot/h)/h$, $\mathcal{K}(\cdot)$ is a nonnegative symmetric kernel function with support on $[-1, 1]$ and h is a bandwidth. By replacing the loss function $\rho(\cdot)$ in (7) with $\log\widehat{f}(\cdot)$, we can estimate $\boldsymbol{\gamma} = (\mathbf{U}, \mathbf{V}, \mathbf{\Psi})$ by the following penalized likelihood

$$\widehat{\boldsymbol{\gamma}} = \underset{\mathbf{U},\mathbf{V},\mathbf{\Psi}\in\mathscr{F}}{\operatorname{argmax}} \left\{ \ell_n(\boldsymbol{\gamma}; \widehat{\mathbf{f}}, \log\widehat{f}) - \lambda \sum_{k=1}^{q} w_k \|\mathbf{V}_{[k]}\|_2 \right\}. \tag{8}$$

3.3. *Algorithm to compute (8).* Since the penalty term is independent of $\mathbf{U}$ and $\mathbf{\Psi}$, in the $(t+1)$-th step, we update $\mathbf{U}$ and $\mathbf{\Psi}$ with a gradient algorithm, that is,

$$\begin{aligned} \mathbf{U}^{(t+1)} &= \mathbf{U}^{(t)} + \eta \bigtriangledown_{\mathbf{U}} \ell_n(\boldsymbol{\gamma}^{(t)}; \widehat{\mathbf{f}}, \log\widehat{f}), \\ \mathbf{\Psi}^{(t+1)}(\cdot) &= \mathbf{\Psi}^{(t)}(\cdot) + \eta \bigtriangledown_{\mathbf{A}} \ell_n(\boldsymbol{\gamma}^{(t)}; \widehat{\mathbf{f}}, \log\widehat{f})\mathbf{M}_2(\cdot), \end{aligned} \tag{9}$$

where $\mathbf{A} = (\mathbf{a}_1, \cdots, \mathbf{a}_d)^{\mathbf{T}}$, $\bigtriangledown_{\mathbf{c}}$ denotes the derivative of $\ell_n(\boldsymbol{\gamma}; \widehat{\mathbf{f}}, \log \widehat{f})$ with respect to $\mathbf{c}$, where $\mathbf{c}$ may represent a matrix or a vector, and η is a constant and may be different in different places.

To deal with the nonsmoothness of the penalty, we use the algorithm proposed by Noah et al. (2013) to update $\mathbf{V}$. In the $(t+1)$-th step, we majorize the negative loss function by $-\ell_n(\mathbf{V}_{[k]}, \boldsymbol{\gamma}^{(t)}_{-\mathbf{V}_{[k]}}; \widehat{\mathbf{f}}, \log \widehat{f}) \leqslant -\ell_n(\boldsymbol{\gamma}^{(t)}; \widehat{\mathbf{f}}, \log \widehat{f}) - (\mathbf{V}_{[k]} - \mathbf{V}^{(t)}_{[k]})^{\mathbf{T}} \bigtriangledown_{\mathbf{V}_{[k]}} \ell_n(\boldsymbol{\gamma}^{(t)}; \widehat{\mathbf{f}}, \log \widehat{f}) + \|\mathbf{V}_{[k]} - \mathbf{V}^{(t)}_{[k]}\|_2^2/(2\eta)$, where $\ell_n(\mathbf{V}_{[k]}, \boldsymbol{\gamma}^{(t)}_{-\mathbf{V}_{[k]}}; \widehat{\mathbf{f}}, \log \widehat{f})$ is $\ell_n(\boldsymbol{\gamma}; \widehat{\mathbf{f}}, \log \widehat{f})$ with $\boldsymbol{\gamma}_{-\mathbf{V}_{[k]}}$ (e.g., $\boldsymbol{\gamma}$ excluding $\mathbf{V}_{[k]}$) replaced by the estimators $\boldsymbol{\gamma}^{(t)}_{-\mathbf{V}_{[k]}}$ from the t-th step, and η is a sufficiently small constant so that the quadratic term dominates the Hessian matrix of the loss function. Considering the penalty term, we estimate $\mathbf{V}_{[k]}$ by minimizing $-\ell_n(\boldsymbol{\gamma}^{(t)}; \widehat{\mathbf{f}}, \log \widehat{f}) - (\mathbf{V}_{[k]} - \mathbf{V}^{(t)}_{[k]})^{\mathbf{T}} \bigtriangledown_{\mathbf{V}_{[k]}} \ell_n(\boldsymbol{\gamma}^{(t)}; \widehat{\mathbf{f}}, \log \widehat{f}) + \|\mathbf{V}_{[k]} - \mathbf{V}^{(t)}_{[k]}\|_2^2/(2\eta) + \lambda w_k \|\mathbf{V}_{[k]}\|_2$, which is equivalent to minimizing $\|\mathbf{V}_{[k]} - \{\mathbf{V}^{(t)}_{[k]} + \eta \bigtriangledown_{\mathbf{V}_{[k]}} \ell_n(\boldsymbol{\gamma}^{(t)}; \widehat{\mathbf{f}}, \log \widehat{f})\}\|_2^2/(2\eta) + \lambda w_k \|\mathbf{V}_{[k]}\|_2$. Then, we update $\mathbf{V}^{(t)}_{[k]}$ $(k = 1, \cdots, q)$ one-by-one by

(10)

$$\mathbf{V}^{(t+1)}_{[k]} = \begin{cases} 0, & \|\mathbf{V}^{(t)}_{[k]} + \eta \bigtriangledown_{\mathbf{V}_{[k]}} \ell_n(\boldsymbol{\gamma}^{(t)}; \widehat{\mathbf{f}}, \log \widehat{f})\|_2 = 0, \\ \left\{1 - \frac{\eta \lambda w_k}{\|\mathbf{V}_{[k]} + \eta \bigtriangledown_{\mathbf{V}_{[k]}} \ell_n(\boldsymbol{\gamma}^{(t)}; \widehat{\mathbf{f}}, \log \widehat{f})\|_2}\right\}_+ \left\{\mathbf{V}_{[k]} + \eta \bigtriangledown_{\mathbf{V}_{[k]}} \ell_n(\boldsymbol{\gamma}^{(t)}; \widehat{\mathbf{f}}, \log \widehat{f})\right\}, & \text{otherwise,} \end{cases}$$

where $(a)_+ = \max(0, a)$. We repeat (9) and (10) until convergence. The selections of initial values and tuning parameters are discussed in Suppl. S1.

4. Analysis of the LDL with the ALSPAC data. We applied our proposed FFRM to analyze LDL levels at age 24 using the ALSPAC dataset. The analysis incorporated both functional and scalar covariates. The functional covariates comprised 11 primary variables measured from ages 7 to 17, including anthropometric measures (height, weight, sitting height, waist circumference, arm circumference, and BMI), body composition indicators (fat percentage and body water), scoliometer readings, axis of astigmatism in the left eye, and impedance measurements. To account for nonlinear relationships, we included squared terms and interaction effects of these functional variables, expanding the functional covariate set to 77 variables. The scalar covariates consisted of genotype data from 455,395 SNPs (including rs4420638, rs11206510, and rs10411594) distributed across the 23 pairs of human chromosomes.

Prior to analysis, we conducted quality control (QC) by excluding individuals with missing functional covariate data, resulting in a final sample size of 1,077 individuals. As illustrated in Figure 1 (left), we analyzed the principal components of $n^{-1} \sum_{i=1}^{n} n_i^{-1} \sum_{l=1}^{n_i} \mathbf{X}_i(t_{il}) \mathbf{X}_i^{\mathbf{T}}(t_{il})$, selecting 30 components for initial examination. Following the criteria detailed in Suppl. S1, we observed that the eigenvalues plateaued at 17 principal components, corresponding to an explained variance ratio of 90.73%. Consequently, we set $q_2 = 17$. Similarly, based on criteria outlined in the same supplementary section, we established $K = 3$.

The matrix $\mathbf{Z}$ contains $\min\{m, n\} = 1077$ singular values. While conventional selection of q_1 based on explained variance proportion would result in an excessive number of factors, we implemented an alternative selection procedure. We computed the relative eigenvalue difference for the j-th eigenvalue as: $d_j = \{\lambda_j(n^{-1}\mathbf{Z}\mathbf{Z}^{\mathbf{T}}) - \lambda_{j-1}(n^{-1}\mathbf{Z}\mathbf{Z}^{\mathbf{T}})\}/\lambda_{j-1}(n^{-1}\mathbf{Z}\mathbf{Z}^{\mathbf{T}})$ with $d_1 = 0$. As illustrated in Figure 1 (right), we initially examined 50 principal components of $n^{-1}\mathbf{Z}\mathbf{Z}^{\mathbf{T}}$. The eigenvalue difference approached zero at 20 principal components, leading us to set $q_1 = 20$. For the final implementation of FFRM, we specified $d = 2, r = 1$ and $\lambda = 0.01$, and employed a bandwidth of $h = n^{-1/3} \approx 0.09$.

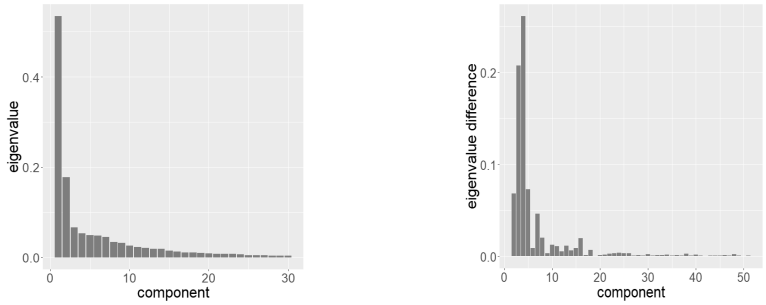


Fig 1: (left): the eigenvalues for the first 30 principal components of $n^{-1}\sum_{i=1}^{n} n_i^{-1}\sum_{l=1}^{n_i}\mathbf{X}_i(t_{il})\mathbf{X}_i^{\mathbf{T}}(t_{il})$; (right): the eigenvalue differences for the first 50 principal components of $n^{-1}\mathbf{Z}\mathbf{Z}^{\mathbf{T}}$.

TABLE 1
PEs and the corresponding standard deviations (reported in parentheses) of LDL

	FFRM	PFLM	PLFAM	SNN
PE	0.6918 (0.1059)	0.7334 (0.1032)	0.7380 (0.1044)	0.7625 (0.1078)

We evaluated the predictive performance of our proposed FFRM method against existing methods, including the partially functional linear regression model (PFLM; Kong et al. 2016), partially linear functional additive model (PLFAM; Wong et al. 2019), and SNN. The SNN architecture parallels FFRM's structure but employs the $\tanh$ function for $\psi_j(\cdot)$ in equation (6). Using optimized tuning parameters, we performed 200 iterations of random data partitioning, with 90% allocated to training and 10% to testing. For SNN, we selected the number of neurons that minimized prediction error, while optimal tuning parameters were chosen for PFLM and PLFAM. Table 1 presents the prediction error (PE), defined as: PE $= \sum_{i\in\mathcal{S}_T}(\widehat{Y}_i - Y_i)^2 / \sum_{i\in\mathcal{S}_T}(Y_i - \bar{Y})^2$, where $\mathcal{S}_T$ denotes the test set. Results demonstrated that FFRM achieved superior predictive performance compared to SNN, PFLM, and PLFAM.

By the identifiability Condition (I1) in Suppl. S6, we had $\mathbf{U}_1 = \mathbf{U}_2$ and $(\boldsymbol{\alpha}_1^{\mathbf{T}}, \boldsymbol{\beta}_1^{\mathbf{T}}) = (\boldsymbol{\alpha}_2^{\mathbf{T}}, \boldsymbol{\beta}_2^{\mathbf{T}})$ when $(d, r) = (2, 1)$. This indicated that the low-rank structure identified only one direction for the effect of $\mathbf{f}_i$ on Y_i, which we denoted by $(\boldsymbol{\alpha}, \boldsymbol{\beta})$. Define $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_1, \ldots, \widetilde{\beta}_m) = \widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\Lambda}}^{\mathbf{T}}\widehat{\boldsymbol{\Lambda}})^{-1}\widehat{\boldsymbol{\Lambda}}^{\mathbf{T}}$. Following Jiang et al. (2019), a simple calculation showed that $\widehat{\boldsymbol{\beta}}\mathbf{F}_i \approx \widetilde{\boldsymbol{\beta}}\mathbf{Z}_i$, implying that $\widetilde{\boldsymbol{\beta}}$ represented the regression coefficients of the SNPs and could be used to quantify their impacts on LDL. We then performed hypothesis testing for each scalar covariate j (i.e., testing $\widetilde{\beta}_j = 0$). To do so, we proposed to use p-values while controlling the false discovery rate. More specifically, we computed the p-values as $p_j = 2[1 - \Phi\{\|\widetilde{\beta}_j\|_1 / \mathrm{SD}(\widetilde{\beta}_j)\}]$, where Φ denoted the cumulative distribution function of the standard normal distribution. Then, following Benjamini and Hochberg (1995), we ordered the p-values as $p_{(1)} \leqslant \cdots \leqslant p_{(m)}$, corresponding to each of the null hypotheses, $H_{0(i)}, i = 1, \ldots, m$. We then identified the threshold $i = \max\{j : p_{(j)} \leqslant 0.05j/m\}$ and rejected all null hypotheses associated with the first i p-values. The analysis, based on 200 bootstrap replications, identified 424 significant SNPs, with key findings presented in Table 2. The majority of these SNPs demonstrated positive associations with LDL levels, suggesting their variants may contribute to elevated LDL concentrations and consequently increase IHD and stroke risk. Our findings corroborated previous research by Sandhu et al. (2008), confirming the association of SNPs, such as rs693, rs1713222, and rs10402271, with LDL levels.

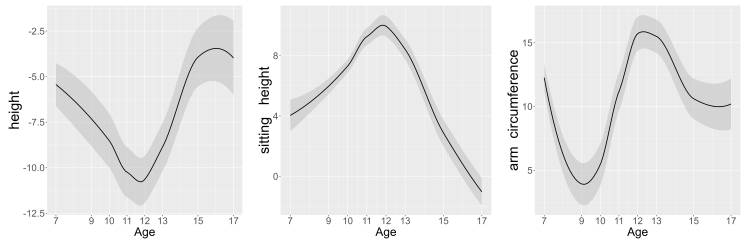
TABLE 2
Estimates, SDs and p-values for $\widetilde{\boldsymbol{\beta}}$ with LDL response using FFRM

	Esitimate	SD	p value
rs6511720	-1.7498	0.5361	1.0978×10^{-3}
rs693	1.7863	0.4063	1.0980×10^{-5}
rs1713222	1.5173	0.3920	1.0843×10^{-4}
rs10402271	1.2598	0.3156	6.5522×10^{-5}
rs1042031	1.4366	0.4315	8.7151×10^{-4}
rs1412444	1.1401	0.3929	3.7112×10^{-3}
rs4803750	1.1233	0.3505	1.3507×10^{-2}
rs646776	1.4138	0.3940	3.3348×10^{-4}
rs1713222	1.5173	0.3920	1.0843×10^{-4}

Based on the theoretical justification in Suppl. S11, we had $p^{-1}\int \widehat{\boldsymbol{\alpha}}\boldsymbol{\Phi}(t)\mathbf{B}^{\mathbf{T}}\mathbf{X}_i(t)dt \approx \widehat{\boldsymbol{\alpha}}\int \boldsymbol{\Phi}(t)\boldsymbol{\Phi}^{\mathbf{T}}(t)dt\boldsymbol{\zeta}_i = \widehat{\boldsymbol{\alpha}}\boldsymbol{\zeta}_i$, which implied that $p^{-1}\widehat{\mathbf{B}}\widehat{\boldsymbol{\Phi}}^{\mathbf{T}}(t)\widehat{\boldsymbol{\alpha}}^{\mathbf{T}}$ was the regression coefficient function of $\mathbf{X}_i(t)$ and could be used to investigate the impact of the functional effects of anthropometric measures on LDL levels. Figures 2 and 3 present the estimated coefficient functions for selected functional covariates, accompanied by confidence bands derived from 200 bootstrap replications. Our analysis revealed no significant associations between LDL levels and several variables, including weight, impedance measurements, scoliometer readings, and the axis of astigmatism in the left eye.

Among anthropometric measures, height, sitting height, and arm circumference (Figures 2 (a)-(c)) – commonly used indicators of growth, body proportions, and skeletal-muscular development – showed distinct associations with LDL levels. Height demonstrated a negative correlation with LDL levels, consistent with Fujita et al. (2016) findings that LDL cholesterol decreases with increasing height during puberty, independent of weight gain, as shown in their three-year follow-up study in Fukuroi. Conversely, both sitting height and arm circumference exhibited positive associations with LDL levels, corroborating previous research by (Schooling et al., 2007; Zhu et al., 2020), who reported higher LDL levels among individuals with greater sitting height and mid-upper arm circumference. Notably, the effects of all three anthropometric measures on LDL peaked at age 12, highlighting the crucial role of pubertal changes in body composition and fat distribution in determining blood lipid profiles.

Waist circumference, BMI, fat percentage, and body water content (Figures 3 (a)-(d)) served as critical indicators of body composition, weight status, and health risk. Our analysis revealed positive associations between LDL levels and waist circumference, BMI, and



(a) (b) (c)

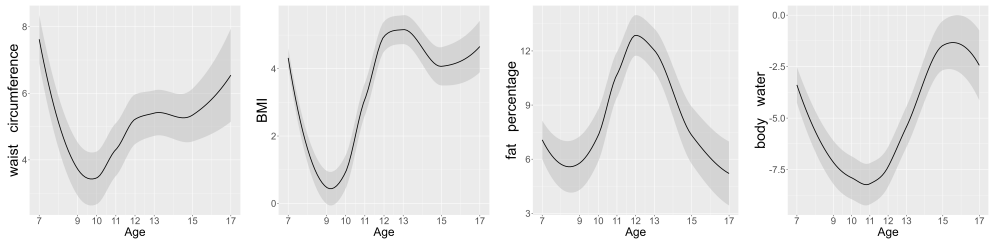
Fig 2: The estimated coefficient function curves corresponding to (a) Height, (b) Sitting Height, (c) Arm Circuference.

fat percentage, while body water content showed an inverse relationship. These findings corroborated previous research (Maffeis et al., 2012; Shirasawa et al., 2013; Oda, 2013; Liu et al., 2023). Notably, the timing of peak influence varied among these anthropometric measures: waist circumference exhibited maximum effect around age 7, while body water content, BMI, and fat percentage showed strongest associations between ages 11 and 13. These temporal patterns suggested the need for age-specific interventions: early childhood (before age 7) should focus on preventing abdominal fat accumulation through dietary modifications, particularly reducing sugary beverage and high-fat food consumption. During puberty, careful monitoring of body composition becomed crucial for obesity prevention, while adolescent health management should emphasize maintaining proper fluid and electrolyte balance through controlled salt and sugar intake.

Our analysis revealed consistent patterns in the ages at which anthropometric measures exhibited minimal influence on LDL levels. Specifically, arm circumference, waist circumference, BMI, and body water content showed their weakest associations between ages 9 and 11, while height and sitting height demonstrated minimal impact at ages 16-17. Fat percentage uniquely displayed reduced influence during both intervals. These temporal patterns align with known developmental phases: children aged 9-10 years typically experience a period of relative metabolic stability during their transition from early childhood to adolescence, resulting in diminished associations between body composition measures and LDL levels. Similarly, by ages 16-17, as physiological development approaches maturity and hormonal profiles stabilize, the influence of anthropometric measures on LDL levels becomes notably reduced.

5. Analysis of the BMI with the ALSPAC data. We further applied FFRM to investigate a critical question in the ALSPAC study, specifically identifying key scalar and functional predictors of adult BMI. To proceed, we set the response variable to be the BMI of the participants at age 24, and included as the covariates 77 functional (and historical) variables, such as the anthropometrics of the individuals (e.g., height and weight), lipid levels (e.g., fat percentage, body water) and their squared or interaction terms, measured between ages 7 and 24. Additionally, 88 scalar covariates, such as maternal and paternal anthropometrics and blood pressure, were included in the analysis. Supplementary Tables S2-S4 provide summary information on these covariates.

We first performed QC by removing individuals with missing values in the scalar and functional covariates. In total, we analyzed 348 individuals. As shown in Supplementary Figure S1, we selected the number of principal components of $n^{-1}\mathbf{Z}\mathbf{Z}^{\mathbf{T}}$ and $n^{-1}\sum_{i=1}^{n} n_i^{-1}\sum_{l=1}^{n_i}\mathbf{X}_i(t_{il})\mathbf{X}_i^{\mathbf{T}}(t_{il})$ to 30, respectively. Based on the criteria outlined in the



(a) (b) (c) (d)

Fig 3: The estimated coefficient function curves corresponding to (a) Waist circumference, (b) BMI, (c) Fat percentage, (d) Body Water.

Suppl. S1, we set $q_1 = 20$, $q_2 = 13$, $K = 2$, $\lambda = 0.001$, $d = 1$, and $r = 1$. Using a bandwidth of $h = n^{-1/3} \approx 0.15$, we estimated $\mathbf{\Omega}$ using the proposed FFRM.

To investigate the effects of parental anthropometrics, blood pressure, and other scalar covariates on response, similar to Section 4, for each scalar covariate j, we performed a hypothesis test of $\widetilde{\beta}_j = 0$ by evaluating its p-value. We identified a total of 28 significant scalar covariates as shown in Table 3. Among them, the positive effects of parental BMI (fm1ms111, ff1ms111) on the BMI of their offspring at 24 years of age were identified by FFRM, which was consistent with positive association between parental BMI and offspring BMI at age 7.5 based on a study using ALSPAC data (George et al., 2007). Furthermore, both maternal and paternal weight (fm1ms110, ff1ms110) and height (fm1ms100, ff1ms100) were significantly identified in positive and negative associations, respectively, with their offspring BMI at age 24.

FFRM also significantly identified the negative association of childbearing age (fm1a011) and preterm delivery (DELP1010) and the positive association of hemoglobin (DELP1047) and gestational weight gain (DELP1128, DELP1129), respectively, with the BMI of the offspring at age 24. Previous studies reported that premature delivery and anemia during pregnancy are the primary risk factors for low birth weight (Rasmussen, 2001), while low birth weight was reported to be significantly associated with increased maternal age in a study of individuals aged 20 to over 40 (Callaway et al., 2005). More recently, a significant association between birth weight and BMI was reported at ages 9 and 7 (Simpson et al., 2017). All this evidence suggests that premature delivery, anemia, and increasing maternal age could lead to a lower BMI of the offspring through a lower birth weight. Furthermore, a systematic review of 15 observational studies revealed a significant positive effect of gestational weight gain on birth weight (Bodnar et al., 2014).

To assess the effects of individual anthropometrics and other assay results on BMI at age 24, similar to Section 4, we used the coefficient functions $p^{-1}\widehat{\mathbf{B}}\widehat{\mathbf{\Phi}}^{\mathbf{T}}(t)\widehat{\boldsymbol{\alpha}}^{\mathbf{T}}$. The estimated coefficient functions for nine selected functional covariates along with the confidence bands based on 200 bootstrap replications are shown in Figure 4, suggesting a significant positive association of weight, fat percentage, cholesterol, arm circumference and systolic pressure and a negative assocation of height and high-density lipoprotein (HDL) in BMI at age 24. These findings were consistent with the existing literature on positive correlations between

TABLE 3
Estimates, SDs and p-values for $\widetilde{\boldsymbol{\beta}}$ with BMI response using FFRM

	Esitimate	SD	p value		Esitimate	SD	p value
fm1a011	-0.0328	0.0066	7.50×10^{-7}	fm1ms100	-0.2410	0.0492	9.68×10^{-7}
fm1ms110	0.4177	0.0209	$< 1 \times 10^{-15}$	fm1ms111	0.5315	0.0220	$< 1 \times 10^{-15}$
fm1ms115	0.4513	0.0170	$< 1 \times 10^{-15}$	fm1ms125	0.4490	0.0174	$< 1 \times 10^{-15}$
fm1dx020	0.4629	0.0184	$< 1 \times 10^{-15}$	fm1dx021	0.1341	0.0263	3.37×10^{-7}
fm1dx391	0.4289	0.0170	$< 1 \times 10^{-15}$	fm1bp112	-0.1197	0.0167	7.78×10^{-13}
fm1bp122	-0.1228	0.0165	1.04×10^{-13}	fm1bp132	-0.1240	0.0170	2.89×10^{-13}
DELP1010	-0.0251	0.0030	$< 1 \times 10^{-15}$	DELP1047	0.0358	0.0041	$< 1 \times 10^{-15}$
DELP1128	0.0451	0.0021	$< 1 \times 10^{-15}$	DELP1129	0.0425	0.0112	1.51×10^{-4}
ff1ms100	-0.0885	0.0203	1.35×10^{-5}	ff1ms110	0.1927	0.0153	$< 1 \times 10^{-15}$
ff1ms111	0.2598	0.0183	$< 1 \times 10^{-15}$	ff1ms120	0.2183	0.0184	$< 1 \times 10^{-15}$
ff1ms125	0.2057	0.0141	$< 1 \times 10^{-15}$	ff1dx020	0.2479	0.0164	$< 1 \times 10^{-15}$
ff1dx030	0.0401	0.0102	8.61×10^{-5}	ff1bp140	-0.2050	0.0206	$< 1 \times 10^{-15}$
ff1bp141	-0.2126	0.0204	$< 1 \times 10^{-15}$	ff1bp142	-0.0253	0.0057	7.90×10^{-6}
ff1bp143	-0.1938	0.0209	$< 1 \times 10^{-15}$	ff1bp144	-0.2025	0.0208	$< 1 \times 10^{-15}$

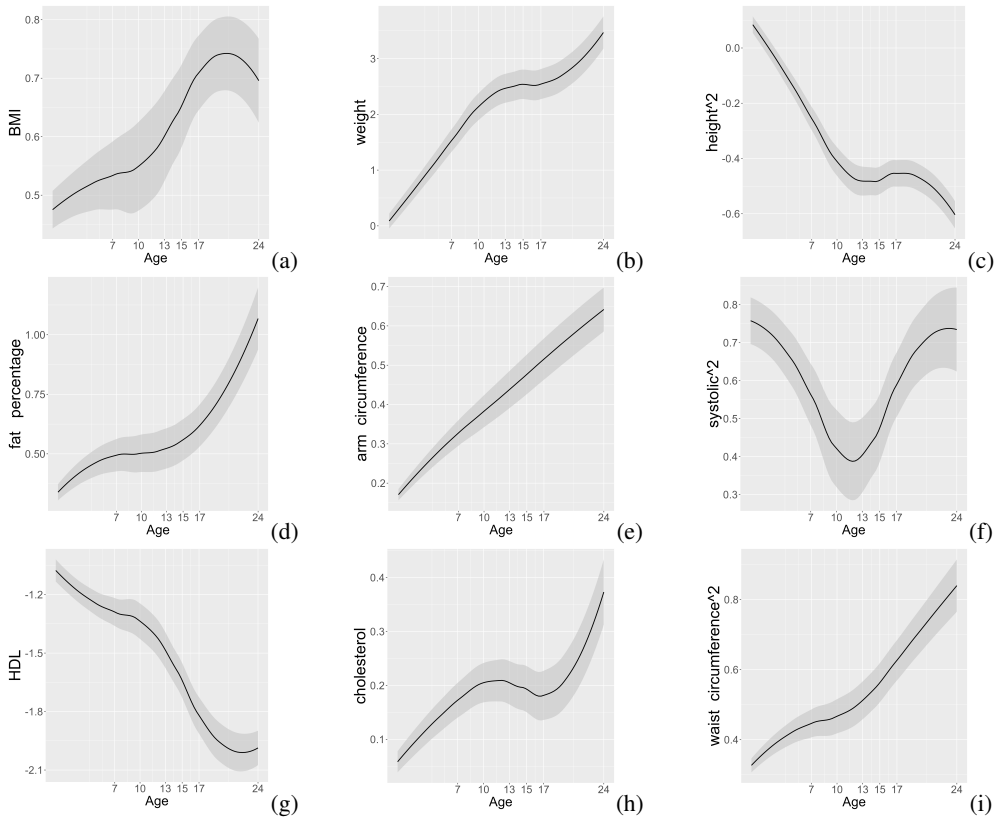


Fig 4: The estimated coefficient function curves corresponding to (a) BMI, (b) Weight, (c) Height^2, (d) Fat Percentage, (e) Arm Circumference, (f) Systolic^2, (g) HDL, (h) Cholesterol, (i) Waist Circumference^2.

body fat percentage, mid-upper arm circumference, and hypertension with BMI (Tang et al., 2013; Demarco et al., 2014) and the negative association between higher BMI and HDL (Shamai et al., 2009).

We further evaluated the prediction performance of the proposed FFRM, PFLM, PLFAM, and SNN on six different response variables, i.e., (a) BMI, (b) average systolic pressure, (c) average diastolic pressure, (d) cortical density (Tibia) (mg/cm^3), (e) cortical density tibia regression correction derived from area and content (mg/cm^3), and (f) subcortical density (Fibula) (mg/cm^3). For each response variable, we first selected the tuning parameters λ, d, r, and h as those for BMI. Using the selected tuning parameters, we divided the dataset into a training set (90% of the data) and a test set (10% of the data) and repeated this process 200 times in each scenario. For SNN, we chose the number of neurons that minimizes the prediction error.

Figure 5 shows the boxplots of the PE for each response. The results demonstrate that our proposed FFRM consistently outperformed SNN, PFLM, and PLFAM in all six responses. In particular, in Figures 5(e) and 5(f), the PE of PFLM exceeded 1, suggesting that PFLM performed even worse than a null model in some cases. Similarly, in Figures 5(c) and 5(f), the PE of PLFAM was also greater than 1, suggesting poor performance compared with a null model. Possible reasons for these observations are as follows (1) the average number of observations per response is around 4, which might be insufficient to provide accurate estimations of uFPCA and mFPCA scores; (2) in both PFLM and PLFAM, the part of regression for the scalar covariates is assumed to be linear, which might not capture the nonlinearity between Y_i and

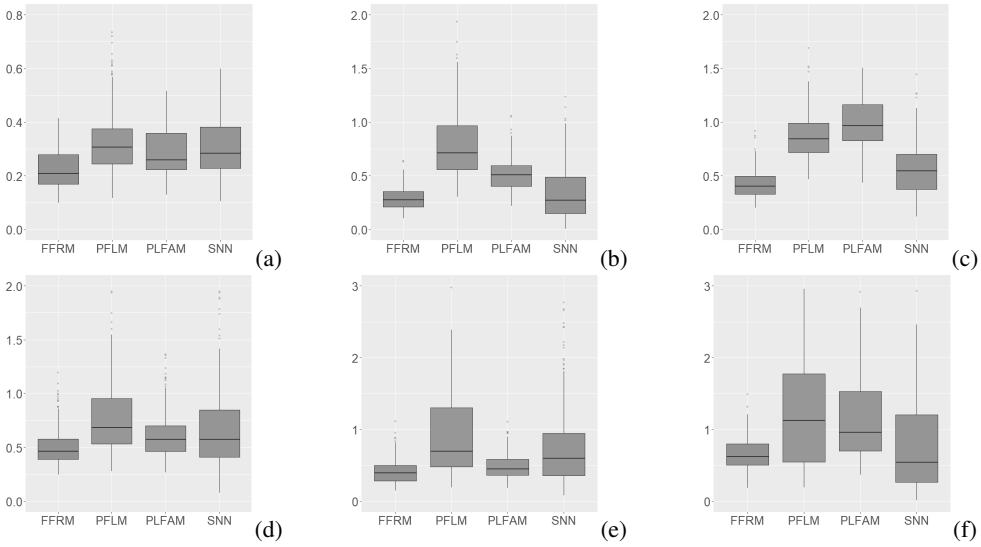


Fig 5: The boxplot of PE using the proposed FFRM, SNN, PFLM and PLFAM for ALSPAC datasets with different responses: (a) BMI, (b) Systolic Pressure, (c) Diastolic Pressure, (d) Cortical Density (Tibia) (e) Cortical Density Tibia Regression Correction derived from area and content, (f) SubCortical Density (Fibula).

$\mathbf{Z}_i$ adequately; (3) Supplementary Figure S1 shows that we can extract the information from the covariates using few latent factors with $(q_1, q_2, K) = (20, 13, 2)$, supporting the existence of low-rank structures in $\mathbf{X}_i(t)$ and $\mathbf{Z}_i$; and (4) both methods may overlook the specific correlation structure in the data. In each scenario, the number of activation functions in SNN is larger than that of component functions in FFRM and FFRM performs better than SNN, which shows the estimated component functions efficiently reduce the number of neurons and improve stability.

6. Numerical Studies. In this section, we evaluate the performance of the proposed FFRM method compared with existing methods, including PFLM, PLFAM and SNN regression model, which is denoted by d-SNN. We also make comparisons with the least square error (LSE) criterion to evaluate the increase in efficiency using a likelihood criterion. We evaluate the performance of the estimators in terms of $\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}\|_F$ for any matrix $\mathbf{\Omega}$ and use the bias, standard deviation (SD), and root mean square error (RMSE), which are defined as Bias $= (n_{grid}^{-1} \sum_{i=1}^{n_{grid}} [E\{\widehat{g}(x_i)\} - g(x_i)]^2)^{1/2}$, SD $= (n_{grid}^{-1} \sum_{i=1}^{n_{grid}} E[\widehat{g}(x_i) - E\{\widehat{g}(x_i)\}]^2)^{1/2}$, and RMSE $= (\text{Bias}^2 + \text{SD}^2)^{1/2}$ for any function $g(\cdot)$, where $x_1, \cdots, x_{n_{grid}}$ are the grid points and $E\{\widehat{g}(x_i)\}$ is approximated using the sample mean of the simulated datasets. In our simulation, we set $n_{grid} = 100$. We evaluate the prediction performance using the PE. In fairness, we considered four examples. Among them, one example is for the case where the assumptions of all the methods were not satisfied, while the other three examples satisfy the model assumptions of FFRM, PFLM and PLFAM, respectively.

6.1. *Simulation setting.* We first generated $\mathbf{f}_i = (\boldsymbol{\zeta}_i^{\mathbf{T}}, \mathbf{F}_i^{\mathbf{T}})^{\mathbf{T}}$ from $N(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{\boldsymbol{\zeta}} & \mathbf{\Sigma}_{\boldsymbol{\zeta},\mathbf{F}} \\ \mathbf{\Sigma}_{\boldsymbol{\zeta},\mathbf{F}}^{\mathbf{T}} & \mathbf{\Sigma}_{\mathbf{F}} \end{pmatrix}$, where $\mathbf{\Sigma}_{\boldsymbol{\zeta}} = (\sigma_{kj}^{(1)})_{Kq_2 \times Kq_2}$ with $\sigma_{kj}^{(1)} = 1/k \cdot 1_{\{k=j\}}$; $\mathbf{\Sigma}_{\mathbf{F}} = (\sigma_{kj}^{(2)})_{q_1 \times q_1}$ with $\sigma_{kj}^{(2)} = 0.5^{\|k-j\|_1}$; and $\mathbf{\Sigma}_{\boldsymbol{\zeta},\mathbf{F}} = (\sigma_{kj}^{(3)})_{Kq_2 \times q_1}$ with $\sigma_{kj}^{(3)} = 0.2^{\|k-j\|_1+1}$.

Given $\mathbf{F}_i$, we generated $\mathbf{Z}_i$ by $\mathbf{Z}_i = \mathbf{\Lambda}\mathbf{F}_i + \mathbf{e}_i$, where $\mathbf{e}_i \sim N(0, 0.1^2\mathbf{I})$. To construct $\mathbf{\Lambda}$, we first generated n samples of the m-dimensional random vector $\mathbf{m}_i$ from $N(\mathbf{0}, (\sigma_{ij})_{m \times m})$

with $\sigma_{ij} = 0.5^{\|i-j\|_1}$. Denote $\mathbf{M} = (\mathbf{m}_1, \cdots, \mathbf{m}_n)^{\mathbf{T}}$. We performed the eigenvalue decomposition on the matrix $\mathbf{M}\mathbf{M}^{\mathbf{T}} = \mathbf{M}_{q_1}\mathbf{\Lambda}_{q_1}\mathbf{M}_{q_1}^{\mathbf{T}}$, where $\mathbf{\Lambda}_{q_1}$ was a diagonal matrix consisting of the q_1 largest eigenvalues and $\mathbf{M}_{q_1} \in \mathbb{R}^{n\times q_1}$ were their corresponding eigenvectors. In addition, $\mathbf{\Lambda}_n = \mathbf{M}^{\mathbf{T}}\mathbf{M}_{q_1}$, singular value decomposition was applied to the matrix $\mathbf{\Lambda}_n = \mathbf{S}_n\mathbf{V}_n\mathbf{D}_n$ and $\mathbf{\Lambda} = \mathbf{S}_n\mathbf{V}_n^{1/2}$.

Given $\boldsymbol{\zeta}_i$, we generated $\mathbf{X}_i(t)$ by $\mathbf{X}_i(t) = \mathbf{B}\mathbf{h}_i(t) + \mathbf{u}_i(t)$, where $\mathbf{u}_i(t) \sim N(0, 0.1^2\mathbf{I})$. The latent process $\mathbf{h}_i(t) = \{h_{ij}(t)\ (j = 1, \cdots, q_2)\}^{\mathbf{T}}$ is generated by $h_{ij}(t) = \sum_{k=1}^{2}\xi_{ijk}\phi_{jk}(t)$ with $\phi_{j1}(t) = \sin\{(2j-1)\pi t\}/\sqrt{5}$ and $\phi_{j2}(t) = \cos\{(2j-1)\pi t\}/\sqrt{5}$. To construct $\mathbf{B}$, we generated $\mathbf{B}_n$ following the process for $\mathbf{\Lambda}_n$ except that (m, q_1) was replaced by (p, q_2). Then, we performed QR decomposition on the matrix $\mathbf{B}_n = \mathbf{Q}_n\mathbf{R}_n$ and took $\mathbf{B} = p^{1/2}\mathbf{Q}_n$. For each trajectory $\mathbf{X}_i(t)$, 50 observation time points were sampled from $U(0, 1)$.

In all the examples, we took $q_1 = 10, q_2 = 5, K = 2, d = 4$ and $r = 2$. For each setting, we considered $n = 100$ or 500, $p = m = 100$ or 500, and conducted a total of 200 replications.

Example 1. To investigate the advantage of using the likelihood-based approach, we generated Y_i from the three settings, where the first two settings followed the model assumptions of the FFRM method with a normal distribution and non-normal distribution, respectively, and the last one was heteroscedastic; that was,
Setting I : $Y_i = \|\mathbf{\Omega}_1^{\mathbf{T}}\mathbf{f}_i + 1\|_1 + (\mathbf{\Omega}_2^{\mathbf{T}}\mathbf{f}_i)^2 + \sin(\mathbf{\Omega}_3^{\mathbf{T}}\mathbf{f}_i) + \cos(\mathbf{\Omega}_4^{\mathbf{T}}\mathbf{f}_i) + 0.2\epsilon_i$,
Setting II : $Y_i = \|\mathbf{\Omega}_1^{\mathbf{T}}\mathbf{f}_i + 1\|_1 + (\mathbf{\Omega}_2^{\mathbf{T}}\mathbf{f}_i)^2 + \sin(\mathbf{\Omega}_3^{\mathbf{T}}\mathbf{f}_i) + \cos(\mathbf{\Omega}_4^{\mathbf{T}}\mathbf{f}_i) + \epsilon_i^*$,
Setting III : $Y_i = \|\mathbf{\Omega}_1^{\mathbf{T}}\mathbf{f}_i + 1\|_1 + (\mathbf{\Omega}_2^{\mathbf{T}}\mathbf{f}_i)^2 + \sin(\mathbf{\Omega}_3^{\mathbf{T}}\mathbf{f}_i) + \cos(\mathbf{\Omega}_4^{\mathbf{T}}\mathbf{f}_i) + 0.1\|\mathbf{\Omega}_1^{\mathbf{T}}\mathbf{f}_i\|_1\epsilon_i$,
where $\epsilon_i \sim N(0, 1)$ is also used in the other examples, and $\epsilon_i^* \sim 2/3 \times N(-4, 2/5) + 2/3 \times N(2, 1/5)$; $\mathbf{\Omega}_k = \mathbf{V}^{\mathbf{T}}\mathbf{U}_k$ with $\mathbf{U}_1 = (3, -2)^{\mathbf{T}}/\sqrt{13}, \mathbf{U}_2 = (3, 2)^{\mathbf{T}}/\sqrt{13}, \mathbf{U}_3 = (2, 1)^{\mathbf{T}}/\sqrt{5}$, $\mathbf{U}_4 = (2, -1)^{\mathbf{T}}/\sqrt{5}$ and $\mathbf{V} = (\mathbf{V}_{[1]}, \cdots, \mathbf{V}_{[q]})$ with $q = 20$; $\mathbf{V}_{[k]} = (1, -1)^{\mathbf{T}}/2\sqrt{2}$ if $k \in \{1, 2, 11, 12\}$; and $\mathbf{V}_{[k]} = (1, 1)^{\mathbf{T}}/2\sqrt{2}$ if $k \in \{3, 4, 13, 14\}$ and $\mathbf{0}$ otherwise.

Example 2 was generated from the following model: $Y_i = 4 + \sum_{k=1}^{10}\cos(f_{ik} + 3) + \sum_{k=11}^{20}\sin(f_{ik} + 3) + 0.5\epsilon_i$, where f_{ik} was the k-th component of $\mathbf{f}_i$. Obviously, Example 2 did not satisfy the assumptions of PFLM and PLFAM. Without the dimension reduction step for features $\mathbf{\Omega}\mathbf{f}_i$, Example 2 also did not follow the assumption of the proposed FFRM method.

Example 3 followed the assumption of PFLM; that was, $Y_i = \sum_{j=1}^{p}\int_0^1 \beta_j(t)X_{ij}(t)dt + \boldsymbol{\eta}^{\mathbf{T}}\mathbf{Z}_i + 0.3\epsilon_i$, where $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_m)^{\mathbf{T}}$ with $\eta_k = 0.1/\sqrt{10}$ if $k \leqslant 20$ and 0 otherwise; $\beta_j(\cdot) = \mathbf{b}_j^{\mathbf{T}}\mathbf{\Phi}(\cdot)$ with $\mathbf{b}_j = 0.1 \times 1_{10} \times \mathbf{1}(j \leqslant 20)$ and $\mathbf{\Phi}(\cdot)$ was formed by the first 10 eigenfunctions of $X_{ij}(t)$. Here, 1_{10} was a 10-dimensional vector with component 1.

Example 4 was generated using the setting of PLFAM, $Y_i = \sum_{k=1}^{10} f_k(\zeta_{ik}^*) + \boldsymbol{\theta}^{\mathbf{T}}\mathbf{Z}_i + 0.5\epsilon_i$, where $f_1(x) = 3x - 1.5, f_3(x) = \sin\{2\pi(x - 0.5)\}, f_5(x) = 4(x - 0.5)^3 - 8/9, f_7(x) = 2\cos(x + 0.5)$ and $f_k(x) = 0$ if $k \notin \{1, 3, 5, 7\}$; $\zeta_{ik}^* = \Phi(\lambda_k^{-1/2}\zeta_{ik})$ with $(\zeta_{i1}, \cdots, \zeta_{i,10})$ being the first 10 multivariate functional principal component analysis (mFPCA) scores of $\mathbf{X}_i(t)$ and $\lambda_k = \text{var}(\zeta_{ik})$; $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_m)^{\mathbf{T}}$ with $\theta_k = 0.1/\sqrt{10}$ if $k \leqslant 5$, $\theta_k = -0.1/\sqrt{10}$ if $5 < k \leqslant 10$ and 0 otherwise.

6.2. *Comparison with the LSE criterion.* To compare the performance of FFRM and LSE, we first extracted features from $\mathbf{Z}_i$ using model (1) and from $\mathbf{X}_i(t)$ using model (4). Subsequently, we computed the LSE of $\widehat{\boldsymbol{\gamma}}$ based on (7), that is,

$$\widehat{\boldsymbol{\gamma}} = \underset{\mathbf{U},\mathbf{V},\mathbf{\Psi}\in\mathscr{F}}{\arg\max}\left\{\ell_n(\boldsymbol{\gamma}; \widehat{\mathbf{f}}, \rho) - \lambda\sum_{k=1}^{q} w_k\|\mathbf{V}_{[k]}\|_2\right\},$$

where $\ell_n(\boldsymbol{\gamma}; \mathbf{f}, \rho) = n^{-1}\sum_{i=1}^{n}\ell(\boldsymbol{\gamma}; Y_i, \mathbf{f}_i, \rho) = n^{-1}\sum_{i=1}^{n}\rho\{Y_i - \sum_{j=1}^{d}\psi_j(\mathbf{U}_j^{\mathbf{T}}\mathbf{V}\mathbf{f}_i)\}$ with the least squares loss function $\rho(x) = -x^2$.

In Example 1, we generated Y_i according to the proposed setting in (6), which looked independent of $\phi_{jk}(t)$. However, we estimated all unknown parameters and functions based on the observations $\{Y_i, \mathbf{X}_i(t), \mathbf{Z}_i\}$ $(i = 1, \cdots, n)$, which were related to $\phi_{jk}(t)$. Therefore, the problem could still be considered as a functional regression problem from the perspective of functional predictors. Figure 6 shows the boxplots of $\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}_0\|_F$ using the proposed FFRM method and LSE with the three settings in Example 1. Clearly, the performance of both methods improved as n or (p, m) increases, which was consistent with Theorem S7.1 in Section S7. In addition, we have some interesting findings in Figure 6. First, since the random error in Setting I followed a normal distribution, it was expected that LSE performs better than FFRM. However, we observed this phenomenon only when $n = 100, p = m = 100$ in Figure 6(a). When n or (p, m) increased, we found that FFRM performed slightly better than LSE in Setting I; see Figure 6(b,c). In fact, superefficiency has been observed in the literature (Zhou et al., 2019; Lin et al., 2021) and may be attributed to the use of structural information, that is, the estimated density function $\widehat{f}(\cdot)$. Since the random error in Setting II was not normally distributed, it was not surprising that FFRM is much better than LSE; see Figure 6(d,e,f). Moreover, although the random error in Setting III was heteroscedastic, which does not satisfy the assumption required by FFRM, FFRM still performed well and was much better than LSE; see Figure 6(g,h,i). Supplementary Table S1 shows the Bias, SD and RMSE for $\widehat{\mathbf{\Psi}}(\cdot)$ of FFRM and LSE. Similar conclusions can be reached as shown in Figure 6.

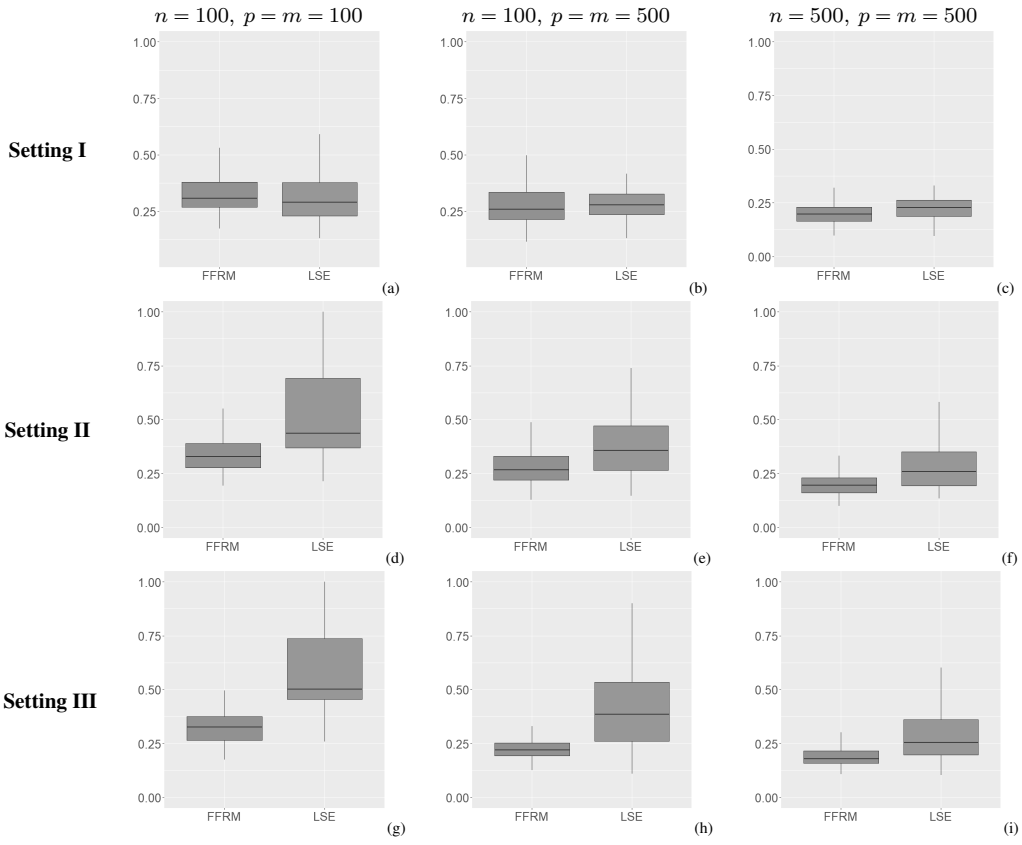


Fig 6: The boxplot of $\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}_0\|_F$ for Example 1 using the proposed FFRM and the estimator based on LSE criterion.

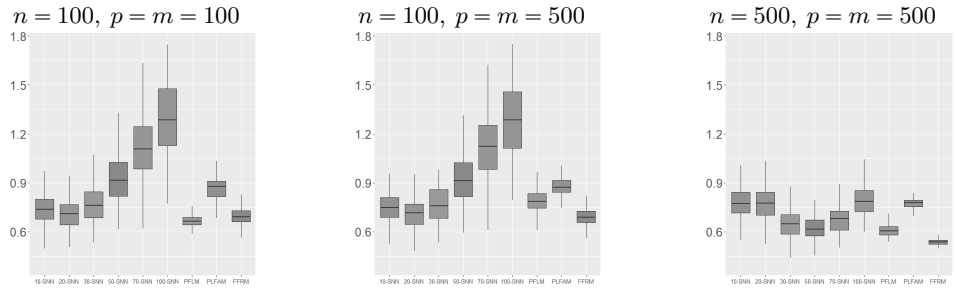


Fig 7: The boxplot of PE for Example 2 using 10-SNN, 20-SNN, 30-SNN, 50-SNN, 70-SNN, 100-SNN, PFLM, PLFAM and FFRM.

6.3. *Comparison with the existing models.* Figure 7 shows boxplots of the PE using the proposed FFRM, PFLM, PLFAM and d -SNN with $d = 10, 20, 30, 50, 70, 100$ for Example 2, where the assumptions of the methods considered were not satisfied. In the FFRM, $d = 2$ and $r = 2$ were selected based on the method described in the Suppl. S1. Figure 7 shows that the proposed FFRM method exhibits the best or almost the best performance, particularly when (p, m) is large. FFRM with only two component functions achieved the same prediction accuracy as 20-SNN for a sample size of $n = 100$ and outperformed every d-SNN when $n = 500$, indicating the importance of estimating the activation function rather than specifying it. Furthermore, the simplicity of FFRM improved its interpretability, while simulation studies demonstrated that the prediction accuracy of d-SNN was sensitive to sample size and did not improve with increasing d, possibly due to the presence of latent sparse features.

Figure 8 shows a boxplot of the PEs using the proposed FFRM, 4-SNN, 10-SNN, PFLM and PLFAM for Example 1 with Setting II, and Examples 3 and 4 with the assumptions satisfied in FFRM, PFLM and PLFAM, respectively. The results of Example 1 in Figure 8 shows that our proposed FFRM outperformed PFLM, PLFAM and SNN in terms of prediction accuracy, as the data aligned with the FFRM assumptions. It was not surprising that PFLM and PLFAM performed best in Examples 3 and 4 respectively. However, FFRM showed robustness to the violation of model assumptions, with similar performance to the methods that favored Examples 3 and 4, respectively. For all examples, FFRM performed better compared to SNN. In summary, under the scenarios examined, it is evident that the proposed FFRM achieves the best performance when the required assumptions are satisfied. In addition, it still delivers reasonable performance even when these assumptions are violated, suggesting promising real-world applications.

7. Discussion. This study examines the relationship between LDL levels and both genetic markers (SNPs) and longitudinal anthropometric measures. To address the methodological challenges of modeling mixed high-dimensional scalar and functional covariates, we developed a novel FFRM approach. Our approach demonstrates superior prediction accuracy compared to existing methods and successfully identifies 424 SNPs significantly associated with LDL levels, several of which confirm previous research findings. We further applied FFRM to investigate the determinants of adult BMI and revealed the significant parental and individual characteristics influencing adult BMI, demonstrating the broader applicability of our methodology beyond LDL prediction.

In summary, the proposed FFRM offers several advantages: (1) Computational efficiency: By deriving closed-form expressions for the extraction of sufficient information from the scalar and functional covariates, we avoid complex and potentially nonconvergent iterations.

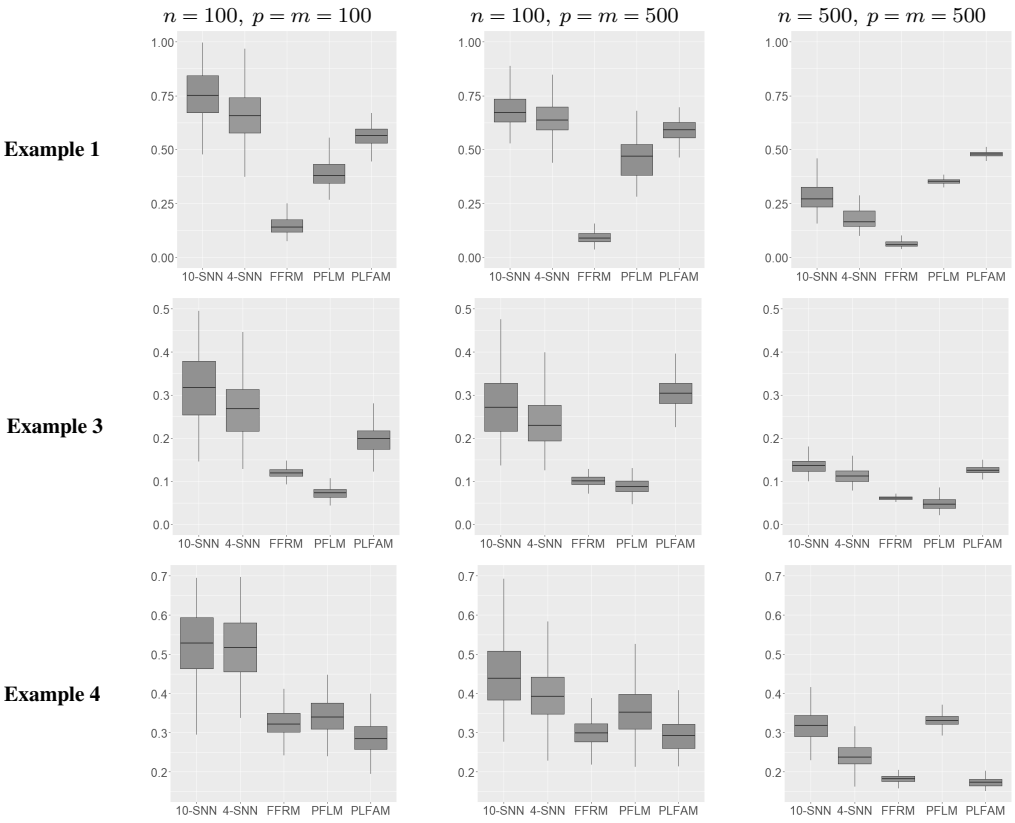


Fig 8: The boxplot of PE using the proposed FFRM, 4-SNN, 10-SNN, PFLM and PLFAM for Examples 1, 3, 4, which satisfy the assumptions of FFRM, PFLM and PLFAM, respectively.

This results in computational efficiency. (2) High prediction accuracy: FFRM allows the number of components to diverge to infinity, enabling a universal approximation for any relationship between the response and the features. Both real data analysis and simulation studies confirm that our method achieves higher prediction accuracy compared to existing methods. It even outperforms shallow neural networks due to the estimated activation functions. (3) Interpretability: Model (6) with a low-rank constraint on $\boldsymbol{\Omega} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$ and a sparse restriction on the columns of $\boldsymbol{\Omega}$ provides simple and interpretable expressions for the relationship between the response and covariates. (4) Efficiency and stability: We provide a framework based on likelihood of sieves for estimating parameters and functions that enhances the efficiency, flexibility, and stability of estimators. This is supported by numerical comparisons with estimators based on the least square error. (5) Theoretical assurance: We establish theoretical properties of the proposed estimator, including selection consistency, estimation consistency, convergence rate, and asymptotic normality.

Several extensions can be considered for future research. We currently focus on continuous responses, but our method can be easily extended to handle discrete responses using a generalized framework. Additionally, the model can be extended to handle multiple responses, taking into account the correlation among them. Furthermore, exploring the FFRM approach in the context of high-dimensional discrete covariates is an avenue for future investigation.

Acknowledgments. We thank the anonymous referees, Associate Editor and the Editor for the constructive comments that improved the quality of this paper. The data supporting the findings of this article are available from the Avon Longitudinal Study of Parents and

Children. Restrictions apply to the availability of these data, which were used under license in this paper. Data are available at https://www.bristol.ac.uk/alspac/ with permission of the Avon Longitudinal Study of Parents and Children.

Funding. The research was partially supported by National Key R&D Program of China (No.2022YFA1003702), National Natural Science Foundation of China (Nos. 12426309, 11931014, 12171374, 12371275, 12371283), New Cornerstone Science Foundation.

SUPPLEMENTARY MATERIAL

S1: Selection of initial values and tuning parameters

S2: Conditions for the asymptotic property of $\widehat{\boldsymbol{\zeta}}_i$

S3: Conditions for the asymptotic property of $\widehat{\mathbf{F}}_i$

S4: Conditions for the asymptotic property of $\widehat{\gamma}$

S5: Notations
In Supplementary Material S5, we define the directional derivatives and the asymptotic variances of the estimator.

S6: Identifiablity of the FFRM
In Supplementary Material S6, we establish the identifiability of model (6) accompanying with (1) and (4).

S7: Theoretical Properties
In Supplementary Material S7, we establish the theoretical properties, including the estimation and selection consistency, and the asymptotic normality.

S8: Lemmas
In Supplementary Material S8, we establish five lemmas, e.g., the covergence rate of the extracted features and the kernel density estimates, for the proofs of the main theorems.

S9: Proofs of the main results
In Supplementary Material S9, we provide detailed proofs of the theoretical results in Supplementary Material S7.

S10: Other results in numerical studies
In Supplementary Material S10, we show the Bias, SD and RMSE for $\widehat{\boldsymbol{\Psi}}(\cdot)$ of the proposed FFRM method and LSE.

S11: The transformation from the regression relationships between LDL and scores to functioanl covariates for analyzing the effects of functional covariates on LDL
In Supplementary Material S11, we show the transformation of $\boldsymbol{\alpha}\boldsymbol{\zeta}_i$ to $\int \boldsymbol{\eta}^{\mathbf{T}}(t)\mathbf{X}_i(t)dt$ in Section 4.

S12: Other results of the analysis of the BMI outcomes with the ALSPAC data
In Supplementary Material S12, we present the feature selection process in the analysis of BMI, along with introductions to the functional and scalar covariates.

REFERENCES

Ash, R. and Gardner, M. (1975). *Topics in stochastic processes*. Academic Press.

Bai, J. and Ng, S. (2013). Principal components estimation and identification of static factors. *J. Econometrics*, 176(1):18–29.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, 57(1):289–300.

Bodnar, L. M., Pugh, S. J., Abrams, B., Himes, K. P., and Hutcheon, J. A. (2014). Gestational weight gain in twin pregnancies and maternal and child health: a systematic review. *J. Perinatol.*, 34(4):252–63.

Cai, T. T. and Hall, P. (2006). Prediction in functional linear regression. *Ann. Stat.*, 34(5):2159–2179.

Callaway, L. K., Lust, K., and Mcintyre, H. D. (2005). Pregnancy outcomes in women of very advanced maternal age. *Obstet. Gynecol. Surv.*, 60(9):562–563.

Chen, D., Hall, P., and Müller, H.-G. (2011). Single and multiple index functional regression models with nonparametric link. *Ann. Stat.*, 39(3):1720–1747.

Demarco, V. G., Aroor, A. R., and Sowers, J. R. (2014). The pathophysiology of hypertension in patients with obesity. *Nat. Rev. Endocrinol.*, 10(6):364–376.

Fan, J., Ke, Y., and Wang, K. (2020). Factor-adjusted regularized model selection. *J. Econometrics*, 216(1):71–85.

Fan, J. and Li, R. (2001). Variable selection via nonconvave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.*, 96(456):1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. B*, 70(5):849–941.

Fan, Y., James, G. M., and Radchenko, P. (2015). Functional additive regression. *Ann. Stat.*, 43(5):2296–2325.

Ference, B. A. Yoo, W., Alesh, I., Mahajan, N., Mirowska, K. K., Mewada, A., Kahn, J., Afonso, L., Williams, K. A., and Flack, J. M. (2012). Effect of long-term exposure to lower low-density lipoprotein cholesterol beginning early in life on the risk of coronary heart disease: A Mendelian randomization analysis. *J. Am. Coll. Cardiol.*, 60(25):2631–2639.

Fujita, Y., Kouda, K., Nakamura, H., and Iki, M. (2016). Inverse association between height increase and LDL cholesterol during puberty: A 3-year follow-up study of the Fukuroi city. *Am. J. Hum. Biol.*, 3(28):330–334.

George, D. S., Colin, S., Sam, L., and Andy, N. (2007). Is there an intrauterine influence on obesity? Evidence from parent-child associations in the Avon Longitudinal Study of Parents and Children (ALSPAC). *Arch. Dis. Child.*, 92(10):871–880.

Hall, P. and Hosseini-Nasab, M. (2006). On properties of functional principal components analysis. *J. R. Stat. Soc. B*, 68(1):109–126.

Hall, P., Marron, J. S., and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *J. R. Stat. Soc. B*, 67(3):427–444.

James, G. M. and Silverman, B. W. (2005). Functional adaptive model estimation. *J. Am. Stat. Assoc.*, 100(470):565–576.

Jansen, H., Samani, N. J., and Schunkert, H. (2014). Mendelian randomization studies in coronary artery disease. *Eur. Heart J.*, (29):1917–1924.

Jiang, F., Ma, Y., and Wei, Y. (2019). Sufficient direction factor model and its application to gene expression quantitative trait loci discovery. *Biometrika*, 106(2):417–432.

Kong, D., Xue, K., Yao, F., and Zhang, H. (2016). Partially functional linear regression in high dimensions. *Biometrika*, 103(1):147–159.

Li, Y. and Hsing, T. (2010). Deciding the dimension of effective dimension reduction space for functional and high-dimensional data. *Ann. Stat.*, 38(5):3028–3062.

Lin, H., Liu, W., and Lan, W. (2021). Regression analysis with individual-specific patterns of missing covariates. *J. Bus. Econ. Stat.*, 39(1):179–188.

Liu, S., Zhang, H., and Zhang, J. (2021). Model averaging estimation for partially linear functional score models. *arXiv:2105.00953*.

Liu, X., Li, J., Xie, J., Ma, G., Xu, K., and Yang, J. (2023). The association between body fluid rate with plasma lipid profile, independent of adiposity in young adults. *Proc. Eur. Nutr. Conf. FENS 2023*, 91(1):355.

Lu, Y., Du, J., and Sun, Z. (2014). Functional partially linear quantile regression model. *Metrika*, 77(2):317–332.

Ma, Y., Li, Y., Lin, H., and Li, Y. (2017). Concordance measure-based feature screening and variable selection. *Stat. Sinica*, 27(4):1967–1985.

Maffeis, C., Pietrobelli, A., Grezzani, A., Provera, S., and Luciano, T. (2012). Waist circumference and cardiovascular risk factors in prepubertal children. *Obesity*, 9(3):179–187.

Müller, H.-G., Wu, Y., and Yao, F. (2013). Continuously additive models for nonlinear functional regression. *Biometrika*, 100(3):607–622.

Noah, S., Jerome, F., Trevor, H., and Robert, T. (2013). A Sparse-Group Lasso. *J. Comput. Graph. Stat.*, 22(2):231–245.

Oda, E. (2013). LDL cholesterol was more strongly associated with percent body fat than body mass index and waist circumference in a health screening population. *Obes. Res. Clin. Pract.*, 12(2):195–203.

Packard, C., Caslake, M., and Shepherd, J. (2000). The role of small, dense low density lipoprotein (LDL): a new look. *Int. J. Cardiol.*, 74(1):S17–S22.

Pinkus, M. A. (1999). Lower bounds for approximation by MLP neural networks. *Neurocomputing*, 25(1-3):81–91.

Rasmussen, K. M. (2001). Is there a causal relationship between iron deficiency or iron-deficiency anemia and weight at birth, length of gestation and perinatal mortality? *J. Nutr.*, 131(2):590–601.

Reiss, P. T. and Ogden, R. T. (2010). Functional generalized linear models with images as predictors. *Biometrics*, 66(1):61–69.

Sandhu, M. S., Waterworth, D. M., Debenham, S. L., Wheeler, E., and Mooser, V. (2008). LDL-cholesterol concentrations: a genome-wide association study. *Lancet*, 371(9611):483–491.

Schooling, C. M., Jiang, C., Lam, T. H., Thomas, G. N., Heys, M., Bmbs, Lao, X., Zhang, W., Adab, P., and Cheng, K. K. (2007). Height, its components, and cardiovascular risk among older chinese: A cross-sectional analysis of the guangzhou biobank cohort study. *Am. J. Public. Health.*, 97(10):1834–1841.

Schwartz, C. J., Valente, A. J., Sorague, E. A., Kelley, J. L., and Nere, R. M. (1991). The pathogenesis of atherosclerosis: An overview. *Clin. Cardiol*, 14(1):1–16.

Shamai, L., Lurix, E., Shen, M., Novaro, G. M., Rosenthal, R., Hernandez, A. V., and Asher, C. R. (2009). Association of body mass index and lipid profiles: Evaluation of a broad spectrum of body mass index patients including the morbidly obese. *Obes. Surg.*, 5(3):42–47.

Shirasawa, T., Ochiai, H., Ohtsu, T., Nishimura, R., Morimoto, A., Hoshino, H., Tajima, N., and Kokaze, A. (2013). LDL-cholesterol and body mass index among Japanese schoolchildren: A population-based cross-sectional study. *Lipids Health Dis.*, 12(77):1–6.

Simpson, J., Smith, A. D. A. C., Fraser, A., Sattar, N., Lindsay, R. S., Ring, S. M., Tilling, K., Smith, G. D., Lawlor, D. A., and Nelson, S. M. (2017). Programming of adiposity in childhood and adolescence: associations with birth weight and cord blood adipokines. *J. Clin. Endocrinol. Metab.*, 102(2):499–506.

Stone, C. J. (1975). Adaptive maximum likelihood estimators of a location parameter. *Ann. Stat.*, 3(2):267–284.

Tang, A. M., Dong, K., Deitchler, M., Chung, M., Maalouf-Manasseh, Z., Tumilowicz, A., and Wanke, C. (2013). *Use of cutoffs for Mid-Upper Arm Circumference (MUAC) as an indicator or predictor of nutritional and health-related outcomes in adolescents and adults: a systematic review*. Washington, DC: FHI 360/FANTA.

Institute for Health Metrics and Evaluation (IHME) (2024). *Global Burden of Disease 2021: Findings from the GBD 2021 Study*. Seattle, WA: IHME.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. B*, 58(1):267–288.

Wang, H. (2012). Factor profiled sure independence screening. *Biometrika*, 99(1):15–28.

Wen, S., Li, Y., Kong, D., and Lin, H. (2025+a). Prediction of cognitive function via brain region volumes with applications to Alzheimer's disease based on Space-Factor-Guided Functional Principal Component Analysis. *J. Am. Stat. Assoc.* online.

Wen, S., Liu, L., Liu, J., Li, Y., and Lin, H. (2025+b). Supplement to "Factor-assisted learning of ultrahigh-dimensional covariates with distributed functional and scalar mixtures with applications to the Avon Longitudinal Study of Parents and Children".

Willer, C. J., Schmidt, E. M., and Sengupta, S. (2013). Discovery and refinement of loci associated with lipid levels. *Nat. Genet.*, 45(11):1274–1283.

Wong, R. K. W., Li, Y., and Zhu, Z. (2019). Partially linear functional additive models for multivariate functional data. *J. Am. Stat. Assoc.*, 114(525):406–418.

Xue, K. and Yao, F. (2021). Hypothesis testing in large-scale functional linear regression. *Stat. Sinica*, 31(2):1101–1123.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005a). Functional data analysis for sparse longitudinal data. *J. Am. Stat. Assoc.*, 100(470):577–590.

Yao, F., Müller, H. G., and Wang, J. L. (2005b). Functional linear regression analysis for longitudinal data. *Ann. Stat.*, 33(6):2873–2903.

Zhang, X. and Wang, J.-L. (2016). From sparse to dense functional data and beyond. *Ann. Statist.*, 44(5):2281–2321.

Zhong, Q., Lin, H., and Li, Y. (2021). Cluster non-gaussian functional data. *Biometrics*, 77(3):852–865.

Zhou, L., Lin, H., Chen, K., and Liang, H. (2019). Efficient estimation and computation of parameters and nonparametric functions in generalized semi/non-parametric regression models. *J. Econometrics*, 213(2):593–607.

Zhou, L., Lin, H., and Liang, H. (2018). Efficient estimation of the nonparametric mean and covariance functions for longitudinal and sparse functional data. *J. Am. Stat. Assoc.*, 113(524):1550–1564.

Zhu, H., Yao, F., and Zhang, H. H. (2014). Structured functional additive regression in reproducing kernel Hilbert spaces. *J. R. Stat. Soc. B*, 76(3):581–603.

Zhu, Y., Lin, Q., Zhang, Y., Deng, H., Hu, X., Yang, X., and Yao, B. (2020). Mid-upper arm circumference as a simple tool for identifying central obesity and insulin resistance in type 2 diabetes. *PLoS One*, 15(5):e0231308.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *J. Am. Stat. Assoc.*, 101(476):1418–1429.