The Annals of Applied Statistics
2020, Vol. 14, No. 1, 276–298
https://doi.org/10.1214/19-AOAS1310

FEATURE SELECTION FOR GENERALIZED VARYING COEFFICIENT MIXED-EFFECT MODELS WITH APPLICATION TO OBESITY GWAS

BY WANGHUAN CHU[1], RUNZE LI[2], JINGYUAN LIU[3] AND MATTHEW REIMHERR[4]

[1]*Google Inc., dqchuwh@gmail.com*
[2]*Department of Statistics and the Methodology Center, Pennsylvania State University, rzli@psu.edu*
[3]*MOE Key Laboratory of Econometrics, Department of Statistics, School of Economics, Wang Yanan Institute for Studies in Economics, and Fujian Key Lab of Statistics, Xiamen University, jingyuan@xmu.edu.cn*
[4]*Department of Statistics, Pennsylvania State University, mreimherr@psu.edu*

Motivated by an empirical analysis of data from a genome-wide association study on obesity, measured by the body mass index (BMI), we propose a two-step gene-detection procedure for generalized varying coefficient mixed-effects models with ultrahigh dimensional covariates. The proposed procedure selects significant single nucleotide polymorphisms (SNPs) impacting the mean BMI trend, some of which have already been biologically proven to be "fat genes." The method also discovers SNPs that significantly influence the age-dependent variability of BMI. The proposed procedure takes into account individual variations of genetic effects and can also be directly applied to longitudinal data with continuous, binary or count responses. We employ Monte Carlo simulation studies to assess the performance of the proposed method and further carry out causal inference for the selected SNPs.

1. Introduction. In genome-wide association studies (GWAS), hundreds of thousands or millions of single nucleotide polymorphisms (SNPs) are explored for their associations with phenotypes or traits of interest, such as body mass index (BMI), blood pressure, asthma and many other complex traits. Traditional GWAS deal with cross-sectional phenotypic data, focusing on a single point in time and primarily explore the mean/fixed effects of SNPs on the phenotype of interest. The present work moves beyond this paradigm in two important aspects. First, we consider longitudinally measured phenotypes, that is, traits measured repeatedly from each individual over time which need not be normally distributed (e.g., binary, counts, etc). Second, we provide a framework for selecting both mean/fixed effects and variance/random effects. In other words, we select SNPs that either affect the mean of the phenotype or the variance of the phenotype. These extensions allow one to develop a much deeper understanding of the temporal-genetic architecture of a complex trait.

The Framingham Heart Study[1] (FHS) is a long-term longitudinal study whose aim is to better understand the risk factors of heart disease. The study began in 1948 and is on its third cohort of subjects. Using this study, we aim to identify genetic factors influencing BMI, either through the mean or variance. Furthermore, we aim to understand how these effects change as subjects age. We focus on the second cohort (children of the first cohort), who were genotyped using the Illumina Omni5 platform, resulting in approximately five million SNPs across approximately 2000 subjects. Repeated measurements include BMI, smoking status and alcohol intake.

To model age-dependent associations, we consider generalized varying coefficient models (Hastie and Tibshirani (1993)). However, estimation of generalized varying coefficient

Received December 2018; revised June 2019.
Key words and phrases. Genome-wide association study, mixed effects, ultrahigh dimensional longitudinal data, varying coefficient models.

[1]https://en.wikipedia.org/wiki/Framingham_Heart_Study

models with ultrahigh dimensional covariates (such as millions of SNPs) becomes infeasible both methodologically and computationally. To reduce the dimensionality of such problems, marginal feature screening procedures serve as computationally efficient solutions to pick potentially important subsets of SNPs. Fan and Lv (2008) developed a marginal sure independence screening procedure (SIS) for ultrahigh dimensional linear models based on Pearson correlations. Several subsequent feature screening procedures have been developed for various models; see a brief review by Liu, Zhong and Li (2015) and references therein. Liu, Li and Wu (2014) extended an SIS for varying coefficient models using conditional Pearson correlations. Fan, Ma and Dai (2014) and Xia, Yang and Li (2016) proposed sure independence screening procedures for varying coefficient models and generalized varying coefficient models, respectively, by extending the B-spline techniques in Fan, Feng and Song (2011) for additive models. Song, Yi and Zou (2014) extended the proposal of Fan, Ma and Dai (2014) for longitudinal data with a continuous response. Chu, Li and Reimherr (2016) further improved the proposal of Song, Yi and Zou (2014) by exploiting within-subject correlation. Our empirical analysis of FHS shows that certain SNPs affect not only the mean of BMI but also its variability. Furthermore, both mean and variance effects vary as subjects age. None of the aforementioned works take into account the longitudinal genetic effects on the variability of traits, despite this being an important goal for geneticists (Aschard et al. (2013), Furlotte and Eskin (2015), Geiler-Samerotte et al. (2013), Paré et al. (2010), Soave et al. (2015)).

In this paper we incorporate a mixed effects structure alongside a generalized varying coefficient model so as to examine longitudinal genetic effects on the mean and variability of traits. More specifically, the fixed effects in the models are responsible for the mean trajectory of the phenotypes, while the random effects deal with the temporally-changing variance of the phenotypes. We develop a new two-step screening procedure in which we first filter out unimportant fixed-effect variables by ranking the deviance of the marginal model, and then the random effects are screened based on the estimated variance of the corresponding random effect term. In both steps, all terms are estimated using a B-splines expansion. Our proposed procedure can be directly applied to longitudinal genetic data with continuous, binary or count responses.

We apply the proposed procedure to FHS data. Several covariates—gender, smoking status and alcohol intake—are incorporated in the model. After the two-step screening stage, 140 SNPs are retained, 70 with fixed effects and 70 with random effects. In the postscreening variable selection and inference stage, we conduct a longitudinal group LASSO (Barber, Reimherr and Schill (2017)) for further selection of fixed effects and forward regression for random effects. The longitudinal effects of the final 52 fixed-effect SNPs, as well as the covariates, are depicted via the corresponding coefficient curves over age.

Detecting the random effects, which is our biggest contribution in this paper, enables us to discover the age trends of BMI variability due to the selected SNPs. The five chosen random effects are illustrated via the variance function. These curves all display a significant age-varying trend. Each covariate or SNP has its particular effect pattern, but some commonalities are observed. For instance, age 40 to 60 is a special period for the covariates and many SNP effects; some effects exhibit dramatic structural changes during this period while others are more stable. A subset of SNPs selected by our method coincide with SNPs previously identified in the literature that have been associated with obesity or certain diseases related to obesity. Thus, the new SNPs discovered in our work are worthy of further scientific investigation. Moreover, according to our causal inference analysis, the SNPs selected by our method have a much larger causal effect size than those previously found in literature.

For the rest of the paper, we introduce the necessary background material and provide some descriptive statistics of the FHS data in Section 2. In Section 3 the statistical framework is

studied, consisting of the model structure and the newly developed two-step mixed-effect screening procedure. The analysis of the FHS data is discussed in Section 4—the two-step screening, the postscreening variable selection are both conducted. Simulation studies are provided in Section 5 to empirically verify the proposed methodology, including comparisons with other related methods. A summary is provided in Section 6. A causal inference study is given in the Supplementary Material (Chu et al. (2020)).

2. Background.

2.1. *Obesity studies*. Obesity is a significant risk factor for various chronic diseases, such as cardiovascular diseases, diabetes, musculoskeletal disorders, etc. Based on the latest World Health Organization (WHO) global estimates,[2] the worldwide prevalence of obesity has more than doubled since 1980. In 2014, over 600 million adults were obese, which composed about 13% of the world adult population, and this number was projected to reach 1.12 billion by 2030 if current trends continue unabated (Kelly et al. (2008)). While obesity was once considered a public health problem only in developed countries, the prevalence rates of obesity in developing countries are rapidly catching up (Bell, Walley and Froguel (2005)). Such a global pandemic of obesity is becoming a major concern to public health worldwide.

Undoubtedly, environmental factors such as changes in dietary and lifestyle patterns during globalization (i.e., increase in energy intake and decrease in physical activity) have contributed to the global pandemic of obesity. However, environmental influences cannot explain the considerable between-individual variation in body weight within a population sharing the same environment (Wardle et al. (2008)). Studies have suggested a strong genetic contribution to the between-individual variation in body weight with an estimated heritability between 40–70% (Ramachandrappa and Farooqi (2011), Allison et al. (1996)). Traditionally, inheritable obesity has been categorized as monogenic (i.e., caused by a single defective gene), syndromic (i.e., inherited in either an autosomal or an X-linked pattern) or polygenic (i.e., caused by the interaction between multiple genes and the environment) (Rankinen et al. (2006)). Although many candidate genes involved in monogenic and syndromic forms of obesity (e.g., Leptin gene) have been successfully identified over the past two decades, genetic studies of the common and most complex polygenic obesity have been characterized by relatively slow progress, and the genetics underlying common obesity has remained elusive (Herrera, Keildson and Lindgren (2011)).

Obesity is defined as the total fat mass of an individual, which ideally is measured by direct fat-measuring methods such as DEXA scans. However, for practical and economic reasons one usually uses a surrogate measurement such as the body mass index (BMI $=$ weight/height^2) or waist circumference (WC) (Fall and Ingelsson (2014)). In this analysis, we examine BMI, and a person is considered obese if his/her corresponding BMI >30 kg/m^2. Another reason for using BMI is that it is highly correlated with body fat, and the WHO has standard categories of it for monitoring populations worldwide. The scientific goal for this paper is to detect significant SNPs that are associated with the developmental trajectory of BMI, as well as its functionally-changing variability.

2.2. *Data description*. The Framingham Heart Study (FHS) is a long-term, ongoing cardiovascular study that began in 1948 under the direction of the National Heart, Lung and Blood Institute (NHLBI) on residents of the town of Framingham, Massachusetts.[3] In this

[2] http://www.who.int/mediacentre/factsheets/fs311/en/

[3] https://en.wikipedia.org/wiki/Framingham_Heart_Study

TABLE 1
Number of subjects for each wave of visit

Number of visits	7	6	5	4	3	2	1	Total #obs.: 12,632
Number of subjects	1453	286	85	44	38	12	6	Total: 1924

study, 913,854 SNPs from 24 chromosomes are genotyped from the second (offspring) cohort study with 1924 subjects. Let "A" and "a" stand for the major and minor allele of a SNP and we code "AA," "Aa" and "aa" into 0, 1 and 2. We use SNPs with a minor allele frequency (MAF) greater than 5%, resulting in 718,867 SNPs.

Each subject participated in up to seven clinical visits where a variety of medical measurements were recorded, including height, weight, age, etc, and survey questions such as alcohol intake and smoking status. Table 1 tabulates the number of subjects that participated in *one*, *two*, ..., *seven* clinical visits. The elapsed time between the first and second waves is *eight* years, with the remaining visits conducted every *three–four* years.

The histograms of BMI at the original and log scale are shown in the top panel of Figure 1, and we adopt $\text{BMI}^* = \log(\text{BMI})$ so that the response variable has a bell-shaped histogram.

Subject age has an approximately normal distribution between five and 85 years old. Instead of using the original scale of age, we use its normalized rank (i.e., its rank dividing by the sample size) to satisfy the assumption of B-spline regression. The histograms of age before and after transformation are shown in the bottom panel of Figure 1.

Based on our preliminary analysis, gender, smoking and alcohol may have significant effects on BMI. Thus, three covariates are included in the model but are not subject to selection.

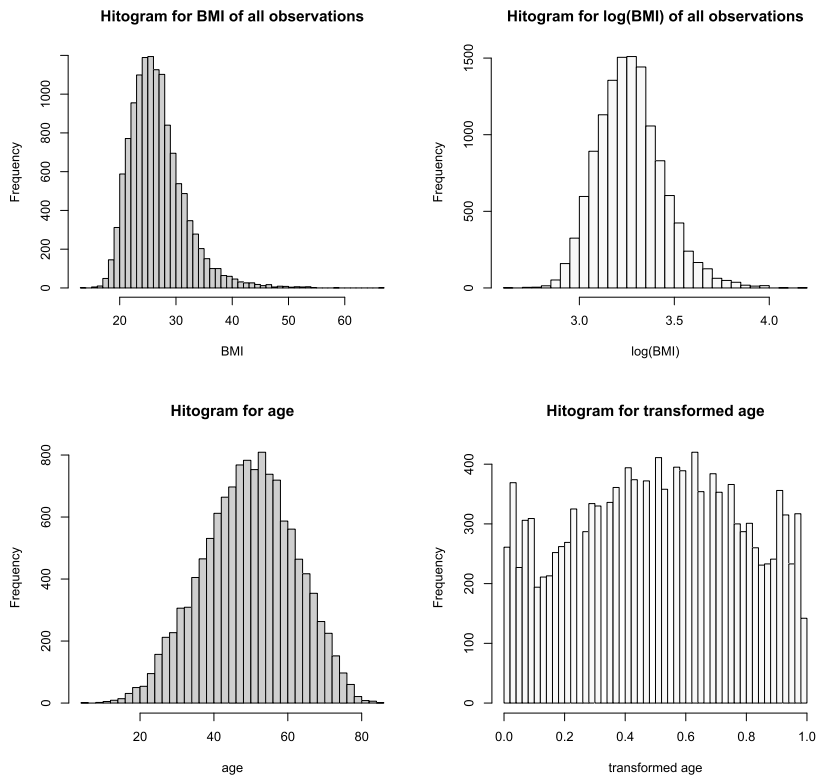
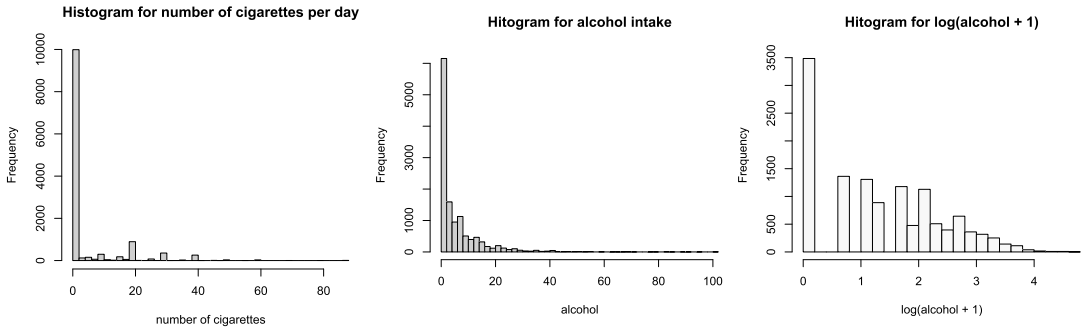


FIG. 1. *Top panel are histograms for BMI on the original and log scale, and the bottom panel are histograms for the age variable before and after transformation.*



(a) Smoking (b) Alcohol intake

FIG. 2. *Histograms for smoking and alcohol covariates.*

For gender, the sample consists of 873 men (reference level) and 1051 women. To evaluate the impact of smoking behavior, we use "the number of cigarettes smoked per day" in the survey questions, with the histogram shown in the left panel of Figure 2a, which suggest converting the smoking status into a binary covariate, smoking vs. nonsmoking. The amount of alcohol intake combines servings of beer, wine and spirits per week for each individual. We use $\text{Alcohol}^* = \log(\text{Alcohol} + 1)$ for further analysis to avoid an over-skewed distribution. The histograms of alcohol intake at the original and transformed scale are shown in the middle and right panels of Figure 2b, respectively.

3. Statistical framework. In this section we develop the relevant statistical model for the aforementioned scientific goal and propose the corresponding two-step screening procedure.

3.1. *Statistical model.* To address the issue of dynamic effects of SNPs on BMI across time (age), it is natural to consider the generalized time-varying coefficient model. To include potential effects on the variability of BMI, we include random effects into the model.

Denote the observed sample at age t as $\{y_i(t), \mathbf{z}_i(t), \mathbf{x}_i(t), \mathbf{u}_i(t), i = 1, \ldots, n\}$. Here, n denotes the sample size, $y_i(t)$ is the BMI for subject i at age t and $\mathbf{z}_i(t)$ is the low-dimensional predictor vector that is not subject to screening, specifically consisting of gender, smoking status and alcohol intake in this empirical analysis of FHS data. The fixed-effect covariate vector, $\mathbf{x}_i(t)$, and the random effect covariate vector, $\mathbf{u}_i(t)$, are ultrahigh dimensional, which, in general, are allowed to depend on t, though in our application they are fixed as they refer to SNPs. Under certain circumstances researchers may have prior knowledge about which covariates affect the response mean and which affect the variance. However, in most cases such information is not available in advance, thus $\mathbf{x}_i(t)$ and $\mathbf{u}_i(t)$ may be the same.

The generalized varying coefficient mixed effect model is described as follows:

$$g\big(\mu_i(t)\big) = \beta_0(t) + \sum_{l=1}^{q} \beta_l(t) z_{il}(t) + \sum_{k=1}^{p} \gamma_k(t) x_{ik}(t) + \sum_{m=1}^{r} b_{im}(t) u_{im}(t), \tag{1}$$

where $\mu_i(t) = \mathrm{E}(y_i(t)|\mathbf{b}_i(t), \mathbf{x}_i(t), \mathbf{u}_i(t))$ and $g(\cdot)$ is a known link function that is commonly taken to be the canonical link. For instance, the canonical links for normal, binomial and Poisson distributions are the identity link $g(\mu) = \mu$, logit link $g(\mu) = \text{logit}(\mu)$ and log link $g(\mu) = \log(\mu)$, respectively. Moreover, $\{\beta_l(t), l = 0, 1, \ldots, q\}$ and $\{\gamma_k(t), k = 1, \ldots, p\}$ are fixed-effect nonparametric smooth coefficient functions. The random coefficient functions $\{b_{im}(t), m = 1, \ldots, r; i = 1, \ldots, n\}$ are assumed to be realizations of Gaussian processes with mean zero and covariance function $\zeta_m(s, t) = \text{cov}(b_{im}(s), b_{im}(t))$ and are independent

between subjects. We define two index sets, $\mathcal{M}_0^f$ and $\mathcal{M}_0^r$, for the true fixed and random effects, respectively,

$$\mathcal{M}_0^f = \{1 \le k \le p : \|\gamma_k\|_2 > 0\}, \quad \text{and} \quad \mathcal{M}_0^r = \{1 \le m \le r : \mathrm{E}\|b_{im}\|_2 > 0\}, \tag{2}$$

where $\|\cdot\|_2$ is the functional L_2 norm.

3.2. *Two-step screening procedure*. We propose the following two-step screening procedure to screen both fixed and random effects:

Step 1: Fixed-effect screening

At the first step, we consider a marginal model for the kth x-variable,

$$g\big(\mu_i(t)\big) = \beta_{0k}^*(t) + \sum_{l=1}^{q} \beta_{lk}^*(t) z_{il}(t) + \gamma_k^*(t) x_{ik}(t), \tag{3}$$

where $\{\beta_{lk}^*(t), l = 0, 1, \ldots, q\}$ and $\gamma_k^*(t)$ are smooth coefficient functions and, thus, can be well approximated by B-splines,

$$\beta_{lk}^*(t) \approx \sum_{m=1}^{M_{ln}} \psi_{lm} B_{lm}(t) \quad \text{and} \quad \gamma_k^*(t) \approx \sum_{h=1}^{L_{kn}} \theta_{kh} B_{kh}(t), \tag{4}$$

where $B_{lm}(\cdot)$ and $B_{kh}(\cdot)$ are B-spline basis functions and M_{ln} and L_{kn} are the numbers of basis functions used for $\beta_{lk}^*(t)$ and $\gamma_k^*(t)$, respectively. Therefore, model (3) approximately becomes the following generalized linear model with working-independent covariance structure and can be fitted accordingly,

$$g\big(\mu_i(t)\big) \approx \sum_{m=1}^{M_{0n}} \psi_{0m} B_{0m}(t) + \sum_{l=1}^{q} \sum_{m=1}^{M_{ln}} \psi_{lm} B_{lm}(t) z_{il}(t) + \sum_{h=1}^{L_{kn}} \theta_{kh} B_{kh}(t) x_{ik}(t). \tag{5}$$

For the marginal model (3) and (5) with the kth x-variable, we use the deviance G_k that compares the current fitted value $\hat{\mu}_i(t)$ with the observed $y_i(t)$ through the difference in log likelihood of the fitted model and the saturated model. For normally distributed response with identity link function, the deviance is computed as the residual sums of squares.

Smaller value of G_k typically indicates greater marginal importance of the kth covariate to the response. Thus, we can obtain a subset index for fixed effects $\widehat{\mathcal{M}}_{\tau_n}^f$ by ranking $\{G_k, k = 1, \ldots, p\}$ in an increasing order:

$$\widehat{\mathcal{M}}_{\tau_n}^f = \{1 \le k \le p : G_k \text{ ranks among the first } \tau_n\}, \tag{6}$$

where τ_n is the data-driven threshold for the number of screened fixed-effect variables, typically less than the sample size n.

Step 2: Random-effect screening

In the second step, we carry out random effects screening. Based on the screened fixed effects in (6), we consider the following generalized varying coefficient regression model,

$$g\big(\mu_i(t)\big) = \beta_{0k}^\star(t) + \sum_{l=1}^{q} \beta_{lk}^\star(t) z_{il}(t) + \sum_{k \in \widehat{\mathcal{M}}_{\tau_n}^{(f)}} \gamma_k^\star(t) x_{ik}(t), \tag{7}$$

and calculate the fitted linear predictor $\hat{\eta}_i^\star(t)$ as the estimates of the right-hand side of model (7). Then, the following marginal model can be utilized to measure the importance of the mth random-effect covariate $u_{im}(t)$:

$$g\big(\mu_i(t)\big) = \hat{\eta}_i^\star(t) + b_{im}^*(t) u_{im}(t), \tag{8}$$

where $\hat{\eta}_i^\star(t)$ enters the model as an offset and $b_{im}^*(t)$ is a Gaussian process with mean zero and covariance function $\zeta_m^*(t,s)$. The marginal importance of the mth u-variable can be evaluated via the magnitude of v_m:

$$v_m = \int \operatorname{var}\{b_{im}^*(t)\}u_{im}^2(t)\,dt = \int E\{b_{im}^*(t)^2\}u_{im}^2(t)\,dt. \tag{9}$$

To estimate v_m in (9), we approximate $b_{im}^*(t)$ using B-splines, with $b_{im}^*(t) \approx \sum_{s=1}^{Sm} \alpha_{ims} \times B_{ms}(t)$; this transforms model (8) into a generalized linear mixed-effects model

$$g\big(\mu_i(t)\big) \approx \hat{\eta}_i^\star(t) + \sum_{s=1}^{S_m} \alpha_{ims} B_{ms}(t) u_{im}(t), \tag{10}$$

where $(\alpha_{im1}, \ldots, \alpha_{imS_m})^T \sim N(0, \mathbf{D}_m)$ and α_{ims}'s are independent across subjects. Thus, v_m in (9) can be estimated by

$$\hat{v}_{nm} = \frac{1}{n}\sum_{i=1}^{n}\left[J_i^{-1} \sum_{j=1}^{J_i} \hat{E}^2\{b_{im}^*(t_{ij})\}u_{im}^2(t_{ij})\right] = \frac{1}{n}\sum_{i=1}^{n} J_i^{-1}\operatorname{tr}(\mathbf{U}_{mi}\widehat{\mathbf{D}}_m\mathbf{U}_{mi}^T), \tag{11}$$

where J_i is the number of visits for subject i, $\mathbf{U}_{mi} = (\mathbf{U}_{mi1}, \ldots, \mathbf{U}_{miJ_i})^T$, $\mathbf{U}_{mij}^T = u_{im}(t_{ij}) \times \mathbf{B}_m(t_{ij})$, $\mathbf{B}_m(t_{ij}) = (B_1(t_{ij}), \ldots, B_{S_m}(t_{ij}))$, t_{ij} is the (transformed) age of subject i at the jth visit, $j = 1, \ldots, J_i$ and $\widehat{\mathbf{D}}_m$ is the maximum likelihood estimator of $\mathbf{D}_m$. Thus, we sort $\hat{v}_{nm}$ in a decreasing order and define the screened random-effect index set as

$$\widehat{\mathcal{M}}_{\xi_n}^r = \{1 \le m \le r : \hat{v}_{nm} \text{ ranks among the first } \xi_n\}, \tag{12}$$

where ξ_n is the threshold number of screened random effects.

In short, we refer this two-step screening procedure for mixed-effect time-varying coefficient models as MEGS in this paper.

4. Framingham Heart Study data analysis. For the Framingham Heart Study (FHS) data, we assume the full model to be

$$\begin{aligned}\mathrm{BMI}_{ij}^* &= \beta_0(\mathrm{age}_{ij}) + \beta_1(\mathrm{age}_{ij})\mathrm{Gender}_i + \beta_2(\mathrm{age}_{ij})\mathrm{Smoke}_{ij} + \beta_3(\mathrm{age}_{ij})\mathrm{Alcohol}_{ij}^* \\ &\quad + \beta_4(\mathrm{age}_{ij})\mathrm{Alcohol}_{ij}^{*2} + \sum_{k=1}^{p}\gamma_k(\mathrm{age}_{ij})\mathrm{SNP}_{ik} + \sum_{m=1}^{p} b_{im}(\mathrm{age}_{ij})\mathrm{SNP}_{im} + \varepsilon_{ij},\end{aligned} \tag{13}$$

where $b_{im}(\mathrm{age}_{ij})$ are random functions of transformed age, $\{\mathrm{SNP}_{ik}, k = 1, \ldots, p\}$ are associated with fixed effects, $\{\mathrm{SNP}_{im}, m = 1, \ldots, p\}$ with random effects and $p = 718{,}867$ as discussed in Section 2.2. We add a quadratic term for alcohol, as Yeomans (2010) showed that moderate amount of alcohol intake actually helps to control body weight but excessive intake can contribute to obesity.

4.1. *Screening.* To reduce the ultrahigh dimensionality, we apply our feature screening procedure from Section 3 to screen both fixed and random effects of SNPs. We first screen through all the SNPs for associations with response mean functions based on the following series of marginal models:

$$\begin{aligned}\mathrm{BMI}_{ij}^* &= \beta_0(\mathrm{age}_{ij}) + \beta_1(\mathrm{age}_{ij})\mathrm{Gender}_i + \beta_2(\mathrm{age}_{ij})\mathrm{Smoke}_{ij} + \beta_3(\mathrm{age}_{ij})\mathrm{Alcohol}_{ij}^* \\ &\quad + \beta_4(\mathrm{age}_{ij})\mathrm{Alcohol}_{ij}^{*2} + \gamma_k(\mathrm{age}_{ij})\mathrm{SNP}_{ik} + \varepsilon_{ij}, \quad k = 1, \ldots, p.\end{aligned}$$

All coefficient functions are approximated by cubic B-splines with equally spaced knots and five basis functions. We adopt G_k to be the residual sum of squares (RSS), since the identity

link function is involved, and select the top $\tau_n = [n^{0.8}/\log(n^{0.8})] = 70$ SNPs by (6). Here, τ_n is chosen following the using a common recommendation from the literature (Liu, Li and Wu (2014)).

At the second step, we proceed to screen random effects. Based on the SNPs remained after the first-step screening, we compute the fitted value

$$\hat{\eta}_i(\text{age}_{ij}) = \widehat{\beta}_0(\text{age}_{ij}) + \widehat{\beta}_1(\text{age}_{ij})\text{Gender}_i + \widehat{\beta}_2(\text{age}_{ij})\text{Smoke}_{ij} + \widehat{\beta}_3(\text{age}_{ij})\text{Alcohol}^*_{ij} + \widehat{\beta}_4(\text{age}_{ij})\text{Alcohol}^{*2}_{ij} + \sum_{k\in\widehat{\mathcal{M}}^{(f)}_{\tau_n}} \hat{\gamma}_k(\text{age}_{ij})\text{SNP}_{ik}.$$

The following marginal model is then constructed as

$$\text{BMI}^*_{ij} = \hat{\eta}_i(\text{age}_{ij}) + b_{im}(\text{age}_{ij})\text{SNP}_{im} + \varepsilon^*_{ij} \tag{14}$$

and approximated by

$$\text{BMI}^*_{ij} \approx \hat{\eta}_i(\text{age}_{ij}) + \sum_{s=1}^{S_m} \alpha_{ims} B_{ms}(\text{age}_{ij})\text{SNP}_{im}(\text{age}_{ij}) + \varepsilon^*_{ij},$$

where $(\alpha_{im1}, \ldots, \alpha_{imS_m})^T \sim N(0, \mathbf{D}_m)$ and $\mathbf{D}_m$ is assumed to be a diagonal matrix. We calculate the marginal utility $\hat{v}_{nm}$ as defined in (11) and keep the top $\xi_n = 70$ SNPs with index set $\widehat{\mathcal{M}}^{(r)}_{\xi_n}$ in (12).

4.2. *Postscreening variable selection*. The screening procedure retains 140 SNPs. We then apply a longitudinal group LASSO (Barber, Reimherr and Schill (2017)) to further select important SNPs that affect the mean response and apply a forward selection to select those SNPs affecting variability.

(1) Group LASSO for fixed effects

Consider the following approximated linear model:

$$\begin{aligned}\text{BMI}^*_{ij} = {} & \sum_{m=1}^{M_{0n}} \psi_{0m} B_{0m}(\text{age}_{ij}) + \sum_{m=1}^{M_{1n}} \psi_{1m} B_{1m}(\text{age}_{ij})\text{Gender}_i \\ & + \sum_{m=1}^{M_{2n}} \psi_{2m} B_{2m}(\text{age}_{ij})\text{Smoke}_{ij} + \sum_{m=1}^{M_{3n}} \psi_{3m} B_{3m}(\text{age}_{ij})\text{Alcohol}_{ij} \\ & + \sum_{m=1}^{M_{4n}} \psi_{4m} B_{4m}(\text{age}_{ij})\text{Alcohol}^2_{ij} + \sum_{k\in\widehat{\mathcal{M}}^{(f)}_{\tau_n}} \sum_{h=1}^{L_{kn}} \theta_{kh} B_{kh}(\text{age}_{ij})\text{SNP}_{ik} + \varepsilon_{ij}.\end{aligned} \tag{15}$$

We apply group LASSO method to select important SNPs without penalizing on the intercept function and coefficient functions of gender, smoke and alcohol terms. The tuning parameter λ is chosen via fivefold cross-validation on minimizing the mean squared prediction error.

The detailed information of the finally selected 52 SNPs are listed in Table 2, where "gene1/gene2" in the "Gene" column indicates that this SNP is intergenic between gene1 and gene2 and the "MAF" column reports the corresponding minor allele frequencies (MAF). The estimated coefficient functions with 95% point-wise confidence band for the baseline predictors, and chosen SNPs are shown in Figures 3 and 4.

The shape of the intercept function in the first plot of Figure 3 shows that BMI increases with age, and the sharpest increase occurs between 30 to 40 years old. Then, it becomes stable between age 40 and 60 and starts to decrease after 60. Previous research has confirmed that those aged 40–60 are most likely to be overweight (Becker et al. (2002), Rand and Kuldau

TABLE 2
52 SNPs selected by group LASSO for BMI

SNP	Chr.	Position	Gene	Risk allele	MAF	SNP	Chr.	Position	Gene	Risk allele	MAF
rs2788611	1	112356814	KCND3	G	48.28%	rs17811806	10	109857751	–	T	12.94%
rs12128031	1	185453884	HMCN1	A	41.76%	rs17774576	10	49292414	C10ORF71	G	12.71%
rs490765	1	57565437	DAB1	T	8.03%	rs592373	11	1890990	LSP1	C	35.32%
rs6701005	1	30644096	PTPRU/MATN1	C	6.11%	rs661348	11	1884062	LSP1	C	42.18%
rs7604277	2	58677274	LINC01122	C	37.06%	rs587961	11	1881256	LSP1	T	33.47%
rs12478317	2	234440320	–	C	24.48%	rs1891288	11	34795957	–	C	6.34%
rs7584036	2	54994399	EML6	A	22.45%	rs11068016	12	116401001	–	G	15.23%
rs11903374	2	81639911	IL1B	T	8.21%	rs7977617	12	104747837	TXNRD1	G	13.49%
rs6552663	4	184524289	–	T	40.9%	rs7152364	14	39272372	LINC00639	G	31.65%
rs12641554	4	183381531	–	A	30.54%	rs8005907	14	96787747	ATG2B	C	8.84%
rs17060468	4	175249433	CEP44	G	22.35%	rs17648549	15	25047848	GABRB3	T	24.35%
rs17007095	4	141293852	–	G	15.25%	rs4785878	16	65955116	CDH5	G	24.97%
rs6879492	5	11205980	CTNND2	A	9.97%	rs17760525	17	56749999	–	A	45.61%
rs199253	6	143004059	LINC01277	C	49.9%	rs2847297	18	12797695	PTPN2	G	33.99%
rs6914292	6	105559027	–	T	45.95%	rs6506838	18	78540074	–	T	30.41%
rs10457128	6	105570101	–	G	37.71%	rs12966015	18	55768664	NFE2L3P1	T	29.39%
rs9446305	6	70888868	B3GAT2	T	22.04%	rs17177666	18	70954324	–	A	12.45%
rs13212642	6	71595571	B3GAT2	G	21.96%	rs4891260	18	71144371	TSHZ1/C18ORF62	T	9.12%
rs17082113	6	120317976	–	T	15.36%	rs405722	18	12797695	MSH5	T	6.65%
rs2529753	7	20838951	SP8/RPL23P8	C	39.09%	rs6510070	19	57450244	–	A	40.18%
rs4722675	7	27243962	HOXA13/HOTTIP	A	7.09%	rs10415880	19	49685899	PRMT1	A	32.51%
rs7859884	9	77417406	TRPM6	T	35.89%	rs12463356	19	50217136	CPT1C	T	20.95%
rs1885167	9	17504515	C9ORF39	A	21.78%	rs4814838	20	19319846	SLC24A3	G	43.63%
rs4244347	10	103452645	BTRC	A	35.63%	rs6053233	20	5242559	CDS2/UBE2D3P1	A	26.33%
rs11192397	10	107060568	SORCS3/YWHAZP5	G	14.84%	rs6014998	20	56003684	RBM38/HMGB1P1	A	12.53%
rs180925	10	115734935	IL13	A	29.31%	rs17624687	22	26178247	MYO18B	T	13.15%

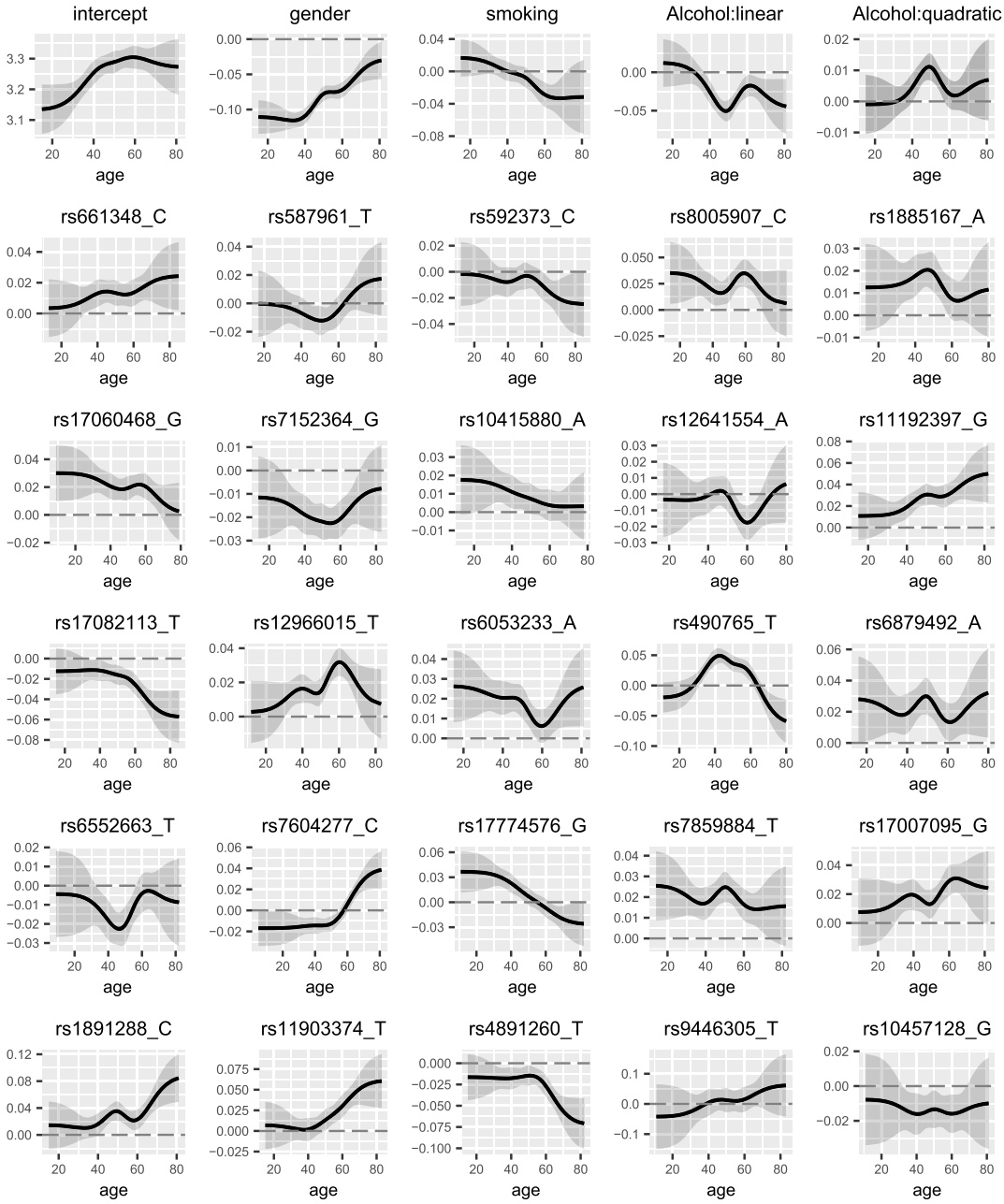


FIG. 3. *Fixed effect coefficients for the covariates and SNPs* 1–25 *selected by group LASSO for BMI.*

(1990), Wang and Beydoun (2007), Ogden et al. (2012)). Examining the gender effect, recall that men are the reference group; we see that women, on average, have a lower BMI than men, but the gap steadily narrows down after age 40. The coefficient function for smoking indicates that smoking before 40 increases BMI, but the effect decreases as subjects' age; after age 50, surprisingly, smoking is negatively associated with BMI. However, the effects of smoking on BMI are not significant at most age ranges except for those between 50 and 70. For alcohol, both linear and quadratic terms have significant time-varying effects. We select five time points and show the alcohol effects in Figure 5. We observe that alcohol intake consistently increases BMI when subjects are young (age 15), while the trend is completely reversed for older subjects (age 60). For the middle-age groups (age 30, 40, 50), moderate (but

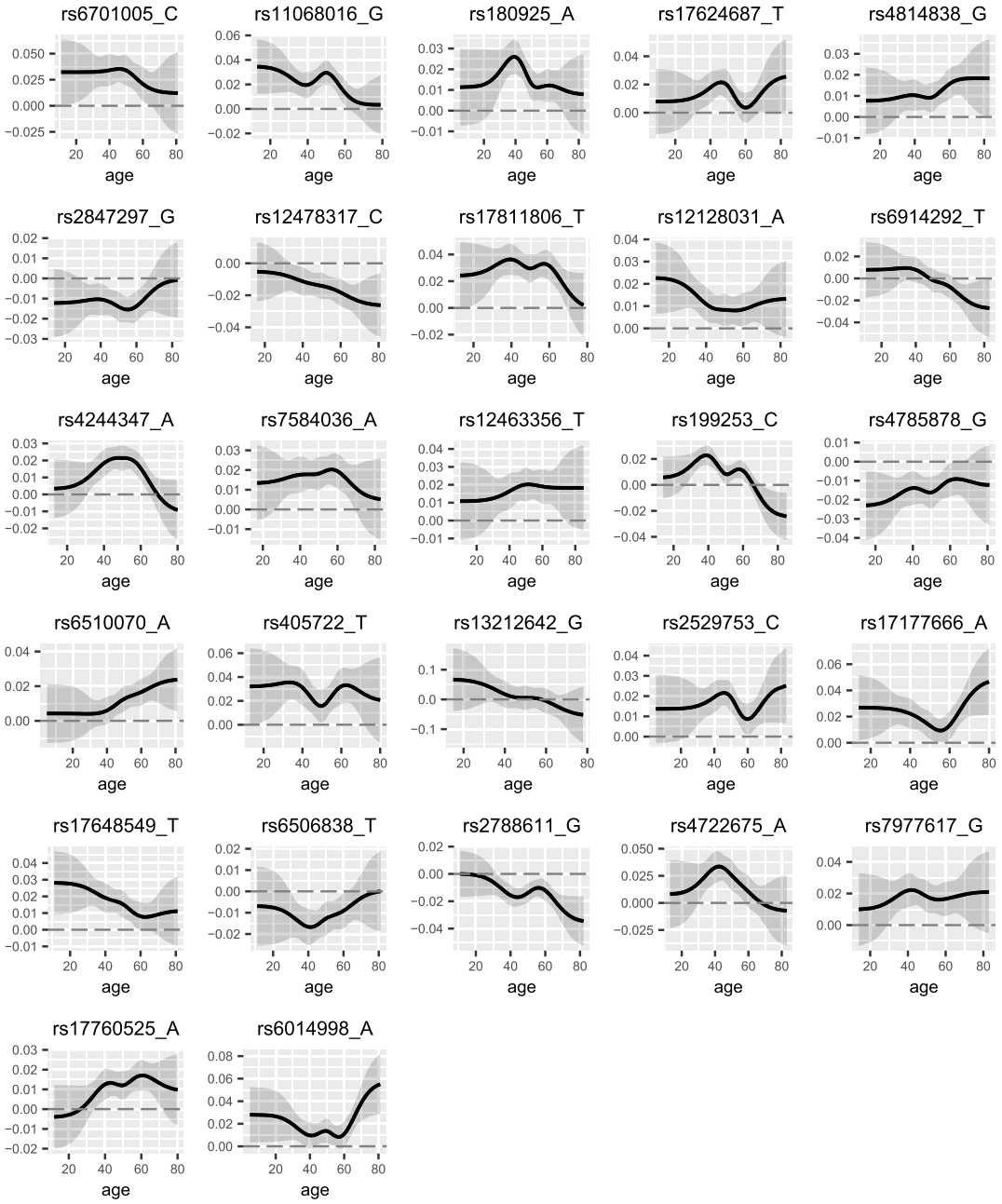


FIG. 4. *Fixed effect coefficients for SNPs* 26–52 *selected by group LASSO for BMI.*

not excessive) alcohol intake could help to control BMI which coincides with the statements in Yeomans (2010).

The coefficients for the selected 52 SNPs take various shapes. Some SNPs (rs11192397, rs17007095, rs1891288, rs11903374, rs6510070, etc.) have significant positive effects on BMI, and the effects increase over time, while some others (rs17060468, rs6701005, rs11068016, rs199253, rs17648549, etc.) have larger impacts at younger ages, but the effects decrease as subjects get older. There are also SNPs that have relatively constant positive (rs6879492) or negative effects (rs10457128) over time. Interestingly, many SNP effects (rs1885167, rs7152364, rs6053233, rs490765, rs180925, rs4722675, etc.) change drastically from age 40 to 60, which is also observed in the linear and quadratic trend of alcohol. This

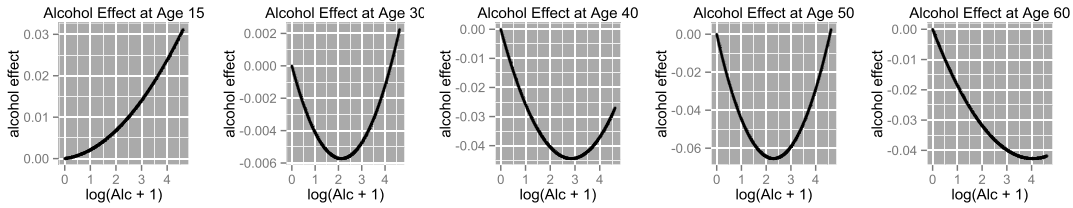


FIG. 5. *Effects of alcohol on BMI at different ages.*

indicates potential interaction effects between genetic markers and individuals' life style such as drinking behavior.

In addition, we looked through previous research results about the selected 52 SNPs. In Table 3 we listed eight SNPs that have been studied in literature. Evidence suggests that the diseases with which these SNPs are associated, such as breast cancer, hypertension and type 2 diabetes, are closely related to BMI and obesity (de Jongh et al. (2004), Weyer et al. (2001), Kaklamani et al. (2008)). Therefore, future research may reveal the mechanisms by which these SNPs influence BMI.

(2) Forward selection of random effects

Next, we apply forward selection to further identify SNPs that are associated with random effects from the 70 SNPs in $\widehat{\mathcal{M}}_{\xi_n}^{(r)}$ obtained from random-effect screening. We start from a model with the baseline covariates (gender, smoking and alcohol behavior) and the 52 selected fixed-effect SNPs (indexed by $\widehat{\mathcal{I}}^{(f)}$). We then add SNPs from $\widehat{\mathcal{M}}_{\xi_n}^{(r)}$, one at a time which result in the largest increase in the restricted log-likelihood (REML). In this way, we can obtain a series of models with random-effect index sets M_q, $q = 1, 2, \ldots$. We then evaluate their mean prediction error using fivefold cross-validation, based on the following model:

$$\begin{aligned}\text{BMI}_{ij}^{*} &= \beta_0(\text{age}_{ij}) + \beta_1(\text{age}_{ij})\text{Gender}_i + \beta_2(\text{age}_{ij})\text{Smoke}_{ij} + \beta_3(\text{age}_{ij})\text{Alcohol}_{ij}^{*} \\ &\quad + \beta_4(\text{age}_{ij})\text{Alcohol}_{ij}^{*2} + \sum_{k\in\widehat{\mathcal{I}}^{(f)}} \gamma_k(\text{age}_{ij})\text{SNP}_{ik} + \sum_{m\in M_q} b_{im}(\text{age}_{ij})\text{SNP}_{im} + \varepsilon_{ij}.\end{aligned} \tag{16}$$

Detailed information about the finally selected five SNPs can be found in Table 4.

According to model (16), the age-varying variance of BMI is computed by the summation over that of all five selected SNPs above, together with the constant variance of the random noise. The estimated variance contribution by the selected five SNPs are depicted in Figure 6. Overall, the effects of all five SNPs on BMI variance increase over age, leading to dramatic rise of the total variance of BMI, and the rise becomes more rapid when people get older. This is consistent with many scientific findings and common sense; BMI, and hence obesity conditions, varies less for people in the early ages, while it diverges more in later life stages. Individally, the contribution of variance by SNP rs4766797 remains steady before age 35 and between age 55–65, while it has a sharp jump around middle age. Thus it is worthwhile studying this SNP for those subjects at age period 35–55. Other SNPs can be interpreted in the similar fashion.

In a case when a replication study is desired, one way is to follow the refitted cross-validation methodology (Fan, Guo and Hao (2012)). Specifically, split the original sample into halves—one half for screening and variable selection, the other half for estimating coefficient functions and drawing inferences; and flip the order. The final results are computed by taking average of the two parts.

5. Simulation studies. We assess the finite sample performance of the proposed two-step screening procedure, MEGS, by Monte Carlo simulation studies. We will consider three

TABLE 3
Some previous findings about selected SNPs for BMI

SNP	Chr.	Position	Gene	Risk allele	MAF[a]	Ranking[b]
rs661348	11	1884062	LSP1	C	42.18%	1
	– Breast cancer studies (Tapper et al. (2008)). – Blood pressure studies (Johnson et al. (2011), Munroe, Barnes and Caulfield (2013)).					
rs592373	11	1890990	LSP1	C	35.32%	3
	– Breast cancer studies (Chen et al. (2015)).					
rs10415880	19	49685899	PRMT1	A	32.51%	8
	– Associated with arteriovenous fistula (AVF) malfunction risk in male hemodialysis (HD) patients (Lee et al. (2015)).					
rs7859884	9	77417406	TRPM6	T	35.89%	22
	– Associated with type 2 diabetes mellitus (Nazıroğlu, Dikici and Dursun ()).					
rs180925	10	115734935	IL13	A	29.31%	34
	– Increase the pneumonitis risk following radiotherapy treatment of nonsmall cell lung cancer (Hildebrandt et al. (2010)).					
rs2847297	18	12797695	PTPN2	G	33.99%	37
	– Associated with risk of Behçet's disease (Wu et al. (2013)). – Correlated with rheumatoid arthritis (RA) (Suzuki et al. (2013)).					
rs405722	18	12797695	MSH5	T	6.65%	53
	– Significantly associated with expression levels of MutS protein homolog 5 (MSH5) gene in lung cancers (Nguyen et al. (2014)).					
rs1150793	18	12797695	MSH5	C	6.37%	66
	– Associated with susceptibility to Kawasaki disease and coronary artery aneurysms (Hsieh et al. (2011)). – Genetic marker for severe cutaneous adverse reactions caused by allopurinol (Hung et al. (2005)).					

a. Minor allele frequency in this study
b. SNP's ranking in the screening procedure.

different scenarios: continuous response in linear regression, binary response in logistic regression and count data response in Poisson regression.

In all situations, we set x- and u-predictors to be the same at the initial stage with $p = 1000$ dimensions. The sample size n is take to be 100 and 200. The number of observations per

TABLE 4
Five SNPs selected for random effects for BMI

SNP	Chromosome	Position	Risk allele	MAF
rs4766797	12	115534970	A	49.77%
rs10121765	9	27362053	A	48.78%
rs2153741	20	2126096–2126097	G	48.70%
rs12471128	2	172863964	A	49.45%
rs4908404	1	28564321	C	49.95%

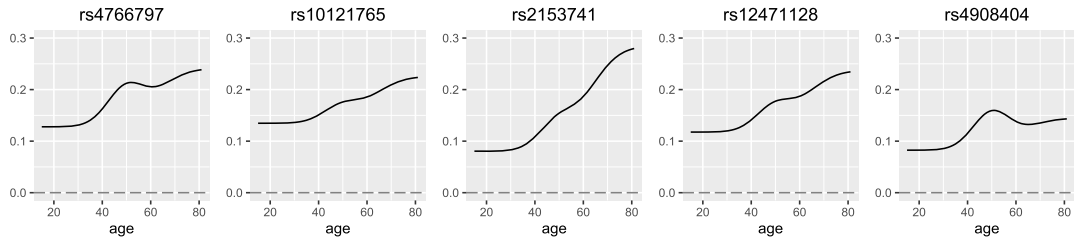


FIG. 6. *Estimated variance functions for the selected five SNPs.*

subject $J_i \equiv J = 10$. The covariate vector $\mathbf{x}_i(t_{ij}) = (x_{i1}(t_{ij}), \ldots, x_{ip}(t_{ij}))^T$ is generated in the following fashion:

$$(t_{ij}^*, \mathbf{x}_{ij}^*) \sim N_{p+1}(0, \Sigma_x), \quad t_{ij} = \Phi(t_{ij}^*), \text{ and } \mathbf{x}_i(t_{ij}) = \mathbf{x}_{ij}^* \tag{17}$$

for $j = 1, \ldots, J$ and $i = 1, \ldots, n$. The (k_1, k_2)th element of Σ_x is set to $\sigma_x \rho_x^{|k_1 - k_2|}$, where $\rho_x = 0.4, 0.8$ and σ_x^2 is the variance of each x-predictor. $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Therefore, t_{ij} is uniformly distributed on $[0, 1]$ and correlated with $\mathbf{x}_i(t_{ij})$. The baseline predictor z_i is generated from a Bernoulli distribution with equal probability.

To evaluate the performance of the proposed method, we employ the following criteria as used in Liu, Li and Wu (2014):

- R_k: The average of ranks of x_k for fixed effects or u_m for random effects in terms of the screening criterion based on 1000 replications.
- M: The minimum size of the submodel so that all true predictors can be selected. The 5%, 25%, 50%, 75% and 95% quantiles of M are reported based on 1000 replications.
- p_a: The proportion of 1000 replications where all true predictors are being selected into $\widehat{\mathcal{M}}_{\tau_n}^{(f)}$ for fixed effects and $\widehat{\mathcal{M}}_{\xi_n}^{(r)}$ for random effects.
- p_k and p_m: The proportion of x_k being selected into the submodel $\widehat{\mathcal{M}}_{\tau_n}^{(f)}$, and u_m being selected into $\widehat{\mathcal{M}}_{\tau_n}^{(r)}$ over 1000 replications.

To calculate p_a, p_k and p_m, we set the selected submodel size $\tau_n = \nu[n/\log n]$ with $\nu = 1, 2, 3$ (Fan and Lv (2008)) at both steps of screening.

We also conduct the GVCM-SIS method proposed by Xia, Yang and Li (2016), which deals with generalized varying coefficient models without longitudinal data structure, and the feature screening procedure for ordinary time-varying coefficient models (Chu, Li and Reimherr (2016)), denoted as TVCM-SIS. Note that these two methods cannot detect random effects; thus, only the fixed effect results are compared. Furthermore, TVCM-SIS are not designed for binary responses and count data; thus, only continuous-response case is considered for this method. GVCM-SIS is designed for generalized varying coefficient model without longitudinal data structure; thus, the within-subject correlation has to be overlooked when implementing this method.

EXAMPLE 1 (Continuous response). We first consider continuous response and generate $y_i(t_{ij})$ by

$$y_i(t_{ij}) = \beta_0(t_{ij}) + \beta_1(t_{ij}) z_i + \sum_{k \in \mathcal{M}^f} \gamma_k(t_{ij}) x_{ik}(t_{ij}) + \sum_{k \in \mathcal{M}^r} b_{ik}(t) x_{ik}(t_{ij}) + \varepsilon_i(t_{ij}).$$

The random errors are drawn from $N(0, 1)$ independently within and across subjects, $\sigma_x = 0.4$. The indices for true fixed and random effects are set to $\mathcal{M}^f = \{100, 400, 700, 900\}$ and

$\mathcal{M}^r = \{300, 800\}$. The fixed coefficient functions are:

$$\beta_0(t) = 0.01t^2, \qquad \beta_1(t) = 0.01t + 0.01t^3, \qquad \gamma_{100}(t) = 0.07 \times \mathbf{1}(t > 0.35) + 0.14,$$
$$\gamma_{400}(t) = 0.06\cos(2\pi t - \pi) + 0.15, \qquad \gamma_{700}(t) = 0.07(1-t)^2 - 0.18,$$
$$\gamma_{900}(t) = -0.06\sin(2\pi t) - 0.15.$$

The random effects are generated as follow:

$$b_{i300}(t) = a_{i1}(t)\big(-0.08\cos(\pi t) - 0.36\big), \qquad b_{i600}(t) = a_{i2}(t)\big(0.08\sin(2\pi t) + 0.3\big),$$

where $\boldsymbol{a}_i(t) = (a_{i1}(t), a_{i2}(t))^T \sim N_2(0, \boldsymbol{\Sigma}_a)$ with $\operatorname{corr}(\boldsymbol{a}_i(t_{ij}), \boldsymbol{a}_i(t_{ij'})) = 0.4^{|j-j'|}$ and

$$\boldsymbol{\Sigma}_a = \begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix}.$$

Tables 5, 6 and 7 report the results. Table 5 consists of the rankings of fixed and random effects for the truly active variables. The newly proposed MEGS performs well and outperforms both GVCM-SIS and TVCM-SIS in all cases, especially when $n = 200$ and $\rho_x = 0.4$: All four fixed effects have average rankings less than 6, and both random effects have average rankings less than 5. The findings imply the ranking consistency property (Zhu et al. (2011)). While the other two methods fail to detect the truly important predictors due to the neglect of within-subject correlation and random effects. When ρ_x increases from 0.4 to 0.8, the average rankings for fixed effects reasonably and slightly increase, but almost remain the same for random effects. Table 6 shows the minimum number of predictors that is required to ensure the inclusion of all truly active predictors. We observe that when sample size $n = 200$, fixed effects ranked top seven and random effects ranked top five among the 1000 predictors in more than 75% of the replications, indicating high accuracy of the screening procedure. The selection proportions are reported in Table 7, and the conclusions are consistent. At sample $n = 200$ and for predictors with both small and large correlations, all true fixed and random effects are retained in more than 900 replications using threshold $\tau_n = 37$.

EXAMPLE 2 (Count response). In this simulation study, we consider count response with Poisson distribution. The fixed and random coefficient functions are generated in the same

TABLE 5
Average ranks R_j of the active predictors: continuous response

			Fixed effects				Random effects	
ρ_x	n	Method	x_{100}	x_{400}	x_{700}	x_{900}	x_{300}	x_{800}
0.4	100	GVCM-SIS	145.375	164.268	223.671	230.735	–	–
		TVCM-SIS	111.815	308.166	190.303	218.460	–	–
		MEGS	13.217	28.221	40.413	39.536	34.393	35.303
	200	GVCM-SIS	41.582	45.546	94.339	110.527	–	–
		TVCM-SIS	33.542	197.033	58.437	67.695	–	–
		MEGS	2.082	3.536	4.602	5.791	4.525	3.670
0.8	100	GVCM-SIS	144.280	169.260	226.535	229.090	–	–
		TVCM-SIS	126.656	325.510	183.042	230.505	–	–
		MEGS	16.744	35.364	43.884	46.349	32.646	26.950
	200	GVCM-SIS	43.134	49.412	109.050	118.953	–	–
		TVCM-SIS	31.152	203.887	66.340	66.946	–	–
		MEGS	2.667	6.374	8.316	8.924	4.251	4.262

TABLE 6
Quantiles of M: continuous response

			Fixed effects					Random effects				
ρ_x	n	Method	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
0.4	100	GVCM-SIS	87.95	229.75	418.00	665.25	895.05	–	–	–	–	–
		TVCM-SIS	119.95	310.25	451.00	669.75	910.20	–	–	–	–	–
		MEGS	5.00	15.00	43.00	113.25	372.10	2.00	5.00	16.00	63.00	317.15
	200	GVCM-SIS	14.00	53.75	139.00	301.00	663.40	–	–	–	–	–
		TVCM-SIS	20.85	58.25	183.00	391.75	735.00	–	–	–	–	–
		MEGS	4.00	4.00	4.00	7.00	33.00	2.00	2.00	2.00	3.25	29.05
0.8	100	GVCM-SIS	91.00	232.00	414.00	646.50	917.15	–	–	–	–	–
		TVCM-SIS	139.00	321.50	515.50	676.75	890.95	–	–	–	–	–
		MEGS	8.00	22.00	51.00	133.00	434.05	2.00	5.00	15.00	52.00	262.05
	200	GVCM-SIS	20.00	62.00	154.00	333.25	682.15	–	–	–	–	–
		TVCM-SIS	14.01	76.56	211.01	415.53	741.34	–	–	–	–	–
		MEGS	5.00	7.00	11.00	17.00	43.05	2.00	2.00	3.00	5.00	20.00

way as Example 1. The data are generated via

$$\log\bigl(\mu_i(t_{ij})\bigr) = \beta_0(t_{ij}) + \beta_1(t_{ij})z_i + \sum_{k\in\mathcal{M}^f} \gamma_k(t_{ij})x_{ik}(t_{ij}) + \sum_{k\in\mathcal{M}^r} b_{ik}(t)x_{ik}(t_{ij}).$$

Table 8 reports the average rankings for all active predictors in different settings, by comparing MEGS and GVCM-SIS. For MEGS, both the fixed and random effects have high average rankings, especially at sample size 200. The minimum model sizes in Table 9 show that the 95% quantiles for both fixed and random effects sets are relatively large around 30 and 46, indicating unstable performance for a few replications. But for most of the simulation replications, the screening procedure can select true predictors with relatively small model sizes. The selection rates reported in Table 10 confirm previous observations: the fixed and random effects are retained into the subset with proportions very close to 1 at sample size 200 using threshold 16. The screening procedure performs slightly but reasonably worse at large correlation ($\rho_x = 0.8$) scenarios, especially for fixed effect x_{900}. Again, GVCM-SIS fails to identify most important predictors by ignoring the within-subject correlation.

EXAMPLE 3 (Binary response). Logistic regression is a commonly used tool for modeling binary response. We consider the binary outcome with the following time-varying coefficient mixed effects model:

$$\log\biggl(\frac{\mu_i(t_{ij})}{1-\mu_i(t_{ij})}\biggr) = \beta_{0k}(t_{ij}) + \beta_1(t_{ij})z_i + \sum_{k\in\mathcal{M}^f} \gamma_k(t_{ij})x_{ik}(t_{ij}) + \sum_{k\in\mathcal{M}^r} b_{ik}(t)x_{ik}(t_{ij}).$$

The fixed coefficient functions are:

$$\beta_0(t) = 0.01t^2, \qquad \beta_1(t) = 0.01t + 0.01t^3, \qquad \gamma_{100}(t) = 0.14 \times \mathbf{1}(t > 0.35) + 0.28,$$
$$\gamma_{400}(t) = 0.12\cos(2\pi t - \pi) + 0.3, \qquad \gamma_{700}(t) = 0.14(1-t)^2 - 0.36,$$
$$\gamma_{900}(t) = -0.12\sin(2\pi t) - 0.3.$$

The random effects are generated as follow:

$$b_{i300}(t) = a_{i1}(t)\bigl(-0.16\cos(\pi t) - 0.72\bigr), \qquad b_{i600}(t) = a_{i2}(t)\bigl(0.16\sin(2\pi t) + 0.6\bigr),$$

TABLE 7
Selection proportion p_j of the active predictors: continuous response

				Fixed effects					Random effects		
ρ_x	n	τ_n & ξ_n	Method	p_{100}	p_{400}	p_{700}	p_{900}	$p_a^{(f)}$	p_{300}	p_{800}	$p_a^{(r)}$
0.4	100	21	GVCM-SIS	0.270	0.213	0.159	0.134	0.000	–	–	–
			TVCM-SIS	0.288	0.021	0.135	0.167	0.000	–	–	–
			MEGS	0.898	0.748	0.710	0.717	0.339	0.740	0.736	0.553
		42	GVCM-SIS	0.358	0.297	0.218	0.183	0.002	–	–	–
			TVCM-SIS	0.387	0.035	0.210	0.272	0.000	–	–	–
			MEGS	0.937	0.833	0.807	0.801	0.498	0.814	0.821	0.674
		63	GVCM-SIS	0.407	0.362	0.261	0.234	0.003	–	–	–
			TVCM-SIS	0.501	0.076	0.277	0.282	0.001	–	–	–
			MEGS	0.955	0.885	0.850	0.856	0.620	0.867	0.870	0.754
	200	37	GVCM-SIS	0.654	0.633	0.430	0.370	0.060	–	–	–
			TVCM-SIS	0.721	0.252	0.480	0.568	0.049	–	–	–
			MEGS	0.997	0.995	0.985	0.982	0.959	0.978	0.986	0.964
		74	GVCM-SIS	0.769	0.733	0.531	0.494	0.151	–	–	–
			TVCM-SIS	0.798	0.334	0.627	0.695	0.112	–	–	–
			MEGS	0.998	0.999	0.996	0.989	0.982	0.989	0.993	0.982
		111	GVCM-SIS	0.809	0.793	0.609	0.570	0.228	–	–	–
			TVCM-SIS	0.848	0.387	0.726	0.766	0.211	–	–	–
			MEGS	1.000	0.999	0.997	0.994	0.990	0.993	0.998	0.991
0.8	100	21	GVCM-SIS	0.243	0.188	0.140	0.116	0.000	–	–	–
			TVCM-SIS	0.273	0.032	0.111	0.116	0.000	–	–	–
			MEGS	0.874	0.699	0.662	0.659	0.242	0.768	0.780	0.585
		42	GVCM-SIS	0.324	0.262	0.207	0.181	0.000	–	–	–
			TVCM-SIS	0.344	0.045	0.191	0.178	0.000	–	–	–
			MEGS	0.929	0.809	0.772	0.780	0.449	0.839	0.851	0.712
		63	GVCM-SIS	0.381	0.319	0.263	0.226	0.001	–	–	–
			TVCM-SIS	0.412	0.085	0.309	0.196	0.000	–	–	–
			MEGS	0.949	0.863	0.827	0.832	0.561	0.877	0.901	0.785
	200	37	GVCM-SIS	0.630	0.593	0.373	0.332	0.032	–	–	–
			TVCM-SIS	0.735	0.264	0.482	0.331	0.081	–	–	–
			MEGS	0.999	0.987	0.981	0.970	0.937	0.986	0.989	0.976
		74	GVCM-SIS	0.735	0.707	0.502	0.467	0.113	–	–	–
			TVCM-SIS	0.826	0.366	0.617	0.521	0.094	–	–	–
			MEGS	1.000	0.997	0.994	0.992	0.983	0.992	0.995	0.987
		111	GVCM-SIS	0.785	0.766	0.564	0.536	0.182	–	–	–
			TVCM-SIS	0.879	0.422	0.711	0.663	0.151	–	–	–
			MEGS	1.000	0.999	0.995	0.995	0.989	0.996	0.997	0.993

where $\boldsymbol{a}_i(t) = (a_{i1}(t), a_{i2}(t))^T \sim N_2(0, \boldsymbol{\Sigma}_a)$ with $\text{corr}(\boldsymbol{a}_i(t_{ij}), \boldsymbol{a}_i(t_{ij'})) = 0.4^{|j-j'|}$ and

$$\boldsymbol{\Sigma}_a = \begin{pmatrix} 2 & 0.2 \\ 0.2 & 2 \end{pmatrix}.$$

Tables 11, 12 and 13 show the results using the evaluation criteria aforementioned. We focus on MEGS, as GVCM-SIS suffers from the same drawback as before. At sample size $n = 100$ in Table 11, the random effects have better ranking results than the fixed effects; when n increases to 200, both fixed and random effects have excellent average rankings.

TABLE 8
Average ranks R_j of the active predictors: Poisson regression

			Fixed effects				Random effects	
ρ_x	n	Method	x_{100}	x_{400}	x_{700}	x_{900}	x_{300}	x_{800}
0.4	100	GVCM-SIS	120.348	156.401	221.250	223.256	–	–
		MEGS	14.522	34.366	42.952	33.993	27.263	21.509
	200	GVCM-SIS	39.794	48.215	86.206	109.687	–	–
		MEGS	2.166	4.771	5.189	7.088	6.816	5.80
0.8	100	GVCM-SIS	127.993	162.958	225.539	226.079	–	–
		MEGS	15.801	37.947	45.937	42.978	24.283	22.775
	200	GVCM-SIS	41.119	50.452	95.871	114.736	–	–
		MEGS	3.443	80	9.820	10.225	8.382	6.548

Table 12 shows the same pattern as that in Table 7, where the minimum model sizes to include all random effects are small at all quantiles and are slightly larger for fixed effects. High correlations among predictors give slightly worse results, but the differences are not significant. The selection proportions for each active predictor and the true model are reported in Table 13, where the selection rates are all close to 1 at sample size 200. Thus, our two-step approach performs very well in selecting both fixed and random effects for binary outcome using logistic regression with time-varying coefficients.

6. Summary. In this work we proposed a two-step screening procedure for time-varying coefficient mixed-effects models. We applied our procedure to analyze data from Framingham Heart Study and used body mass index (BMI) to study the genetic and environmental effects on obesity. We further selected fixed-effect SNPs by group LASSO and random-effect SNPs using forward regression. Many of the selected SNPs had previously been identified in the literature, and we observed some novel time-varying patterns for both fixed and random effects. We also applied causal inference techniques and compared the causal effect estimates of the SNPs by our method with those in literature. Results show that the SNPs we found have a stronger causal influence on BMI than those that are previously identified. As future work, the interaction between fixed and random effects can be further explored by including the

TABLE 9
Quantiles of M: Poisson regression

			Fixed effects					Random effects				
ρ_x	n	Method	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
0.4	100	GVCM-SIS	74.80	231.00	388.00	613.25	870.00	–	–	–	–	–
		MEGS	5.00	14.00	40.00	113.00	422.05	2.00	3.00	12.00	42.25	214.10
	200	GVCM-SIS	12.95	54.75	142.50	306.75	642.00	–	–	–	–	–
		MEGS	4.00	4.00	4.00	7.00	39.05	2.00	2.00	2.00	4.00	46.00
0.8	100	GVCM-SIS	84.90	236.00	415.00	624.25	895.20	–	–	–	–	–
		MEGS	9.00	22.00	54.50	137.00	392.10	2.00	4.00	11.00	42.00	205.15
	200	GVCM-SIS	19.00	66.00	146.50	309.75	673.20	–	–	–	–	–
		MEGS	5.00	8.00	11.00	17.00	57.05	2.00	2.00	3.00	6.00	44.10

TABLE 10
Selection proportion p_j of the active predictors: Poisson regression

				Fixed effects					Random effects		
ρ_x	n	τ_n & ξ_n	Method	p_{100}	p_{400}	p_{700}	p_{900}	$p_a^{(f)}$	p_{300}	p_{800}	$p_a^{(r)}$
0.4	100	21	GVCM-SIS	0.290	0.219	0.135	0.135	0.000	–	–	–
			MEGS	0.880	0.730	0.770	0.714	0.354	0.787	0.824	0.631
		42	GVCM-SIS	0.371	0.310	0.206	0.213	0.006	–	–	–
			MEGS	0.930	0.820	0.826	0.815	0.517	0.863	0.881	0.750
		63	GVCM-SIS	0.433	0.363	0.254	0.250	0.012	–	–	–
			MEGS	0.952	0.873	0.860	0.864	0.621	0.894	0.910	0.806
	200	37	GVCM-SIS	0.685	0.632	0.461	0.411	0.074	–	–	–
			MEGS	0.997	0.985	0.985	0.971	0.945	0.972	0.969	0.941
		74	GVCM-SIS	0.781	0.732	0.590	0.525	0.167	–	–	–
			MEGS	0.999	0.993	0.993	0.988	0.977	0.982	0.982	0.964
		111	GVCM-SIS	0.827	0.778	0.651	0.580	0.230	–	–	–
			MEGS	1.000	0.996	0.995	0.993	0.986	0.988	0.991	0.979
0.8	100	21	GVCM-SIS	0.242	0.203	0.122	0.117	0.000	–	–	–
			MEGS	0.851	0.687	0.718	0.637	0.248	0.784	0.821	0.620
		42	GVCM-SIS	0.358	0.286	0.199	0.184	0.004	–	–	–
			MEGS	0.929	0.789	0.795	0.759	0.422	0.857	0.891	0.751
		63	GVCM-SIS	0.426	0.342	0.238	0.234	0.008	–	–	–
			MEGS	0.948	0.837	0.837	0.830	0.541	0.896	0.915	0.812
	200	37	GVCM-SIS	0.659	0.583	0.409	0.361	0.038	–	–	–
			MEGS	0.998	0.985	0.967	0.964	0.923	0.966	0.979	0.945
		74	GVCM-SIS	0.765	0.696	0.527	0.497	0.115	–	–	–
			MEGS	0.999	0.991	0.982	0.990	0.966	0.980	0.988	0.968
		111	GVCM-SIS	0.803	0.759	0.600	0.568	0.186	–	–	–
			MEGS	0.999	0.997	0.986	0.994	0.980	0.984	0.992	0.976

TABLE 11
Average ranks R_j of the active predictors: Logistic regression

			Fixed effects				Random effects	
ρ_x	n	Method	x_{100}	x_{400}	x_{700}	x_{900}	x_{300}	x_{800}
0.4	100	GVCM-SIS	167.077	175.549	244.828	265.104	–	–
		MEGS	31.228	42.946	74.052	40.094	32.295	23.339
	200	GVCM-SIS	53.170	56.677	116.959	130.558	–	–
		MEGS	3.029	5.173	11.145	3.763	3.121	2.738
0.8	100	GVCM-SIS	166.537	191.909	252.176	274.457	–	–
		MEGS	33.515	51.361	72.886	36.448	26.001	18.655
	200	GVCM-SIS	60.623	67.873	121.443	146.156	–	–
		MEGS	4.942	8.145	16.517	6.599	2.992	2.990

TABLE 12
Quantiles of M: Logistic regression

			Fixed effects					Random effects				
ρ_x	n	Method	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
0.4	100	GVCM-SIS	106.9	274.0	471.5	687.0	918.0	–	–	–	–	–
		MEGS	7.00	30.25	81.50	193.00	534.55	2.00	4.00	10.00	49.00	256.55
	200	GVCM-SIS	20.00	78.00	178.00	359.25	721.10	–	–	–	–	–
		MEGS	4.00	4.00	5.00	10.00	55.00	2.00	2.00	2.00	3.00	12.00
0.8	100	GVCM-SIS	105.90	276.75	514.00	712.25	921.00	–	–	–	–	–
		MEGS	12.00	35.00	79.00	187.75	493.20	2.00	4.00	12.00	38.00	195.65
	200	GVCM-SIS	28.00	87.75	191.00	384.00	779.40	–	–	–	–	–
		MEGS	5.00	8.00	13.00	23.00	76.00	2.00	2.00	2.00	4.00	12.00

TABLE 13
Selection proportion p_j of the active predictors: Logistic regression

				Fixed effects					Random effects		
ρ_x	n	τ_n & ξ_n	Method	p_{100}	p_{400}	p_{700}	p_{900}	$p_a^{(f)}$	p_{300}	p_{800}	$p_a^{(r)}$
0.4	100	21	GVCM-SIS	0.194	0.145	0.109	0.074	0.000	–	–	–
			MEGS	0.775	0.671	0.555	0.702	0.185	0.773	0.813	0.634
		42	GVCM-SIS	0.285	0.223	0.174	0.129	0.001	–	–	–
			MEGS	0.841	0.792	0.668	0.788	0.332	0.842	0.860	0.726
		63	GVCM-SIS	0.335	0.295	0.220	0.169	0.002	–	–	–
			MEGS	0.872	0.836	0.718	0.839	0.419	0.881	0.896	0.794
	200	37	GVCM-SIS	0.579	0.546	0.371	0.310	0.032	–	–	–
			MEGS	0.992	0.976	0.947	0.991	0.908	0.990	0.989	0.980
		74	GVCM-SIS	0.695	0.678	0.471	0.435	0.096	–	–	–
			MEGS	1.000	0.993	0.979	0.998	0.971	0.995	0.998	0.993
		111	GVCM-SIS	0.760	0.744	0.543	0.508	0.153	–	–	–
			MEGS	1.000	0.998	0.986	0.998	0.981	0.996	1.000	0.996
0.8	100	21	GVCM-SIS	0.165	0.168	0.095	0.079	0.000	–	–	–
			MEGS	0.744	0.635	0.537	0.690	0.150	0.780	0.834	0.650
		42	GVCM-SIS	0.256	0.232	0.153	0.129	0.000	–	–	–
			MEGS	0.817	0.744	0.654	0.811	0.308	0.863	0.888	0.761
		63	GVCM-SIS	0.310	0.283	0.202	0.169	0.005	–	–	–
			MEGS	0.866	0.795	0.716	0.856	0.430	0.911	0.925	0.840
	200	37	GVCM-SIS	0.530	0.519	0.326	0.261	0.015	–	–	–
			MEGS	0.990	0.969	0.921	0.976	0.861	0.995	0.990	0.986
		74	GVCM-SIS	0.651	0.639	0.452	0.388	0.065	–	–	–
			MEGS	0.999	0.989	0.965	0.993	0.946	0.998	0.996	0.994
		111	GVCM-SIS	0.714	0.701	0.522	0.461	0.120	–	–	–
			MEGS	0.999	0.994	0.975	0.996	0.965	0.998	0.998	0.997

products between the fixed-effect SNPs and random-effect SNPs in the model. The interaction terms are also random that account for variance of BMI. In addition, interaction terms between the behavioral variables, such as smoking, and SNPs can reveal the mediation effects of these behavioral variables. All simulation results indicate the capability of this two-step screening approach for effectively reducing feature dimensions that are associated with both fixed and random effects. It is of interest to establish sure screening property of the proposed feature screening procedure, but it seems to be very challenging and out of scope of this paper. This might be a good research topic for future research.

Acknowledgments. The authors thank the Editor, Associate Editor and reviewers for their constructive comments and Mr. Xiang Li in School of Economics, Xiamen University for his technical support. All authors equally contributed to this paper, and the authors are listed in alphabetical order.

Li's research is supported by NIDA Grant P50-DA 039838, NIH Grant R01CA229542 and NSF Grant DMS-1820702.

Liu is corresponding author, and her research is supported by NNSFC Grant 11771361, 11871409, 11671334 and JAS14007.

Reimherr's research is supported by NSF Grant DMS-1712826 and NIDA Grant P50-DA 039838.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIDA, the NIH, the NNSFC or the NSF.

SUPPLEMENTARY MATERIAL

Supplement to "Feature selection for generalized varying coefficient mixed-effect models with application to obesity GWAS" (DOI: 10.1214/19-AOAS1310SUPP; .pdf). The causal inference results of the selected fixed and random effects of SNPs are given in the online supplement.

REFERENCES

ALLISON, D. B., KAPRIO, J., KORKEILA, M., KOSKENVUO, M., NEALE, M. C. and HAYAKAWA, K. (1996). The heritability of body mass index among an international sample of monozygotic twins reared apart. *Int. J. Obes.* **20** 501–506.

ASCHARD, H., ZAITLEN, N., TAMIMI, R. M., LINDSTRÖM, S. and KRAFT, P. (2013). A nonparametric test to detect quantitative trait loci where the phenotypic distribution differs by genotypes. *Genet. Epidemiol.* **37** 323–333.

BARBER, R. F., REIMHERR, M. and SCHILL, T. (2017). The function-on-scalar LASSO with applications to longitudinal GWAS. *Electron. J. Stat.* **11** 1351–1389. MR3635916 https://doi.org/10.1214/17-EJS1260

BECKER, J., NORA, D. B., GOMES, I., STRINGARI, F. F., SEITENSUS, R., PANOSSO, J. S. and EHLERS, J. A. C. (2002). An evaluation of gender, obesity, age and diabetes mellitus as risk factors for carpal tunnel syndrome. *Clin. Neurophysiol.* **113** 1429–1434.

BELL, C. G., WALLEY, A. J. and FROGUEL, P. (2005). The genetics of human obesity. *Nat. Rev. Genet.* **6** 221–234.

CHEN, H., QI, X., QIU, P. and ZHAO, J. (2015). Correlation between LSP1 polymorphisms and the susceptibility to breast cancer. *Int. J. Clin. Exp. Pathol.* **8** 5798–5802.

CHU, W., LI, R. and REIMHERR, M. (2016). Feature screening for time-varying coefficient models with ultrahigh-dimensional longitudinal data. *Ann. Appl. Stat.* **10** 596–617. MR3528353 https://doi.org/10.1214/16-AOAS912

CHU, W., LI, R., LIU, J. and REIMHERR, M. (2020). Supplement to "Feature selection for generalized varying coefficient mixed-effect models with application to obesity GWAS." https://doi.org/10.1214/19-AOAS1310SUPP.

DE JONGH, R. T., SERNÉ, E. H., IJZERMAN, R. G., DE VRIES, G. and STEHOUWER, C. D. (2004). Impaired microvascular function in obesity implications for obesity-associated microangiopathy, hypertension, and insulin resistance. *Circulation* **109** 2529–2535.

FALL, T. and INGELSSON, E. (2014). Genome-wide association studies of obesity and metabolic syndrome. *Mol. Cell. Endocrinol.* **382** 740–757. https://doi.org/10.1016/j.mce.2012.08.018

FAN, J., FENG, Y. and SONG, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *J. Amer. Statist. Assoc.* **106** 544–557. MR2847969 https://doi.org/10.1198/jasa.2011.tm09779

FAN, J., GUO, S. and HAO, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 37–65. MR2885839 https://doi.org/10.1111/j.1467-9868.2011.01005.x

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911. MR2530322 https://doi.org/10.1111/j.1467-9868.2008.00674.x

FAN, J., MA, Y. and DAI, W. (2014). Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *J. Amer. Statist. Assoc.* **109** 1270–1284. MR3265696 https://doi.org/10.1080/01621459.2013.879828

FURLOTTE, N. A. and ESKIN, E. (2015). Efficient multiple-trait association and estimation of genetic correlation using the matrix-variate linear mixed model. *Genetics* **200** 59–68. https://doi.org/10.1534/genetics.114.171447

GEILER-SAMEROTTE, K., BAUER, C., LI, S., ZIV, N., GRESHAM, D. and SIEGAL, M. (2013). The details in the distributions: Why and how to study phenotypic variability. *Curr. Opin. Biotechnol.* **24** 752–759.

HASTIE, T. and TIBSHIRANI, R. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. Ser. B* **55** 757–796. MR1229881

HERRERA, B. M., KEILDSON, S. and LINDGREN, C. M. (2011). Genetics and epigenetics of obesity. *Maturitas* **69** 41–49.

HILDEBRANDT, M. A., KOMAKI, R., LIAO, Z., GU, J., CHANG, J. Y., YE, Y., LU, C., STEWART, D. J., MINNA, J. D. et al. (2010). Genetic variants in inflammation-related genes are associated with radiation-induced toxicity following treatment for non-small cell lung cancer. *PLoS ONE* **5** e12402.

HSIEH, Y.-Y., CHANG, C.-C., HSU, C.-M., CHEN, S.-Y., LIN, W.-H. and TSAI, F.-J. (2011). Major histocompatibility complex class I chain-related gene polymorphisms: Associated with susceptibility to Kawasaki disease and coronary artery aneurysms. *Genet. Test. Mol. Biomark.* **15** 755–763.

HUNG, S.-I., CHUNG, W.-H., LIOU, L.-B., CHU, C.-C., LIN, M., HUANG, H.-P., LIN, Y.-L., LAN, J.-L., YANG, L.-C. et al. (2005). HLA-B* 5801 allele as a genetic marker for severe cutaneous adverse reactions caused by allopurinol. *Proc. Natl. Acad. Sci. USA* **102** 4134–4139.

JOHNSON, T., GAUNT, T. R., NEWHOUSE, S. J., PADMANABHAN, S., TOMASZEWSKI, M., KUMARI, M., MORRIS, R. W., TZOULAKI, I., O'BRIEN, E. T. et al. (2011). Blood pressure loci identified with a gene-centric array. *Am. J. Hum. Genet.* **89** 688–700. https://doi.org/10.1016/j.ajhg.2011.10.013

KAKLAMANI, V. G., SADIM, M., HSI, A., OFFIT, K., ODDOUX, C., OSTRER, H., AHSAN, H., PASCHE, B. and MANTZOROS, C. (2008). Variants of the adiponectin and adiponectin receptor 1 genes and breast cancer risk. *Cancer Res.* **68** 3178–3184.

KELLY, T., YANG, W., CHEN, C.-S., REYNOLDS, K. and HE, J. (2008). Global burden of obesity in 2005 and projections to 2030. *Int. J. Obes.* **32** 1431–1437.

LEE, K.-H., TSAI, W.-J., CHEN, Y.-W., YANG, W.-C., LEE, C.-Y., OU, S.-M., CHEN, Y.-T., CHIEN, C.-C., LEE, P.-C. et al. (2015). Genotype polymorphisms of genes regulating nitric oxide synthesis determine long-term arteriovenous fistula patency in male hemodialysis patients. *Ren. Fail.* **38** 1–10.

LIU, J., LI, R. and WU, R. (2014). Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *J. Amer. Statist. Assoc.* **109** 266–274. MR3180562 https://doi.org/10.1080/01621459.2013.850086

LIU, J., ZHONG, W. and LI, R. (2015). A selective overview of feature screening for ultrahigh-dimensional data. *Sci. China Math.* **58** 2033–2054. MR3400642 https://doi.org/10.1007/s11425-015-5062-9

MUNROE, P. B., BARNES, M. R. and CAULFIELD, M. J. (2013). Advances in blood pressure genomics. *Circ. Res.* **112** 1365–1379.

NAZIROĞLU, M., DIKICI, D. M. and DURSUN, S. Role of oxidative stress and Ca^{2+} signaling on molecular pathways of neuropathic pain in diabetes: Focus on TRP channels.

NGUYEN, J. D. U., LAMONTAGNE, M., COUTURE, C., CONTI, M., PARÉ, P. D., SIN, D. D., HOGG, J. C., NICKLE, D., POSTMA, D. S. et al. (2014). Susceptibility loci for lung cancer are associated with mRNA levels of nearby genes in the lung. *Carcinogenesis* **35** 2653–2659.

OGDEN, C. L., CARROLL, M. D., KIT, B. K. and FLEGAL, K. M. (2012). Prevalence of obesity in the United States, 2009–2010.

PARÉ, G., COOK, N. R., RIDKER, P. M. and CHASMAN, D. I. (2010). On the use of variance per genotype as a tool to identify quantitative trait interaction effects: A report from the Women's Genome Health Study. *PLoS Genet.* **6** e1000981. https://doi.org/10.1371/journal.pgen.1000981

RAMACHANDRAPPA, S. and FAROOQI, I. S. (2011). Genetic approaches to understanding human obesity. *J. Clin. Invest.* **121** 2080–2086.

RAND, C. S. and KULDAU, J. M. (1990). The epidemiology of obesity and self-defined weight problem in the general population: Gender, race, age, and social class. *Int. J. Eat. Disord.* **9** 329–343.

RANKINEN, T., ZUBERI, A., CHAGNON, Y. C., WEISNAGEL, S. J., ARGYROPOULOS, G., WALTS, B., PÉRUSSE, L. and BOUCHARD, C. (2006). The human obesity gene map: The 2005 update. *Obesity* **14** 529–644. https://doi.org/10.1038/oby.2006.71

SOAVE, D., CORVOL, H., PANJWANI, N., GONG, J., LI, W., BOËLLE, P.-Y., DURIE, P. R., PATERSON, A. D., ROMMENS, J. M. et al. (2015). A joint location-scale test improves power to detect associated SNPs, gene sets, and pathways. *Am. J. Hum. Genet.* **97** 125–138.

SONG, R., YI, F. and ZOU, H. (2014). On varying-coefficient independence screening for high-dimensional varying-coefficient models. *Statist. Sinica* **24** 1735–1752. MR3308660

SUZUKI, T., IKARI, K., YANO, K., INOUE, E., TOYAMA, Y., TANIGUCHI, A., YAMANAKA, H. and MOMOHARA, S. (2013). PADI4 and HLA-DRB1 are genetic risks for radiographic progression in RA patients, independent of ACPA status: Results from the IORRA cohort study. *PLoS ONE* **8** e61045. https://doi.org/10.1371/journal.pone.0061045

TAPPER, W., HAMMOND, V., GERTY, S., ENNIS, S., SIMMONDS, P., COLLINS, A., ECCLES, D. and PROSPECTIVE STUDY OF OUTCOMES IN SPORADIC VERSUS HEREDITARY BREAST CANCER (POSH) STEERING GROUP (2008). The influence of genetic variation in 30 selected genes on the clinical characteristics of early onset breast cancer. *Breast Cancer Res.* **10** 1–10.

WANG, Y. and BEYDOUN, M. A. (2007). The obesity epidemic in the United States—gender, age, socioeconomic, racial/ethnic, and geographic characteristics: A systematic review and meta-regression analysis. *Epidemiol. Rev.* **29** 6–28. https://doi.org/10.1093/epirev/mxm007

WARDLE, J., CARNELL, S., HAWORTH, C. M. and PLOMIN, R. (2008). Evidence for a strong genetic influence on childhood adiposity despite the force of the obesogenic environment. *Am. J. Clin. Nutr.* **87** 398–404.

WEYER, C., FUNAHASHI, T., TANAKA, S., HOTTA, K., MATSUZAWA, Y., PRATLEY, R. E. and TATARANNI, P. A. (2001). Hypoadiponectinemia in obesity and type 2 diabetes: Close association with insulin resistance and hyperinsulinemia. *J. Clin. Endocrinol. Metab.* **86** 1930–1935.

WU, Z., CHEN, H., SUN, F., XU, J., ZHENG, W., LI, P., CHEN, S., SHEN, M., ZHANG, W. et al. (2013). PTPN2 rs1893217 single-nucleotide polymorphism is associated with risk of Behcet's disease in a Chinese Han population. *Clin. Exp. Rheumatol.* **32** S20–S26.

XIA, X., YANG, H. and LI, J. (2016). Feature screening for generalized varying coefficient models with application to dichotomous responses. *Comput. Statist. Data Anal.* **102** 85–97. MR3506984 https://doi.org/10.1016/j.csda.2016.04.008

YEOMANS, M. R. (2010). Alcohol, appetite and energy balance: Is alcohol intake a risk factor for obesity? *Physiol. Behav.* **100** 82–89.

ZHU, L.-P., LI, L., LI, R. and ZHU, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *J. Amer. Statist. Assoc.* **106** 1464–1475. MR2896849 https://doi.org/10.1198/jasa.2011.tm10563