The Annals of Applied Statistics
2016, Vol. 10, No. 2, 596–617
DOI: 10.1214/16-AOAS912

FEATURE SCREENING FOR TIME-VARYING COEFFICIENT MODELS WITH ULTRAHIGH-DIMENSIONAL LONGITUDINAL DATA

BY WANGHUAN CHU[1], RUNZE LI[2] AND MATTHEW REIMHERR

Pennsylvania State University

Motivated by an empirical analysis of the Childhood Asthma Management Project, CAMP, we introduce a new screening procedure for varying coefficient models with ultrahigh-dimensional longitudinal predictor variables. The performance of the proposed procedure is investigated via Monte Carlo simulation. Numerical comparisons indicate that it outperforms existing ones substantially, resulting in significant improvements in explained variability and prediction error. Applying these methods to CAMP, we are able to find a number of potentially important genetic mutations related to lung function, several of which exhibit interesting nonlinear patterns around puberty.

1. Introduction. Over the last several decades we have seen the rapid development of high-dimensional techniques fueled by precipitous advances in technology. As our computing power has increased, so has our ability to obtain and examine ever larger and more complicated data sets. One of the primary examples of such data come from genetic association studies. In traditional genome-wide association studies (GWAS) hundreds of thousands or even millions of single nucleotide polymorphisms (SNPs) are explored to find associations with some phenotypes of interest, for example, blood pressure, height, asthma, etc. Companies are developing cheaper and cheaper sequencing technologies while also providing increasingly larger pictures of an individual's genome. Indeed, the next technological step consists of high throughput sequencing technologies which are capable of complete genome sequencing. Such studies result in millions of genetic mutations which include not only SNPs, but also insertions or deletions of segments of DNA.

The present work was motivated by the Childhood Asthma Management Program (CAMP), a 4 year clinical trial which explored the impact of daily asthma medications on lung development in growing children. We consider 540 subjects, each of whom contributed up to 16 clinical visits. Our aim is to determine which

Received January 2016.

[1]Supported in part by NSF IGERT Award #DGE-1144860, Big Data Social Science.

[2]Supported in part by NIDA Grants P50 DA039838 and P50 DA036107, and NSF Grant DMS-15-12422. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIDA, the NIH or the NSF.

Key words and phrases. Feature selection, time-varying coefficient models, ultrahigh-dimensional longitudinal data, genome-wide association study, functional linear model.

genetic markers, among hundreds of thousands, affect lung development. The specific outcome we focus on here is FEV1, a common proxy for lung development, which represents the volume of air one can expel out of their lungs in one second. Given that the subjects may change rather rapidly over the course of the trial, we also wish to understand how the effect of significant SNPs changes over time. Analyzing such longitudinal genetic data poses a substantial challenge for data scientists and necessitates the development of new data analytic tools to address scientific questions and test important hypotheses.

The so-called "large p, small n" problems require tools that are not only powerful, but also computationally efficient. While association methods such as the LASSO [Tibshirani (1996)] and SCAD [Fan and Li (2001)] are powerful, they require significant computational resources for large-dimensional data sets. One of the main issues stems from having to handle all of the predictors jointly, which is an enormous computational burden when dealing with hundreds of thousands of predictors simultaneously. An elegant and effective solution is to incorporate screening rules. A screening rule is a method which analyzes much smaller subsets of the predictors and attempts to filter out those that are clearly unimportant. Such rules may attempt to pick the "best" subset of predictors or just a substantially smaller subset which could in turn be analyzed by other methods. By using such screening rules, it is not unusual to see full-day computation times reduced to minutes. The primary goal of this work is to develop an effective screening procedure for longitudinal genetic studies such as CAMP.

A number of feature screening procedures have been developed in various contexts. Fan and Lv (2008) developed a sure independence screening procedure (SIS) for ultrahigh-dimensional linear models. Furthermore, they showed that SIS possesses the *sure screening property*, that is, with probability tending to one, it produces a subset which contains the true underlying predictors. Fan and Song (2010) extended SIS to ultrahigh-dimensional generalized linear models by ranking the maximum marginal likelihood estimates. Fan, Feng and Song (2011) proposed an SIS for ultrahigh-dimensional additive models by ranking the magnitude of each nonparametric component. In addition, model-free SIS procedures have been advocated in more recent literature. Zhu et al. (2011) proposed an SIS for the multi-index model setting. Li, Zhong and Zhu (2012) developed a distance-correlation-based SIS, which is directly applicable for a multivariate response and grouped predictors. He, Wang and Hong (2013) proposed a quantile-adaptive model-free feature screening procedure for heterogeneous data. Screening procedures have also been developed for varying coefficient models. Liu, Li and Wu (2014) developed an SIS for varying coefficient models with ultrahigh-dimensional predictor variables (ultrahigh-dimensional varying coefficient models for short) by using a conditional Pearson correlation coefficient to rank the importance of predictors. Fan, Ma and Dai (2014) proposed an SIS for ultrahigh-dimensional varying coefficient models by extending the B-spline techniques in Fan, Feng and Song (2011) for additive models. Song, Yi and Zou (2014) further extended the proposal of

Fan, Ma and Dai (2014) for longitudinal data. Both the work of Liu, Li and Wu (2014) and of Fan, Ma and Dai (2014) were developed based on independent and identically observed data, while the proposal of Song, Yi and Zou (2014) did not incorporate within-subject correlation and dynamic error variance at the screening stage, a key ingredient of our proposed methodology.

While the vast majority of GWAS are cross-sectional, there are numerous longitudinal studies which also have genetic measurements. However, high-dimensional methods for longitudinal outcomes have only been sparsely studied. In a longitudinal genetic study such as CAMP, it is typical that researchers collect many baseline variables, a huge number of genetic markers and longitudinal predictor variables/phenotypic traits. Some baseline variables and longitudinal predictors should be included in the analysis based on prior knowledge. None of the aforementioned works on feature screening for ultrahigh-dimensional varying coefficient models have studied this situation, and this work intends to fill this gap. This work also makes a substantial improvement to the B-spline methods in Fan, Ma and Dai (2014) and Song, Yi and Zou (2014) for ultrahigh-dimensional varying coefficient models by incorporating within-subject correlation and dynamic error structure. This is now straightforward for standard multivariate regression models because it is reasonable to assume that the working models are true or well approximate the truth. However, feature screening procedures focus on cycling through very small submodels, which are inherently misspecified. This poses a substantial challenge for constructing effective screening rules using longitudinal data. The main contributions of this paper are to present an effective screening rule based on B-spline regression and to demonstrate how within-subject variability can be harnessed for increased screening accuracy by Monte Carlo simulation. We support the methodology with accompanying theory and illustrate it via an empirical analysis and comparison of the CAMP data. Our empirical analysis clearly shows that the proposed nonparametric approach is especially useful for such studies with highly nonlinear patterns and intricate within-subject dependencies, as one might expect for rapidly changing populations such as children or the elderly.

The rest of this paper is organized as follows. In Section 2, we propose a new screening rule for longitudinal genetic data. We also discuss how to incorporate within-subject variability and correlation into the screening procedure to increase screening accuracy. A theoretical justification for our methods is provided in Section S1 of the supplementary material [Chu, Li and Reimherr (2016)]. In Section 3, we conduct a Monte Carlo simulation to examine the finite sample performance of the proposed screening procedures, and to compare with existing ones. In Section 4, we present our empirical analysis and comparison of CAMP data using the newly proposed procedure and existing procedures. Some concluding remarks and discussions are given in Section 5.

2. A feature screening procedure. Assume that we collect a random sample from n subjects and, for the ith subject, we observe the response $y_i(t)$ along with

its covariate vectors $\{\mathbf{z}_i(t), \mathbf{x}_i(t)\}$ at times t_{ij}, $j = 1, \ldots, J_i$, where J_i is the total number of observations from the ith subject. While, for complete generality, we always include the argument t, a particular covariate need not change with time, such as gender. The covariate vector $\mathbf{z}_i(t)$ is a low-dimensional predictor consisting of variables that are believed to impact the response based on empirical evidence or relevant theories. Thus, $\mathbf{z}_i(t)$ should be included into the model, and is not subject to be screened. The covariate vector $\mathbf{x}_i(t)$ is ultrahigh-dimensional and contains a vast number of covariates such as hundreds of thousands of SNPs. We assume that the dimension p of $\mathbf{x}_i(t)$ is allowed to grow with sample size n at an exponential rate, that is, $\log p = O(n^a)$ for some $a \in (0, 1/2)$. We will discuss this point further in Section S1 of the supplementary material [Chu, Li and Reimherr (2016)]. It is believed that a relatively small number of x-variables have an impact on the response, and most of the x-variables are likely to be irrelevant. To explore potential time-varying effects, we consider the following time-varying coefficient model:

$$y_i(t) = \beta_0(t) + \sum_{l=1}^{q} \beta_l(t) z_{il}(t) + \sum_{k=1}^{p} \gamma_k(t) x_{ik}(t) + \varepsilon_i(t), \tag{1}$$

where $\{\beta_l(t), l = 0, \ldots, q\}$ and $\{\gamma_k(t), k = 1, \ldots, p\}$ are nonparametric smooth coefficient functions, and $\varepsilon_i(t)$ is the error term with conditional mean $\mathrm{E}\{\varepsilon_i(t)|\mathbf{x}_i(t), \mathbf{z}_i(t)\} = 0$. It is assumed throughout this paper that $\varepsilon_i(t)$ have a variance that varies across time, are independent across i (between subjects) and correlated across t (within same subject). In model (1), t need not to be calendar time. For example, we may set t to be the age of a subject in order to explore potential age-dependent genetic effects and examine whether genetic effect changes across developmental stages. In general, it is assumed that $t \in \mathcal{T}$, where $\mathcal{T}$ is a closed and bounded interval in $\mathbb{R}$.

The goal of a screening procedure is to effectively filter out as many unimportant x-variables as possible while retaining all of the important ones. To denote the significant variables, we define the index set

$$\mathcal{M}_0 = \{1 \le k \le p : \|\gamma_k(\cdot)\|_2 > 0\}, \tag{2}$$

where $\|\cdot\|_2$ is the (functional) L^2 norm. The screening procedure proposed by Liu, Li and Wu (2014), based on conditional correlation, cannot be used for feature screening in model (1) because of the inclusion of z-variables. The screening procedures developed in Fan, Ma and Dai (2014) and Song, Yi and Zou (2014) may be directly applicable for model (1) by assuming that within-subject observations are independent. However, in their original works, a sure screening property has not been established when z-variables are included. In this section, we introduce a more effective screening procedure, which improves the proposal of Fan, Ma and Dai (2014) by including baseline variables, incorporating within-subject correlation and taking into account the time-varying error variance. Furthermore, our

procedure achieves the sure screening property, which is established in Section S1 and proved in Section S2 of the supplementary material [Chu, Li and Reimherr (2016)].

We now describe our procedure. For each k, we define a marginal (a single x-variable) nonparametric regression model with the kth x-predictor:

$$y_i(t_{ij}) = \beta_{0k}^*(t_{ij}) + \sum_{l=1}^{q} \beta_{lk}^*(t_{ij}) z_{il}(t_{lj}) + \gamma_k^*(t_{ij}) x_{ik}(t_{ij}) + \varepsilon_i^*(t_{ij}), \tag{3}$$

where $\{\beta_{lk}^*(t), l = 0, 1, \ldots, q\}$ and $\gamma_k^*(t)$ are smooth coefficient functions. Intuitively, the residual sum of squares of model (3) may be used to measure the importance of the kth x-variable. A smaller residual sum of squares implies that the corresponding x-variable explains more variation of the response variable, and therefore would be more important.

We employ a regression spline method to estimate the coefficient functions and obtain the residuals. Using cubic B-splines, we approximate $\{\beta_{lk}^*(t), l = 0, 1, \ldots, q\}$ and $\gamma_k^*(t)$ as follows:

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

where $\{B_{hm}(\cdot), m = 1, \ldots, M_{hn}\}$ is a set of B-splines which may differ across h, and M_{ln} and L_{kn} are the numbers of basis functions used for $\beta_{lk}^*(t)$ and $\gamma_k^*(t)$ respectively. Larger M_{kn} and L_{ln} lead to more accurate approximations of the varying coefficients, but at the cost of a higher variance (i.e., the classic bias/variance trade-off). Model (3) becomes, approximately, a linear regression model:

$$\begin{aligned} y_i(t_{ij}) \approx{}& \sum_{m=1}^{M_{0n}} \eta_{0m} B_{0m}(t_{ij}) + \sum_{l=1}^{q} \sum_{m=1}^{M_{ln}} \eta_{lm} B_{lm}(t) z_{il}(t_{ij}) \\ &+ \sum_{h=1}^{L_{kn}} \theta_{kh} B_{kh}(t) x_{ik}(t_{ij}) + \varepsilon_i^*(t_{ij}). \end{aligned} \tag{5}$$

The error term $\varepsilon_i^*(t_{ij})$ is assumed to be independent between subjects and correlated within subject. Moreover, the variance of $\varepsilon_i^*(t_{ij})$ is assumed to be time-varying. Incorporating the error covariance structure into the model estimation is expected to increase screening accuracy.

Intuitively, one may use weighted least squares (WLS) or generalized estimating equations (GEE) [Liang and Zeger (1986)] to estimate the coefficients, however, the situation here is much more challenging because (a) the working marginal model (5) is a misspecified model, and (b) the total computational cost for estimating the error variance and correlation matrix in each marginal model would be extremely expensive in the presence of ultrahigh-dimensional x-covariates. Instead

of estimating the covariance matrix of $\boldsymbol{\varepsilon}_i^* = (\varepsilon_i^*(t_{i1}), \ldots, \varepsilon_i^*(t_{iJ_i}))^T$, we propose an approach to construct the weights for WLS using the following working model:

$$y_i(t_{ij}) = \beta_{0k}^w(t_{ij}) + \sum_{l=1}^{q} \beta_{lk}^w(t_{ij}) z_{il}(t_{lj}) + \varepsilon_i^w(t_{ij}). \tag{6}$$

This model includes only the baseline covariates $\mathbf{z}_i(t)$ without any x-covariates. Although it is misspecified, we can still gain valuable information about the response covariance structure that can be incorporated into the screening procedure for better performance.

We construct $V(t_{ij})$, a working variance function for $\varepsilon_i^*(t_{ij})$, using the techniques in Huang, Wu and Zhou (2004). We apply the ordinary least squares method and regression spline technique to model (6), and obtain the corresponding residuals $\{r_i(t_{ij})\}$. Assuming that $V(t)$ is a smooth function of t, we can approximate $V(t_{ij}) \approx \sum_{h=1}^{H_n} \alpha_h B_{hn}(t_{ij})$. Minimizing the following least squares function

$$\sum_{i=1}^{n} \sum_{j=1}^{J_i} \left(r_i^2(t_{ij}) - \sum_{h=1}^{H_n} \alpha_h B_{hn}(t_{ij}) \right)^2 \tag{7}$$

leads to an estimate of the coefficients: $\{\hat{\alpha}_h, h = 1, \ldots, H_n\}$. We then define $\widehat{V}(t_{ij}) = \sum_{h=1}^{H_n} \hat{\alpha}_h B_{hn}(t_{ij})$.

We use a parametric model for the working correlation matrix. Denote by $\mathbf{R}_i(\boldsymbol{\lambda}) = (R_{jk})$ the $J_i \times J_i$ working correlation matrix for the ith subject, where $\boldsymbol{\lambda}$ is an $s \times 1$ vector that fully characterizes the correlation structure. Commonly used correlation structures include autoregressive (AR) correlation structure, stationary or nonstationary M-dependent correlation structures, as well as parametric families such as the Matérn. In practice, we propose to employ moment estimators for the parameters $\boldsymbol{\lambda}$ in the correlation structure based on the residuals $r_i(t_{ij})$s in feature screening procedures. Denote by $\hat{\boldsymbol{\lambda}}$ the resulting moment estimate of $\boldsymbol{\lambda}$.

We propose the following weight matrix for the ith subject:

$$\mathbf{W}_i = J_i^{-1} \widehat{\mathbf{V}}_i^{-1/2} \mathbf{R}_i^{-1}(\hat{\boldsymbol{\lambda}}) \widehat{\mathbf{V}}_i^{-1/2}, \tag{8}$$

where $\widehat{\mathbf{V}}_i$ is the $J_i \times J_i$ diagonal matrix consisting of the time-varying variance

$$\widehat{\mathbf{V}}_i = \begin{pmatrix} \widehat{V}(t_{i1}) & 0 & \ldots & 0 \\ 0 & \widehat{V}(t_{i2}) & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \widehat{V}(t_{iJ_i}) \end{pmatrix}. \tag{9}$$

We can then obtain the WLS estimate for regression coefficients in model (5), and calculate the fitted value $\hat{y}_i^{(k)}(t_{ij})$. More specifically, let $\mathbf{B}(t) = (B_1(t), \ldots, B_{L_n}(t))$, where we ignore the difference among using different numbers of basis

functions M_{l_n} for $l = 0, 1, \ldots, q$ and L_{k_n} for $k = 1, \ldots, p$. Let $\mathbf{B}_z(t)$ be an $(q + 1) \times L_n(q + 1)$ matrix defined by

$$\mathbf{B}_z(t) = \begin{pmatrix} \mathbf{B}(t) & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}(t) & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{B}(t) \end{pmatrix},$$

$\mathbf{U}_{zij}^T = (1, \mathbf{z}_i(t_{ij})^T)\mathbf{B}_z(t_{ij})$, $\mathbf{U}_{zi} = (\mathbf{U}_{zi1}, \ldots, \mathbf{U}_{ziJ_i})^T$, $\mathbf{U}_{xkij}^T = x_{ik}(t_{ij})\mathbf{B}(t_{ij})$, $\mathbf{U}_{xki} = (\mathbf{U}_{xki1}, \ldots, \mathbf{U}_{xkiJ_i})^T$ and $\mathbf{U}_{ki} = (\mathbf{U}_{zi}, \mathbf{U}_{xki})$. For the B-spline coefficients, let $\boldsymbol{\eta}_{kl} = (\eta_{kl1}, \ldots, \eta_{klL_n})^T$ and $\boldsymbol{\eta}_k = (\boldsymbol{\eta}_{k1}, \ldots, \boldsymbol{\eta}_{kq})^T$ for z-variables, $\boldsymbol{\theta}_{xk} = (\theta_{xk1}, \ldots, \theta_{xkL_n})^T$ for the kth x-variable, and $\boldsymbol{\theta}_k = (\boldsymbol{\eta}_k, \boldsymbol{\theta}_{xk})^T$. Then the WLS estimate for $\boldsymbol{\theta}_k$ is

$$\hat{\boldsymbol{\theta}}_k = \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{U}_{ki}^T \mathbf{W}_i \mathbf{U}_{ki}\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{U}_{ki}^T \mathbf{W}_i \mathbf{y}_i\right), \tag{10}$$

and $\mathbf{y}_i$ can be estimated by $\hat{\mathbf{y}}_i^{(k)} = \mathbf{U}_{ki}\hat{\boldsymbol{\theta}}_k$. This enables us to calculate the weighted mean squared errors denoted by $\hat{u}_{nk}$:

$$\hat{u}_{nk} = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{y}_i - \hat{\mathbf{y}}_i^{(k)})^T \mathbf{W}_i (\mathbf{y}_i - \hat{\mathbf{y}}_i^{(k)}). \tag{11}$$

Note that the smaller value of $\hat{u}_{nk}$ indicates stronger marginal association between the kth covariate and the response. Thus, we sort $\{\hat{u}_{nk}, k = 1, \ldots, p\}$ in an increasing order, and define the screened submodel as

$$\widehat{\mathcal{M}}_{\tau_n} = \{1 \le k \le p : \hat{u}_{nk} \text{ ranks among the first } \tau_n\}, \tag{12}$$

where τ_n is the submodel size chosen to be smaller than the sample size n. Following Fan and Lv (2008), we set $\tau_n = [n/\log(n)]$, where $[a]$ refers to the integer part of a. This procedure has sure screening properties, which means that, with probability tending to one, all true covariates are included in the sub-screened model defined by $\widehat{\mathcal{M}}_{\tau_n}$ provided that certain conditions are satisfied. More detailed description of the sure screening property and theoretical proof can be found in Sections S1 and S2 in the supplementary material [Chu, Li and Reimherr (2016)].

3. Simulation studies. To make our simulation results more generalizable to real-world applications, we generate data mimicking the CAMP data, and compare the finite sample performance of the new method with that of sure independence screening (SIS) [Fan and Lv (2008)], nonparametric independence screening (NIS) with varying coefficient models [Fan, Ma and Dai (2014)] and the procedure proposed in Song, Yi and Zou (2014) (NIS2). In our simulation setting where the number of observations for each subject is the same for all subjects, the last two

procedures essentially use the same marginal models, which are time-varying coefficient models assuming working independence and constant variance. Thus, they only differ in the criteria used for ranking the importance of covariates. Song, Yi and Zou's (2014) marginal utility is defined by

$$w_k = \frac{1}{\mathcal{T}} \int_{\mathcal{T}} \hat{\gamma}_k^*(t)^2 \, \mathrm{d}t,$$

and, as suggested in their paper, we take $N = 10{,}000$ equally spaced time points $t_1 < t_2 \cdots < t_N$ on the time interval $\mathcal{T}$ and estimate w_k by

$$w_{Nk} = \frac{1}{N} \sum_{i=1}^{N} \hat{\gamma}_k^*(t_i)^2.$$

For Fan, Ma and Dai (2014), we use residual sums of squares of the marginal model as ranking criteria.

We set the feature dimension, p, to 2000, 5000 and 10,000. We first randomly choose p SNPs from CAMP as the x-variables and set gender as the only z-variable, as only gender among the baseline variables has a significant impact on the response based on our preliminary analysis of the CAMP data using an age-varying coefficient model. The distribution of the age variable is approximately normal over the range [5, 17.2]. To achieve better numerical stability, we make a transformation on the time points $\{\tilde{t}_{ij}, j = 1, \ldots, J_i; i = 1, \ldots, n\}$ so that they are approximately uniformly distributed on [0, 1] by $t_{ij} = \Phi((\tilde{t}_{ij} - \bar{t})/s_t)$, where $\Phi(\cdot)$ is the cumulative distribution function of $N(0, 1)$, and $\bar{t}$ and s_t are the sample mean and standard deviation of all time points $\tilde{t}_{ij}$ in the CAMP data. We generate the simulated data from

$$y_i(t_{ij}) = \beta_0(t_{ij}) + \beta_1(t_{ij})\mathrm{Gender}_i + \sum_{k=1}^{p} \gamma_k(t_{ij})\mathrm{SNP}_{ik} + \varepsilon_i(t_{ij}). \tag{13}$$

In each replication, we randomly select $n = 200$ subjects with $J_i = 16$ observations from the CAMP data, and directly use their values on the gender variable, time t_{ij} and the p selected SNPs in this simulation.

The error term $\varepsilon_i(t_{ij})$ is generated from a zero mean Gaussian process with variance and correlation defined by

$$\begin{aligned} \mathrm{Var}\big(\varepsilon_i(t_{ij})\big) &\equiv V(t_{ij}) = 0.5 + 3t_{ij}^3 \quad \text{and} \\ \mathrm{cor}\big(\varepsilon_i(t_{ij}), \varepsilon_i(t_{ik})\big) &= 0.5\rho_1^{|j-k|} + 0.5\rho_2, \end{aligned} \tag{14}$$

where we use a correlation structure as a combination of AR(1) and compound symmetry with equal weights. We set $(\rho_1, \rho_2) = (0.6, 0.4)$ and $(0.8, 0.6)$ in our simulation. To investigate the sensitivity of using different correlation structures in estimating the weighted matrix, we consider three options in the analysis. The first is stationary M-dependent, where $M = J_i - 1 = 15$ in the simulation. This

is the true structure generated by the second equation in (14), but the parameters are estimated using the working model specified in (6). The other two are AR(1) and compound symmetry, which are commonly used in practice. Their results are labeled as "stat-M," "AR(1)" and "C.P.S.," respectively.

We set x_1, x_2, x_3, x_4 to be significant, and all others are inactive. To ensure fair comparisons, we consider two examples for nonzero coefficients. In the first example, the nonzero coefficients for x-variables are time-varying, while they are time-invariant in the second example. The specific nonzero coefficient functions are given below.

- *Example* I. The nonzero coefficient functions are defined by

$$\gamma_1(t) = 0.5\cos(\pi t)\mathbf{1}_{\{t\le 0.5\}}, \qquad \gamma_2(t) = -0.4\cos(2\pi t)\mathbf{1}_{\{t\le 0.5\}},$$
$$\gamma_3(t) = -0.3\sin(2\pi t), \qquad \gamma_4(t) = 0.5(1.2 - t).$$

- *Example* II. The nonzero coefficient functions are defined by

$$\gamma_1(t) = 0.4, \qquad \gamma_2(t) = 0.5, \qquad \gamma_3(t) = -0.3, \qquad \gamma_4(t) = -0.5.$$

We set $\beta_0(t)$ and $\beta_1(t)$ to be the coefficient functions estimated from $y_i(t_{ij}) = \beta_0(t_{ij}) + \beta_1(t_{ij})\text{Gender}_i + \varepsilon_i(t_{ij})$ using the CAMP data. Their plots are shown in Figure 1. The baseline predictor gender is also considered in the SIS and the NIS method in our numerical comparison.

For NIS [Fan, Ma and Dai (2014)], NIS2 [Song, Yi and Zou (2014)] and our procedure, we apply B-splines approximations to the time-varying coefficients, which involves three tuning parameters: the degree of the splines, the number of knots, and the positions of the interior knots. In all simulation studies, we set the degrees of spline functions to be three, that is, cubic spline, which is the most commonly used option. Since the time points are transformed to be approximately uniform over [0, 1], we use equally spaced knots. The number of interior knots is set to be four, which gives eight degrees of freedom for each varying coefficient. We believe

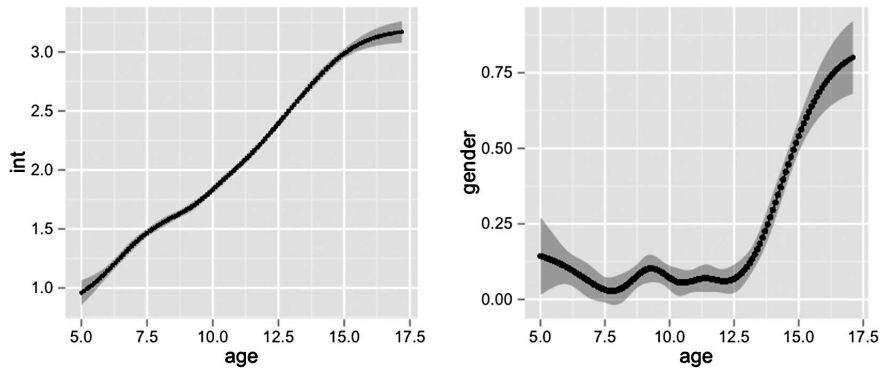

FIG. 1. *Coefficient functions for intercept and gender.*

that this is good enough to capture the time-varying effects. Alternatively, cross-validation methods can be applied to better select the number of interior knots. However, the computational costs are substantial in the presence of ultrahigh dimensionality, and pilot analysis shows that results are not sensitive to different numbers of knots.

Following Liu, Li and Wu (2014), the following four criteria are used to evaluate the performance of different screening methods:

- R_k: The average of ranks of x_k (or SNP_k in our case) 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 from 1000 replications.
- p_a: The proportion of 1000 replications where all true predictors are being selected into $\hat{\mathcal{M}}_{\tau_n}$.
- p_k: The proportion of x_k being selected into the submodel $\hat{\mathcal{M}}_{\tau_n}$ over 1000 replications.

To calculate p_a and p_k, we set the selected submodel size $\tau_n = \nu[n/\log n]$, $\nu = 1, 2, 3$ [Fan and Lv (2008)]. All the simulation results are summarized over 1000 replications.

Results for $p = 2000$ are shown in Tables 1, 2 and 3 for R_j's, quantiles of M, and p_j's and p_a, respectively. Results for $p = 5000$ and 10,000 are shown in

TABLE 1
R_j of the active SNPs for $p = 2000$

	Example 1: $\gamma(t)$'s are time-varying				**Example 2: $\gamma(t)$'s are time-invariant**			
Method	R_1	R_2	R_3	R_4	R_1	R_2	R_3	R_4
				$\rho_1 = 0.6, \rho_2 = 0.4$				
SIS	141.531	1030.235	1021.058	1.499	4.318	3.598	1.132	5.572
NIS	17.835	140.741	94.589	1.387	4.438	3.892	1.132	6.326
NIS2	27.101	147.466	94.407	1.409	4.59	3.663	1.139	6.381
stat-M	4.553	4.351	13.295	1.499	6.892	4.492	1.047	15.004
AR(1)	3.539	19.456	30.644	1.09	4.632	3.88	1.036	6.727
C.P.S.	4.877	3.58	12.391	2.065	6.389	4.35	1.049	14.312
				$\rho_1 = 0.8, \rho_2 = 0.6$				
SIS	255.711	1025.682	1035.404	5.804	4.765	3.937	1.232	6.756
NIS	53.929	234.304	169.961	5.047	5.746	4.569	1.233	9.837
NIS2	73.091	239.601	164.635	4.874	6.281	4.381	1.234	10.242
stat-M	7.913	2.93	12.68	2.726	12.318	6.976	1.093	23.308
AR(1)	5.975	14.892	35.182	1.39	5.275	4.08	1.068	7.985
C.P.S.	8.837	3.225	14.549	4.425	14.877	8.593	1.112	26.956

TABLE 2
The quantiles of M for $p = 2000$

	Example 1: $\gamma(t)$'s are time-varying					Example 2: $\gamma(t)$'s are time-invariant				
Method	**5%**	**25%**	**50%**	**75%**	**95%**	**5%**	**25%**	**50%**	**75%**	**95%**
					$\rho_1 = 0.6, \rho_2 = 0.4$					
SIS	522.9	1059.5	1464.5	1744.5	1956.1	4	5	7.5	8	10
NIS	17	55.75	124	258	610.2	4	5	8	9	15
NIS2	17	57.75	131	276.5	612.1	4	5	8	9	15.05
stat-M	4	5	6	11	58.05	4	7	9	15	62.05
AR(1)	5	8	16	40.25	168.15	4	5	8	9	12
C.P.S.	4	5	6	11	60	4	6	8	13	51.2
					$\rho_1 = 0.8, \rho_2 = 0.6$					
SIS	544.85	1065	1459.5	1755	1949.1	4	5	8	9	18
NIS	46.95	128	251	436	862.05	4	5	8	11	40
NIS2	48.95	130	253	456	858.55	4	5	8	12	42
stat-M	4	5	7	13	67.2	4	8	13	28	118.1
AR(1)	5	8	17	41	190	4	5	8	9	19
C.P.S.	4	5	8	17	88.2	4	8	14	35.25	151.25

Section S3 of the supplementary material [Chu, Li and Reimherr (2016)]. When $p = 2000$, outputs of the first example show that SIS is able to identify SNP_4, with average ranks (R_4) of 1.499 and 5.804, and selection proportions (p_4) 0.997 and 0.971 under $\tau_n = 38$, for the two correlation cases respectively, but fails to select the other three SNPs. This is likely due to $\gamma_4(t)$ being a strong stable signal, with a slope (-0.5) and relatively large intercept (0.6), while the other three coefficients have significant time-varying effects and cannot be detected by SIS. NIS also identifies SNP_4 very well. Furthermore, it selects SNP_1 into the submodel with a relatively large probability, especially under the $(0.6, 0.4)$ correlation scenario and using the more conservative submodel size ($\tau_n = 76$ or 114). However, it tends to give low ranks to SNP_2 and SNP_3 (R_2 and R_3 of NIS from Table 1) and low selection rates (p_2 and p_3 of NIS from Table 3). This is because the signal magnitudes of $\gamma_2(t)$ and $\gamma_3(t)$ are not large enough for NIS to detect. As expected, results of NIS2 are very similar to NIS.

Our procedure has excellent performance for all four SNPs, generating consistently high ranking and large selection rates under all scenarios. It identifies the relatively constant signal (like SIS and NIS), SNP_4, the larger time-varying signal SNP_1 (like NIS, but not SIS), and, unlike the other methods, gains enough power from exploiting the covariance structure to select SNP_2 and SNP_3. Among the three working correlation structures, stationary-M dependent gives the best results as expected (since it is the truth). The AR(1) structure performs better in capturing signals of the first and fourth SNPs, while the compound symmetry structure gives more accurate ranking of the second and third SNPs. This indicates that, under

TABLE 3
Selection proportion p_j's and p_a for true SNPs for $p = 2000$

		Example 1: $\gamma(t)$'s are time-varying					**Example 2: $\gamma(t)$'s are time-invariant**				
τ_n	**Method**	p_1	p_2	p_3	p_4	p_a	p_1	p_2	p_3	p_4	p_a
						$\rho_1 = 0.6, \rho_2 = 0.4$					
38	SIS	0.563	0.013	0.012	0.997	0	1.000	1.000	1.000	0.998	0.998
	NIS	0.889	0.333	0.532	1.000	0.171	0.999	1.000	1.000	0.995	0.994
	NIS2	0.822	0.324	0.541	1.000	0.150	1.000	1.000	1.000	0.993	0.993
	stat-M	0.995	0.994	0.935	1.000	0.924	0.987	0.994	1.000	0.924	0.906
	AR(1)	0.996	0.907	0.819	1.000	0.741	0.999	1.000	1.000	0.994	0.993
	C.P.S.	0.994	0.995	0.947	1.000	0.936	0.991	0.996	1.000	0.936	0.924
76	SIS	0.671	0.025	0.025	1.000	0.002	1.000	1.000	1.000	0.999	0.999
	NIS	0.946	0.519	0.683	1.000	0.356	1.000	1.000	1.000	0.999	0.999
	NIS2	0.91	0.491	0.684	1.000	0.337	1.000	1.000	1.000	0.999	0.999
	stat-M	0.997	0.996	0.971	1.000	0.964	0.995	1.000	1.000	0.965	0.961
	AR(1)	0.999	0.951	0.902	1.000	0.856	1.000	1.000	1.000	0.999	0.999
	C.P.S.	0.999	0.996	0.968	1.000	0.963	0.997	1.000	1.000	0.972	0.970
114	SIS	0.735	0.041	0.037	1.000	0.003	1.000	1.000	1.000	1.000	1.000
	NIS	0.974	0.620	0.768	1.000	0.476	1.000	1.000	1.000	0.999	0.999
	NIS2	0.939	0.608	0.769	1.000	0.449	1.000	1.000	1.000	0.999	0.999
	stat-M	0.997	0.996	0.982	1.000	0.975	0.996	1.000	1.000	0.981	0.977
	AR(1)	1.000	0.969	0.945	1.000	0.915	1.000	1.000	1.000	1.000	1.000
	C.P.S.	0.999	0.996	0.982	1.000	0.977	0.997	1.000	1.000	0.984	0.981
						$\rho_1 = 0.8, \rho_2 = 0.6$					
38	SIS	0.381	0.019	0.013	0.971	0.001	0.997	0.999	1.000	0.988	0.984
	NIS	0.662	0.159	0.318	0.979	0.037	0.992	1.000	1.000	0.956	0.948
	NIS2	0.575	0.164	0.34	0.981	0.029	0.986	0.999	1.000	0.955	0.941
	stat-M	0.974	0.994	0.942	0.996	0.910	0.949	0.976	1.000	0.871	0.806
	AR(1)	0.98	0.920	0.809	0.999	0.732	0.996	1.000	1.000	0.987	0.983
	C.P.S.	0.962	0.991	0.928	0.992	0.878	0.925	0.967	1.000	0.852	0.764
76	SIS	0.5	0.037	0.028	0.981	0.001	1.000	1.000	1.000	0.998	0.998
	NIS	0.806	0.298	0.491	0.993	0.129	0.995	1.000	1.000	0.989	0.984
	NIS2	0.735	0.295	0.508	0.994	0.131	0.994	1.000	1.000	0.986	0.980
	stat-M	0.986	0.999	0.970	1.000	0.957	0.975	0.990	1.000	0.932	0.901
	AR(1)	0.991	0.969	0.893	0.999	0.861	0.997	1.000	1.000	0.993	0.99
	C.P.S.	0.984	0.998	0.958	0.996	0.938	0.959	0.978	1.000	0.925	0.871
114	SIS	0.569	0.05	0.041	0.989	0.003	1.000	1.000	1.000	0.999	0.999
	NIS	0.872	0.416	0.587	0.997	0.216	0.998	1.000	1.000	0.994	0.992
	NIS2	0.819	0.403	0.604	0.997	0.202	0.998	1.000	1.000	0.992	0.990
	stat-M	0.994	0.999	0.983	1.000	0.977	0.984	0.995	1.000	0.967	0.948
	AR(1)	0.996	0.981	0.924	1.000	0.905	0.999	1.000	1.000	0.998	0.997
	C.P.S.	0.988	1.000	0.975	1.000	0.963	0.978	0.993	1.000	0.954	0.929

longitudinal settings, better screening results can be gained from accounting for varying variance and within-subject correlation, even with misspecified correlation structures.

As for the second example, all methods have good performance, with SIS performing the best. Thus, our screening method is also valid for linear models. However, if the underlying model is known to be linear, then SIS would be the best option due to the smallest computational cost. By comparing the results of the two correlation scenarios, we can observe that all methods perform slightly worse when the error correlations get larger.

The simulation results for $p = 5000$ and $10{,}000$ are shown in the supplementary material [Chu, Li and Reimherr (2016)]. As feature dimension p increases to 5000 and $10{,}000$, the aforementioned patterns can still be observed, although the overall performance for all procedures deteriorate with the increase of extra noise. Given the performance of these methods, we can definitively recommend our procedure in practice for longitudinal data. Especially in the setting where further analyses are to be performed, our method truly shines. While our rankings for constant effects are slightly worse, they are still very high, and thus very likely to make it past any reasonable screening threshold. Our performance for truly time-varying effects and dynamic errors is substantially better, and it is clear that SIS and NIS run the risk of missing such signals.

A reasonable concern for our procedure is its computational time. Table 4 shows the average computing time of one replication for all cases. The standard deviations calculated over 1000 replications shown in the parenthesis are quite large because different replications are run at different cluster nodes on HPC at Penn State University. Overall, the new procedures take 4 to 5 times longer than NIS, which is still acceptable when considering the gain in screening accuracy.

4. Application. The Childhood Asthma Management Program (CAMP) was a longitudinal study designed to explore the long-term impact of several daily treatments for mild to moderate asthma in children [The Childhood Asthma Management Program Research Group (1999, 2000)]. Here, we consider $n = 540$ Caucasian subjects, each of whom contributed up to 16 clinical visits over 4 years; 81% made all 16 and 92% missed at most one. The primary outcome variable examined here is lung growth, as assessed by the change in forced expiratory volume in one second (FEV1). There is strong evidence of large within-subject correlations, where the pairwise correlations among all visits range from 0.83 to 0.96. As our procedure incorporates this information when screening, we believe that it can outperform other methods that assume independence when analyzing such highly correlated longitudinal data. Genome-wide SNP data and phenotypic information were downloaded from dbGaP (http://www.ncbi.nlm.nih.gov/gap) study accession phs000166.v2.p1. There are in total eight hundred and seventy thousand SNPs to be screened. We set the age of the ith subject at the jth measurements to be the

TABLE 4
Average computing time of one replication (*standard deviation*)

	Example 1: $\gamma(t)$'s are time-varying		**Example 2: $\gamma(t)$'s are time-invariant**	
(ρ_1, ρ_2)	**(0.6, 0.4)**	**(0.8, 0.6)**	**(0.6, 0.4)**	**(0.8, 0.6)**
		$p = 2000$		
SIS	46.96 (16.19)	93.21 (38.28)	93.21 (38.28)	73.85 (34.49)
NIS	191.51 (66.77)	363.84 (146.24)	363.84 (146.24)	297.01 (141.01)
NIS2	279.77 (100.2)	523.96 (209.59)	523.96 (209.59)	431.3 (207.44)
New method (stat-M)	1000.23 (349.94)	2053.66 (860.9)	2053.66 (860.9)	1569.46 (748.27)
New method [AR(1)]	1006.03 (354.93)	2063.61 (867.21)	2063.61 (867.21)	1581.29 (758.93)
New method (ex)	1020.96 (357.55)	2092.43 (870.79)	2092.43 (870.79)	1606.75 (765.69)
		$p = 5000$		
SIS	61.04 (18.29)	56.05 (16.43)	68.76 (25.8)	61.51 (19.21)
NIS	223.99 (67.93)	200.67 (56.73)	241.61 (90.1)	214.91 (66.45)
NIS2	393.75 (128.53)	359.79 (110.3)	428.7 (167.08)	383.78 (127.43)
New method (stat-M)	1213.21 (365.09)	1097.04 (310.52)	1330.07 (496.37)	1177.26 (361.74)
New method [AR(1)]	1193.5 (360.84)	1076.76 (306.43)	1312.13 (492.97)	1159.08 (358.36)
New method (ex)	1203.5 (363.75)	1087.15 (309.87)	1327.84 (496.8)	1173.61 (362.01)
		$p = 10{,}000$		
SIS	136.27 (52.32)	137.04 (50.51)	185.9 (111.18)	382.63 (112.86)
NIS	584.04 (229.47)	588.2 (221.07)	865.28 (566.91)	1844.63 (562.13)
NIS2	897.5 (350.37)	903.39 (337.47)	1360.19 (912.93)	2951.83 (876.64)
New method (stat-M)	3338.73 (1319.84)	3325.84 (1254.73)	4172.04 (2550.17)	8685.42 (2599.42)
New method [AR(1)]	3311.22 (1306.11)	3298.33 (1242.96)	4174.32 (2569.67)	8714.44 (2607.33)
New method (ex)	3303.42 (1301.17)	3291.55 (1241.23)	4133.03 (2535.05)	8623.91 (2591.1)

time variable t_{ij}, and consider the following model:

$$\begin{aligned}\text{FEV}_i(\text{age}_{ij}) &= \beta_0(\text{age}_{ij}) + \beta_1(\text{age}_{ij})\text{Gender}_i \\ &\quad + \sum_{k=1}^{p} \gamma_k(\text{age}_{ij})\text{SNP}_{ik} + \varepsilon_i(\text{age}_{ij}),\end{aligned} \tag{15}$$

where gender is the baseline predictor and $\{\text{SNP}_{ik}\}$ are the SNP variables. Throughout this empirical analysis, it is assumed that $\varepsilon_i(\text{age}_{ij})$ is a Gaussian process with mean zero and variance $\text{Var}(\varepsilon_i(\text{age}_{ij})) = V(\text{age}_{ij})$, a smoothing function of age.

We apply the feature screening procedure introduced in Section 3 and the NIS method to this data set. We set the tuning parameters the same as in the simulation studies, that is, we use the cubic spline with four equally spaced interior knots. For the error covariance structure, we use the M-dependent correlation with $M = J_i - 1$ to incorporate more flexibility. Both methods select $\tau_n = [540/\log(540)] = 85$ SNPs. The two submodels obtained have 15 overlapping SNPs. Since the purpose of screening procedures is to remove as many irrelevant SNPs as possible and to retain all important SNPs, the screening procedures are typically conservative. Thus, we apply a stepwise regression to remove more irrelevant SNPs.

In the forward step, we choose the SNP which results in the greatest decrease in the weighted residual sum of squares (WRSS), and then use an F-test to determine if this SNP should be added to the model. The F statistic can be calculated by

$$F = \frac{(\text{WRSS}_1 - \text{WRSS}_2)/(\text{df}_2 - \text{df}_1)}{\text{WRSS}_2/\text{df}_2}, \tag{16}$$

where WRSS_1 and WRSS_2 are the weighted residual sum of squares of the model without and with the candidate SNP, respectively, and defined as

$$\text{WRSS} = \sum_{i=1}^{n} (\mathbf{y}_i - \hat{\mathbf{y}}_i)^T \widehat{\boldsymbol{\Sigma}}_i^{-1} (\mathbf{y}_i - \hat{\mathbf{y}}_i), \tag{17}$$

where $\widehat{\boldsymbol{\Sigma}}_i^{-1}$ is the estimated covariance matrix for subject i from its corresponding model. In the backward step, to determine if an existing SNP should be excluded, we check if its contribution is smaller than the newly added SNP. Specifically, let $\{\text{SNP}_{(k)}, k = 1, \ldots, K\}$ be the existing SNPs and $\text{SNP}_{(K+1)}$ be the new one. Then delete $\text{SNP}_{(j)}$ if WRSS of the model based on $\{\text{SNP}_{(k)}, k = 1, \ldots, K\}$ is greater than WRSS of the model based on $\{\text{SNP}_{(k)}, k \in \{1, \ldots, K+1\} \setminus \{j\}\}$. This procedure automatically stops when no SNP can make a significant contribution to the model. By applying this procedure to the two submodels obtained from screening, a final model with 23 SNPs is selected for the new method and a model with 6 SNPs for NIS.

We further compare these two models by conducting leave-one-subject-out cross-validation (LooCV) and assessing their predication performance. At each

TABLE 5
LooCV results

	Number of SNPs	PRESS
New method	23	873.37
NIS	6	992.01

evaluation, we leave the data of one subject out, and predict his/her FEV. Letting $y_i(t_{ij})$ and $\hat{y}_i^{(i)}(t_{ij})$, $j = 1, \ldots, J_i$ be the observed and predicted values for subject i, then we calculate the prediction sum of squares (PRESS):

$$\text{PRESS} = \sum_{i=1}^{n} \sum_{j=1}^{J_i} \big(y_i(t_{ij}) - \hat{y}_i^{(i)}(t_{ij})\big)^2. \tag{18}$$

Table 5 shows the results of the two models selected and the new method outperforms NIS by more than 10%.

We show in Figure 2 the estimated coefficient functions of the best model selected for the new method. The first two panels are the coefficients for the intercept and gender (with female as the baseline); the others are for the 23 SNPs. Detailed information about these 23 SNPs is also shown in Table 6. The shape of the intercept function is as expected; as subjects age, their lungs develop and FEV1 increases. We see that there is a slight tapering around 16–17 years old as teenagers get closer to their adult heights. The shape of the gender function is especially interesting. We see that at younger ages boys have slightly higher (recall female is the baseline) lung function. However, we see a dip and the two groups begin to converge starting around age 10, which is right around the time girls begin entering puberty. Boys, on average, enter puberty about a year after girls, which we can also see as the plot rebounds around age 12 as the boys begin growing larger than the girls. Finally, around age 16 when both groups are closer to their adult heights, we see the plots settle on a more pronounced difference between the genders.

The shapes we see in the SNP functions take a variety of forms. Most are primarily protective (1, 6, 10, 14, 16, 17, 18, 20, 23) or deleterious (2, 3, 4, 5, 7, 9, 11, 13, 15, 19, 21, 22), though SNPs 8 and 12 do not clearly fall into one category. We also see that the impact of many of the SNPs seems to fundamentally change before and after puberty. The plot we see for SNP14 might be what one would expect for a protective SNP: a steady increase which accelerates during puberty and then tapers off. Shapes that are more surprising are ones like SNP1. This SNP starts off as protective, but when children hit puberty, it seems to decrease in effect. SNP3 only seems to be active during puberty, but otherwise does not seem to have an effect. In many of the plots we see more chaotic or rapid behavior around puberty. This makes sense as a rapid growth in the children should rapidly change how SNPs are

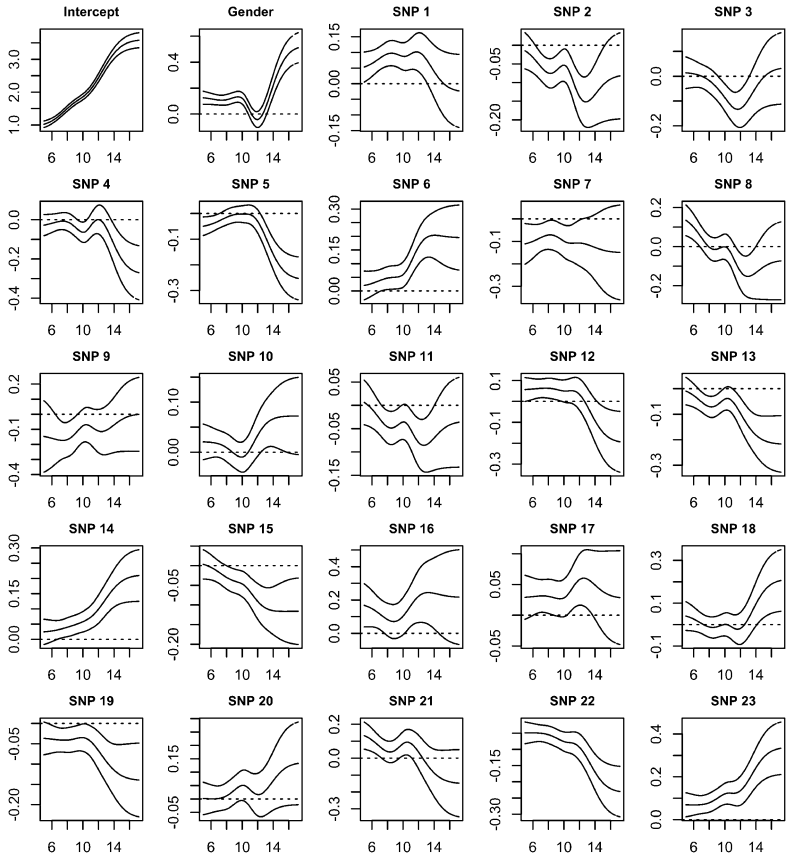

FIG. 2. *Estimated coefficient functions for best model selected by our procedure.*

affecting lung function. What is not so obvious is that puberty also seems to fundamentally change the nature of certain SNPs. Some seem to change the direction of the effect, while some seem most active during puberty. It is these types of patterns which make nonparametric longitudinal methods so powerful. By allowing very general structures for the coefficient functions, we can better find nonlinear patterns.

We conclude this section by examining the heritability discovered by the models, as well as the heritability explained by individual SNPs. Heritability is a concept that summarizes the proportion of variation in a trait due to genetic factors. Examining heritability is an important step in understanding the genetic architecture of complex diseases, and so is commonly measured in GWAS. Statistically, heritably is a type of regression R^2, and thus it is also important for evaluating the fit of a model. Since we are selecting a relatively small subset of SNPs, the heritability we examine here is not the overall heritability of the disease but only the heritability due to our submodel. The heritability of FEV1 was explored in

TABLE 6
Information of the 23 SNPs selected by the new method

No.	Chromosome	SNP name	Chr. position	No.	Chromosome	SNP name	Chr. position
SNP_1	22	rs5992809	16601985	SNP_{13}	2	rs2894456	222765340
SNP_2	16	rs17766975	74968977	SNP_{14}	1	rs1499663	55830578
SNP_3	5	rs4704894	157136938	SNP_{15}	1	rs7530486	64955414
SNP_4	8	rs16924622	60197787	SNP_{16}	5	rs16902245	85806442
SNP_5	10	rs293286	52889045	SNP_{17}	19	rs11673302	9462362
SNP_6	4	rs17444879	41429386	SNP_{18}	11	rs10501066	26724528
SNP_7	15	rs12050625	30751109	SNP_{19}	13	rs12716713	67431310
SNP_8	5	rs17167077	98947056	SNP_{20}	14	rs4904757	41274666
SNP_9	2	rs12469442	195233905	SNP_{21}	4	rs10433674	71980590
SNP_{10}	18	rs1459497	52150550	SNP_{22}	1	rs12734254	77180853
SNP_{11}	2	rs1481387	157598327	SNP_{23}	6	rs7751381	117037951
SNP_{12}	5	rs1013193	169131901				

Reimherr and Nicolae (2014), where they found that the heritability of FEV1 in asthmatic children was around 46%. However, they also discovered that heritability can vary substantially with age. In their methods, "time" was study time (i.e., number of weeks of the trial), where as here we let time be the age of the child. This is especially important as we can get a more direct handle on how heritability changes with age. For model (15), we consider the total heritability of all selected SNPs and the heritability of a single SNP. The heritability of all SNPs is calculated by

$$\mathrm{H(FEV)} = \frac{\mathrm{RSS(FEV|Gender)}}{\mathrm{RSS(FEV|Gender)}} - \frac{\mathrm{RSS(FEV|Gender, SNP}_1, \ldots, \mathrm{SNP}_p)}{\mathrm{RSS(FEV|Gender)}}. \tag{19}$$

Here, RSS is the unweighted residual sum squares defined by

$$\mathrm{RSS} = \sum_{i=1}^{n} \sum_{j=1}^{J_i} \big(y_i(t_{ij}) - \hat{y}_i(t_{ij})\big)^2,$$

where $\hat{y}_i(t_{ij})$ is the fitted value from the model using weighted least square estimation, that is, accounting for time-varying variance and within-subject correlation. The total heritability for our model and the best model of NIS is, respectively, 34.673% and 17.977%. We also estimate the time-varying heritability for all SNPs using a B-splines approximation, and the results are shown in Figure 3. There we see a similar result of Reimherr and Nicolae (2014) that the heritability seems to change quite substantially with age. In particular, we see rapid increases in the

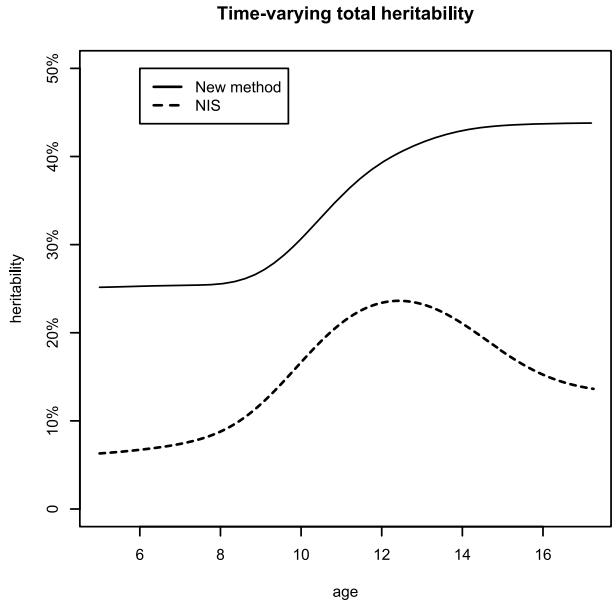

FIG. 3. *Time-varying total heritability.*

heritability as children enter puberty. It seems to level off at around ages 16–17. While we know that the heritability of the NIS submodel is lower than ours, we see another remarkable difference in their time-varying heritability patterns. The NIS model plot looks similar to ours, except at the later ages as it decreases as puberty ends. This suggests that the NIS model has missed SNPs which play a larger role at the later ages.

Finally, we calculate the heritability of single SNPs. This is determined by the order in which each SNP is selected into the model in the stepwise selection procedure. Let $\mathrm{SNP}_{(k)}$ be the kth SNP to be selected into the model, then its heritability is calculated by

$$\mathrm{H}(\mathrm{SNP}_{(k)}) = \frac{\mathrm{RSS}(\mathrm{FEV}|\mathrm{Gender}, \mathrm{SNP}_{(1)}, \ldots, \mathrm{SNP}_{(k-1)})}{\mathrm{RSS}(\mathrm{FEV}|\mathrm{Gender})} - \frac{\mathrm{RSS}(\mathrm{FEV}|\mathrm{Gender}, \mathrm{SNP}_{(1)}, \ldots, \mathrm{SNP}_{(k-1)}, \mathrm{SNP}_{(k)})}{\mathrm{RSS}(\mathrm{FEV}|\mathrm{Gender})}.$$

Tables 7 and 8 show the heritability of a single SNP in the two best models. We see that the heritability of the SNPs ranges fairly evenly between zero and four percent. Interestingly, SNP22 or rs12734254 on gene ST6GALNAC5 was also discovered in Reimherr and Nicolae (2014) using a very different and stringent statistical approach, which reaffirms that this gene is influencing lung function.

TABLE 7
Heritability of single SNPs by new method

Selecting order	SNP name	H(SNP$_{(k)}$)	Selecting order	SNP name	H(SNP$_{(k)}$)
1	rs5992809	1.321%	13	rs2894456	3.196%
2	rs17766975	3.494%	14	rs1499663	1.53%
3	rs4704894	0.68%	15	rs7530486	1.267%
4	rs16924622	1.515%	16	rs16902245	0.825%
5	rs293286	2.412%	17	rs11673302	0.53%
6	rs17444879	4.231%	18	rs10501066	0.198%
7	rs12050625	1.408%	19	rs12716713	1.689%
8	rs17167077	1.327%	20	rs4904757	0.204%
9	rs12469442	0.286%	21	rs10433674	0.388%
10	rs1459497	1.245%	22	rs12734254*	3.454%
11	rs1481387	0.9%	23	rs7751381	2.051%
12	rs1013193	0.524%			

*SNP discovered in Reimherr and Nicolae (2014).

5. Concluding remarks. We developed a screening procedure for ultrahigh-dimensional varying coefficient models motivated by longitudinal genetic studies. From our numerical comparison, the proposed procedure outperforms the SIS proposed in Fan and Lv (2008) and the NIS proposed in Fan, Ma and Dai (2014) for longitudinal data. This implies that incorporating within-subject variability and within-subject correlation is important for increasing the accuracy of a screening rule. We applied the proposed procedure in an analysis of CAMP. The newly proposed screening procedure is able to select a model with much higher heritability and lower prediction error than other methods. Our methodology allows for time-varying SNP effects which revealed that many seem to fundamentally change as children age and enter puberty.

There are a number of ways in which this methodology can be expanded. One that we briefly explored is allowing the correlation structure to also take a smooth nonparametric form. However, our initial attempts showed that the resulting estimates were too noisy to be of much use, and resulted in inconsistent screening results. Thus, finding a nonparametric estimation method for the correlation struc-

TABLE 8
Heritability of single SNPs by NIS

Selecting order	SNP name	H(SNP$_{(k)}$)	Selecting order	SNP name	H(SNP$_{(k)}$)
1	rs1522621	4.201%	4	rs2894456	3.423%
2	rs17766975	3.137%	5	rs4323745	2.698%
3	rs17444879	4.183%	6	rs12734069	0.336%

ture which results in efficient and stable screening would be useful. Another useful generalization would be to allow for more smoothing procedures such as local polynomial smoothing, smoothing splines, etc. Regression splines allow for nice statistical tests, which we exploit in the Application Section. To achieve a similar effect, other smoothing methods would need to be incorporated with care.

SUPPLEMENTARY MATERIAL

Supplement to "Feature screening for time-varying coefficient models with ultrahigh-dimensional longitudinal data" (DOI: 10.1214/16-AOAS912SUPP; .pdf). Theoretical property with technical proofs and additional simulation results for $p = 5000$ and 10,000 are given in the online supplement.

REFERENCES

CHU, W., LI, R. and REIMHERR, M. (2016). Supplement to "Feature screening for time-varying coefficient models with ultrahigh-dimensional longitudinal data." DOI:10.1214/16-AOAS912SUPP.

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

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

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

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

FAN, J. and SONG, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.* **38** 3567–3604. MR2766861

HE, X., WANG, L. and HONG, H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *Ann. Statist.* **41** 342–369. MR3059421

HUANG, J. Z., WU, C. O. and ZHOU, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statist. Sinica* **14** 763–788. MR2087972

LI, R., ZHONG, W. and ZHU, L. (2012). Feature screening via distance correlation learning. *J. Amer. Statist. Assoc.* **107** 1129–1139. MR3010900

LIANG, K. Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22. MR0836430

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

REIMHERR, M. and NICOLAE, D. (2014). A functional data analysis approach for genetic association studies. *Ann. Appl. Stat.* **8** 406–429. MR3191996

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

THE CHILDHOOD ASTHMA MANAGEMENT PROGRAM RESEARCH GROUP (1999). The Childhood Asthma Management Program (CAMP): Design, rationale, and methods. *Control. Clin. Trials* **20** 91–120.

THE CHILDHOOD ASTHMA MANAGEMENT PROGRAM RESEARCH GROUP (2000). Long-term effects of budesonide or nedocromil in children with asthma. *N. Engl. J. Med.* **343** 1054–1063.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

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

W. CHU
M. REIMHERR
DEPARTMENT OF STATISTICS
PENNSYLVANIA STATE UNIVERSITY
STATE COLLEGE, PENNSYLVANIA 16801
USA
E-MAIL: wxc228@psu.edu
mreimherr@psu.edu

R. LI
DEPARTMENT OF STATISTICS
AND THE METHODOLOGY CENTER
PENNSYLVANIA STATE UNIVERSITY
STATE COLLEGE, PENNSYLVANIA 16801
USA
E-MAIL: rzli@psu.edu