Statistica Sinica **30** (2020), 1049-1067
doi:https://doi.org/10.5705/ss.202017.0362

FEATURE SCREENING IN ULTRAHIGH-DIMENSIONAL GENERALIZED VARYING-COEFFICIENT MODELS

Guangren Yang[1], Songshan Yang[2] and Runze Li[2]

[1]*Jinan University and* [2]*Pennsylvania State University*

Abstract: Generalized varying-coefficient models are particularly useful for examining the dynamic effects of covariates on a continuous, binary, or count response. This study examines feature screening for generalized varying-coefficient models with ultrahigh-dimensional covariates. The proposed screening procedure is based on the joint quasi-likelihood of all predictors, which differentiates it from the marginal screening procedures proposed in the literature. In particular, the proposed procedure effectively identifies active predictors that are jointly dependent, but marginally independent of the response. We provide an algorithm for the proposed procedure, and establish the ascent property of the proposed algorithm. Furthermore, we prove that the proposed procedure possesses the sure screening property. That is, with probability tending to one, the selected variable set includes the actual active predictors. We examine the finite-sample performance of the proposed procedure, and compare it with that of several Monte Carlo simulations. Lastly, we illustrate our procedure using a real-data example.

Key words and phrases: Generalized varying-coefficient models, ultrahigh-dimensional data, variable screening.

1. Introduction

Generalized linear models have been well studied in the literature. Penalized likelihood methods have been developed for variable selection in such models with high-dimensional covariates (Tibshirani (1996); Fan and Li (2001)). Ultrahigh-dimensional data are becoming increasingly common in research areas such as genome-wide association studies, proteomics studies, finance, tumor classification, and biomedical imaging. However, variable-selection methods based on penalized likelihood methods may not perform well for ultrahigh-dimensional data because of the methods algorithmic stability, computational cost, and statistical accuracy (Fan, Samworth and Wu (2009)). Fan and Lv (2008) advocate a two-stage approach: (a) reduce the ultrahigh-dimensional covariates to high-dimensional covariates by filtering out those that are irrelevant covariates, using

a marginal screening procedure, and (b) apply variable selection methods to the reduced model. Fan and Lv (2008) proposed a sure independence screening (SIS) procedure for linear models, using the Pearson correlation coefficient as the marginal utility. They also established the sure screening property of their procedure under a Gaussian linear model framework. Hall and Miller (2009) proposed a feature screening procedure for the transformation linear model based on a generalized correlation, and Li et al. (2012) advocated using a rank correlation for screening to deal with a heavy-tailed distribution and the presence of outliers. Fan, Samworth and Wu (2009) proposed an SIS procedure for generalized linear models based on a marginal likelihood estimate. Further details about these procedures can be found in the recent review paper on feature screening by Liu, Zhong and Li (2015).

Varying-coefficient models (VCMs) were proposed to deal with "curse of dimensionality" (Cleveland, Grasse and Shyu (1992); Hastie and Tibshirani (1993)). As a natural extension of linear regression models that allow coefficients to vary over a variable such as age and time, VCMs are particularly useful for exploring dynamic patterns of effects, and have been used in various research fields (e.g., Zhu et al. (2011); Tan et al. (2012); Liu, Li and Wu (2014)). Feature screening procedures for VCMs with ultrahigh-dimensional covariates (referred to as ultrahigh-dimensional VCMs) have been proposed in the literature. Liu, Li and Wu (2014) developed an SIS procedure for ultrahigh-dimensional VCMs that uses conditional Pearson correlation coefficients to denote marginal utility in order to rank the importance of the predictors. Fan, Ma and Dai (2014) proposed an SIS procedure for ultrahigh-dimensional VCMs that extends the B-spline techniques of Fan, Feng and Song (2011) for additive models. Xia, Yang and Li (2016) further extends the SIS procedure proposed in Fan, Ma and Dai (2014) to include generalized varying-coefficient models (GVCMs). Cheng, Honda and Zhang (2016) proposed a forward variable-selection procedure for ultrahigh-dimensional VCMs based on techniques that use B-spline regressions and grouped variable-selection. Song, Yi and Zou (2014) extended the proposal of Fan, Ma and Dai (2014) for longitudinal data, without taking into account within-subject correlation. Then, Chu, Li and Reimherr (2016) proposed an SIS procedure for longitudinal data based on a weighted residual sum of squares that uses within-subjection correlation to improve the accuracy of feature screening. However, while feature screening for ultrahigh-dimensional VCMs is an active research topic in the literature, few studies investigate joint feature screening for ultrahigh-dimensional GVCMs, which is particularly useful for examining the dy-

namic effects of covariates on a binary, count, or continuous response. Exceptions include the work of Li and Zhang (2011), who proposed a new semiparametric threshold model for censored longitudinal data analyses. Then, Cheng et al. (2014) developed a procedure that automatically identifies sparse semivarying coefficient models, which are widely used for longitudinal data analyses. This study intends to fill this gap.

We propose a new feature screening procedure for ultrahigh-dimensional GVCMs. The proposed procedure is based on the joint likelihood of potential active predictors. This differentiates it from existing SIS procedures (Fan, Ma and Dai (2014); Liu, Li and Wu (2014); Xia, Yang and Li (2016)) in that the proposed procedure is not a marginal screening procedure. Wang (2009) proposed a forward regression approach for feature screening in ultrahigh-dimensional linear models, which Cheng, Honda and Zhang (2016) then extended using B-spline regressions and grouped variable-selection. Xu and Chen (2014) proposed a feature screening procedure for generalized linear models based on the sparsity-restricted maximum likelihood estimator. As demonstrated in Wang (2009), Xu and Chen (2014), and Cheng, Honda and Zhang (2016), their approaches outperform the SIS procedures, and effectively identify predictors that are jointly dependent, but marginally independent of the response. We develop a new screening procedure for ultrahigh-dimensional GVCMs based on the joint likelihood of the potential active predictors. The proposed procedure effectively identifies active predictors that are jointly dependent, but marginally independent of the response, without performing an iterative procedure. We develop a computationally efficient algorithm to implement the proposed procedure and establish the ascent property of the proposed algorithm. Furthermore, we prove that this procedure possesses the sure screening property. That is, with probability tending to one, the selected variable set includes the actual active predictors. In summary, this work makes the following major contributions to the literature. (a) We propose a sure joint screening (SJS) procedure for ultrahigh-dimensional GVCMs. In addition, we provide an efficient algorithm to implement the proposed screening procedure, and demonstrate the ascent property of the proposed algorithm. (b) We establish the screening property for the proposed joint screening procedure.

The rest of this paper is organized as follows. In Section 2, we present the proposed feature screening procedure for ultrahigh-dimensional GVCMs, as well as an algorithm for the proposed procedure. Here, we also investigate the theoretical properties of the proposed procedure and algorithm. In Section 3, we present numerical comparisons and an empirical analysis of a real-data example.

Section 4 concludes the paper. All technical proofs are provided in the online Supplementary Material.

2. Screening Procedure for GVCMs

Let Y be the response variable, and let $\{\mathbf{x}, U\}$ denote its associated covariates, where $\mathbf{x} = (X_1, \ldots, X_p)$ and U are p-dimensional and univariate covariates, respectively. Further, let $\mu(\mathbf{x}, U) = E(Y|\mathbf{x}, U)$. The GVCM assumes that

$$\eta(\mathbf{x}, U) \hat{=} g\{\mu(\mathbf{x}, U)\} = \mathbf{x}^T \boldsymbol{\alpha}(U), \tag{2.1}$$

where $g(\cdot)$ is a known link function, and $\boldsymbol{\alpha}(\cdot)$ is a vector consisting of unspecified smooth regression coefficient functions. Here, it is assumed that all $\alpha_j(\cdot)$ are nonparametric functions, and that the support of U is finite and denoted by $[a, b]$.

Suppose that $\{U_i, \mathbf{x}_i, Y_i\}$, for $i = 1, \ldots, n$, constitute an independent and identically distributed (i.i.d.) sample, and that, conditionally on $\{U_i, \mathbf{x}_i\}$, the conditional quasi-likelihood of Y_i is $Q\{\mu(U_i, \mathbf{x}_i), Y_i\}$, where the quasi-likelihood function is defined by $Q(\mu, y) = \int_{\mu}^{y} (s - y)/(V(s))ds$, or equivalently, $(\partial Q(\mu, y))/(\partial \mu) = (y - \mu)/(V(\mu))$, for a specific variance function $V(s)$. Denote by $\ell\{\boldsymbol{\alpha}(\cdot)\}$ the quasi-likelihood (McCullagh and Nelder (1989)) of the collected data $\{(U_i, \mathbf{x}_i, Y_i), i = 1, \ldots, n\}$. That is,

$$\ell\{\boldsymbol{\alpha}(\cdot)\} = \sum_{i=1}^{n} Q[g^{-1}\{\mathbf{x}_i^T \boldsymbol{\alpha}(U_i)\}; Y_i]. \tag{2.2}$$

To estimate the nonparametric regression coefficient, we use the B-spline regression method. Let $\mathcal{S}_n$ be the space of polynomial splines of degree $l \geq 1$, and let $\{\psi_{jk}, k = 1, \ldots, d_{n_j}\}$ denote a normalized B-spline basis with $\|\psi_{jk}\|_\infty \leq 1$ and $d_{nj} = O(n^{1/5})$, where $\|\cdot\|_\infty$ is the sup norm. For any $\alpha_{nj} \in \mathcal{S}_n$, we have

$$\alpha_{nj}(U) = \sum_{k=1}^{d_{n_j}} \beta_{jk} \psi_{jk}(U) = \boldsymbol{\beta}_j^T \boldsymbol{\psi}_j(U), \quad j = 1, \ldots, p, \tag{2.3}$$

for some coefficients $\{\beta_{jk}\}_{k=1}^{d_{n_j}}$. Here, d_{n_j} increases with n. We allow d_{n_j} to vary with j because the coefficient functions may have varying smoothness. Under some conditions, each nonparametric coefficient function $\alpha_j(U)$, for $j = 1, \ldots, p$, can be well approximated by functions in $\mathcal{S}_n$.

Substituting (2.3) into (2.2), the maximum quasi-likelihood estimate of (2.2) maximizes

$$\ell(\boldsymbol{\beta}) \hat{=} \sum_{i=1}^{n} Q\left[g^{-1}\left\{\sum_{j=1}^{p} \boldsymbol{\beta}_j^T \boldsymbol{\psi}_j(U_i) X_{ij}\right\}; Y_i\right] = \sum_{i=1}^{n} Q[g^{-1}(\mathbf{z}_i^T \boldsymbol{\beta}); Y_i] \tag{2.4}$$

with respect to $\boldsymbol{\beta}$, where $\mathbf{z}_i = (X_{i1}\boldsymbol{\psi}_1(U_i)^T, \ldots, X_{ip}\boldsymbol{\psi}_p(U_i)^T)^T$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_p^T)^T$. With a slight abuse of notation, we use $\ell\{\boldsymbol{\alpha}(\cdot)\}$ in (2.2) and $\ell(\boldsymbol{\beta})$ in (2.4). However, the notation will be clear in the relevant context. In the presence of ultrahigh-dimensional covariate $\mathbf{x}$, the corresponding optimization problem becomes ill-posed. It is typical to assume sparsity. That is, only a few x-covariates are significant, with the remainder having no impact on the response. We next propose a feature screening procedure for model (2.1).

2.1. The proposed feature screening procedure

Denote $\|\alpha_j(U)\|_2 = [E\alpha_j^2(U)]^{1/2}$ as the L_2-norm of $\alpha_j(U)$. For ease of presentation, s denotes an arbitrary subset of $\{1, \ldots, p\}$, $\mathbf{x}_s = \{x_j, j \in s\}$, and $\boldsymbol{\alpha}_s(U) = \{\alpha_j(U), j \in s\}$. For a set s, $\tau(s)$ denotes the cardinality of s. Suppose the effect of $\mathbf{x}$ is sparse, and the true value of $\boldsymbol{\alpha}(U)$ is $\boldsymbol{\alpha}^*(U)$; thus, $\boldsymbol{\beta}$ corresponds to $\boldsymbol{\beta}^*$. Denote $s^* = \{j : \|\alpha_j(U)\|_2 > 0\}$. By sparsity, we mean that $\tau(s^*)$ is much less than p. The goal of feature screening is to identify a subset s, such that $s^* \subset s$ with overwhelming probability and $\tau(s)$ is also much less than p. From a theoretical perspective, we can formulate this problem as the following optimization problem:

$$\max_{\boldsymbol{\alpha}(\cdot)} \ell\{\boldsymbol{\alpha}(\cdot)\} \quad \text{subject to } \tau(\{j : \|\alpha_j(\cdot)\|_2^2 > 0\}) \le m, \tag{2.5}$$

for a prespecified m, which is presumed to be much less than p.

When the approximation error is negligible, we construct a feature screening procedure by considering the following maximization problem:

$$\max_{\boldsymbol{\alpha}_n(\cdot)} \ell\{\boldsymbol{\alpha}_n(\cdot)\} \quad \text{subject to } \tau(\{j : \|\alpha_{nj}(\cdot)\|_2^2 > 0\}) \le m. \tag{2.6}$$

Note that $\|\alpha_{nj}(U)\|_2^2 = \boldsymbol{\beta}_j^T E\{\boldsymbol{\psi}_j(U)\boldsymbol{\psi}_j(U)^T\}\boldsymbol{\beta}_j$. Under the assumption that $E\{\boldsymbol{\psi}_j(U)\boldsymbol{\psi}_j(U)^T\}$ is finite positive-definite, for all $j = 1, \ldots, p$, the maximization problem in (2.6) is equivalent to

$$\max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) \quad \text{subject to } \tau(\{j : \|\boldsymbol{\beta}_j\|_2^2 > 0\}) \le m. \tag{2.7}$$

For high-dimensional problems, it becomes almost impossible to solve the constrained maximization problem in (2.7) directly. As an alternative, we consider a proxy for the quasi-likelihood function. It follows from the Taylor expansion for the quasi-likelihood function $\ell(\boldsymbol{\gamma})$ at $\boldsymbol{\beta}$, within the neighborhood of $\boldsymbol{\gamma}$,

that

$$\ell(\boldsymbol{\gamma}) \approx \ell(\boldsymbol{\beta}) + (\boldsymbol{\gamma}-\boldsymbol{\beta})^T \ell'(\boldsymbol{\beta}) + \frac{1}{2}(\boldsymbol{\gamma}-\boldsymbol{\beta})^T \ell''(\boldsymbol{\beta})(\boldsymbol{\gamma}-\boldsymbol{\beta}),$$

where $\ell'(\boldsymbol{\beta}) = \partial \ell(\boldsymbol{\gamma})/\partial \boldsymbol{\gamma}|_{\boldsymbol{\gamma}=\boldsymbol{\beta}}$ and $\ell''(\boldsymbol{\beta}) = \partial^2 \ell(\boldsymbol{\gamma})/\partial \boldsymbol{\gamma}\partial \boldsymbol{\gamma}^T|_{\boldsymbol{\gamma}=\boldsymbol{\beta}}$. Denote $P_t = \sum_{j=1}^{p} d_{nj}$. If $\ell''(\boldsymbol{\beta})$ is invertible, the computational complexity of calculating the inverse of $\ell''(\boldsymbol{\beta})$ is $O(P_t^3)$. For problems with large P_t and small n (i.e., $P_t \gg n$), $\ell''(\boldsymbol{\beta})$ becomes not invertible. A low computational cost is always desirable for feature screening. To cope with the singularity of the Hessian matrix and to minimize the computational cost, we propose using the following approximation for $\ell''(\boldsymbol{\gamma})$:

$$h(\boldsymbol{\gamma}|\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) + (\boldsymbol{\gamma}-\boldsymbol{\beta})^T \ell'(\boldsymbol{\beta}) - \frac{u}{2}(\boldsymbol{\gamma}-\boldsymbol{\beta})^T W(\boldsymbol{\beta})(\boldsymbol{\gamma}-\boldsymbol{\beta}), \tag{2.8}$$

where u is a scaling constant (to be specified), and $W(\boldsymbol{\beta}) = \operatorname{diag}(W_1(\boldsymbol{\beta}), \ldots, W_p(\boldsymbol{\beta}))$ is a block diagonal matrix, with $W_j(\boldsymbol{\beta})$ a $d_{nj} \times d_{nj}$ matrix. Here, we allow $W(\boldsymbol{\beta})$ to depend on $\boldsymbol{\beta}$. This implies that we approximate $\ell''(\boldsymbol{\beta})$ by $-uW(\boldsymbol{\beta})$. Throughout this paper, we use $W_j(\boldsymbol{\beta}) = -\partial^2 \ell(\boldsymbol{\beta})/\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_j^T$.

Clearly, $h(\boldsymbol{\beta}|\boldsymbol{\beta}) = \ell(\boldsymbol{\beta})$. Furthermore, under some conditions, $h(\boldsymbol{\gamma}|\boldsymbol{\beta}) \le \ell(\boldsymbol{\beta})$, for all $\boldsymbol{\gamma}$. This ensures the ascent property. See Theorem 1 below for more details. Because $W(\boldsymbol{\beta})$ is a block diagonal matrix, $h(\boldsymbol{\gamma}|\boldsymbol{\beta})$ is an additive function of γ_j, for any given $\boldsymbol{\beta}$. This additivity enables us to have a closed-form solution for the following maximization problem:

$$\max_{\boldsymbol{\gamma}} h(\boldsymbol{\gamma}|\boldsymbol{\beta}) \qquad \text{subject to } \tau(\{j : \|\boldsymbol{\gamma}_j\|_2^2 > 0\}) \le m, \tag{2.9}$$

for given $\boldsymbol{\beta}$ and m. Define $\tilde{\boldsymbol{\gamma}}_j = \boldsymbol{\beta}_j + u^{-1} W_j^{-1}(\boldsymbol{\beta}_j) \partial \ell(\boldsymbol{\beta})/\partial \boldsymbol{\beta}_j$, for $j = 1, \ldots, p$, and $\tilde{\boldsymbol{\gamma}} = (\tilde{\boldsymbol{\gamma}}_1^T, \ldots, \tilde{\boldsymbol{\gamma}}_p^T)^T = \boldsymbol{\beta} + u^{-1} W^{-1}(\boldsymbol{\beta}) \ell'(\boldsymbol{\beta})$ is the maximizer of $h(\boldsymbol{\gamma}|\boldsymbol{\beta})$. Denote $g_j = \tilde{\boldsymbol{\gamma}}_j^T W_j(\boldsymbol{\beta}_j) \tilde{\boldsymbol{\gamma}}_j$, for $j = 1, \ldots, p$, and sort g_j such that $g_{(1)} \ge g_{(2)} \ge \cdots \ge g_{(p)}$. The solution to the maximization problem given in (2.9) is the hard-thresholding rule defined as follows:

$$\hat{\boldsymbol{\gamma}}_j = \tilde{\boldsymbol{\gamma}}_j I\{g_j > g_{(m+1)}\}.$$

This enables us to screen features effectively using the following algorithm:

Step 1. Set the initial value $\boldsymbol{\beta}_j^{(0)} = \mathbf{0}$, for $j = 1, \ldots, p$.

Step 2. Set $t = 0, 1, 2, \ldots$, and iteratively perform Step 2a and Step 2b until the algorithm converges.

Step 2a. Calculate $\tilde{\boldsymbol{\gamma}}_j^{(t)} = \boldsymbol{\beta}_j^{(t)} + u_t^{-1} W_j^{-1}(\boldsymbol{\beta}_j) \partial \ell(\boldsymbol{\beta}^{(t)})/\partial \boldsymbol{\beta}_j$, and $g_j^{(t)} = \{\tilde{\boldsymbol{\gamma}}_j^{(t)}\}^T W_j(\boldsymbol{\beta}^{(t)}) \tilde{\boldsymbol{\gamma}}_j^{(t)}$. Let $g_{(1)}^{(t)} \ge g_{(2)}^{(t)} \ge \cdots \ge g_{(p)}^{(t)}$, the order statistics of $g_j^{(t)}$s. Set $S_t = \{j : g_j^{(t)} \ge g_{(m+1)}^{(t)}\}$, the nonzero index set.

Step 2b. Update $\boldsymbol{\beta}$ by $\boldsymbol{\beta}^{(t+1)} = (\boldsymbol{\beta}_1^{(t+1)}, \ldots, \boldsymbol{\beta}_p^{(t+1)})^T$, as follows. If $j \notin S_t$, set $\boldsymbol{\beta}_j^{(t+1)} = \mathbf{0}$; otherwise, set $\{\boldsymbol{\beta}_j^{(t+1)} : j \in S_t\}$ as the maximum likelihood estimate of the submodel S_t.

Remark 1. Unlike the screening procedures based on marginal partial likelihood methods, our proposed procedure iteratively updates $\boldsymbol{\beta}$ in Step 2. This enables the proposed screening procedure to incorporate information on the correlations between the predictors by updating $\ell_p'(\boldsymbol{\beta})$ and $\ell_p''(\boldsymbol{\beta})$. Thus, the proposed procedure is expected to outperform the marginal screening procedures when some predictors are marginally independent. At the same time, because Step 2 does not include a large-scale matrix inversion, it incurs a low computational cost.

Theorem 1. *Let $\{\boldsymbol{\beta}^{(t)}\}$ be the sequence defined in Step 2b of the above algorithm. Denote*

$$\rho^{(t)} = \sup_{\boldsymbol{\beta}} \Big[\lambda_{\max}\{W^{-1/2}(\boldsymbol{\beta}^{(t)})\{-\ell''(\boldsymbol{\beta})\}W^{-1/2}(\boldsymbol{\beta}^{(t)})\}\Big].$$

Here, and hereafter, $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote the maximal and minimal eigenvalues of a matrix A, respectively. If $u_t \geq \rho^{(t)}$, then

$$\ell\left(\boldsymbol{\beta}^{(t+1)}\right) \geq \ell\left(\boldsymbol{\beta}^{(t)}\right),$$

where $\boldsymbol{\beta}^{(t+1)}$ is defined in Step 2b of the above algorithm.

Theorem 1 claims the ascent property of the proposed algorithm if u_t is chosen appropriately. That is, the proposed algorithm may improve the current estimate within the feasible region (i.e., $\tau(\{j : \|\alpha_j(U)\|_2 > 0\}) \leq m$), and the resulting estimate in the current step may serve as a refinement of the previous step. This theorem also provides insight into choosing u_t in a practical implementation. For VCMs: $E(Y|U, \mathbf{x}) = \mathbf{x}^T\boldsymbol{\alpha}(U)$, and we may set $\ell\{\boldsymbol{\alpha}(\cdot)\} = -2^{-1}\sum_{i=1}^n\{Y_i - \mathbf{x}_i\boldsymbol{\alpha}(U_i)\}^2$. In this case, $\ell(\boldsymbol{\beta})$ in (2.4) is $\ell(\boldsymbol{\beta}) = -2^{-1}\sum_{i=1}^n(Y_i - \mathbf{z}_i^T\boldsymbol{\beta})^2$. Thus, $-\ell''(\boldsymbol{\beta}) = \sum_{i=1}^n \mathbf{z}_i\mathbf{z}_i^T = \mathbf{Z}^T\mathbf{Z}$, where $\mathbf{Z}$ is $n \times p_t$ matrix, with the ith row being $\mathbf{z}_i^T$. Thus,

$$\rho^{(t)} = \lambda_{\max}(\mathrm{diag}(\mathbf{Z}^T\mathbf{Z})^{-1/2}(\mathbf{Z}^T\mathbf{Z})\mathrm{diag}(\mathbf{Z}^T\mathbf{Z})^{-1/2}),$$

which does not depend on iteration t. If $\mathbf{z}_i$ is marginally standardized such that its marginal sample mean and sample standard deviation are equal to zero and one, respectively, then $\mathrm{diag}(\mathbf{Z}^T\mathbf{Z})^{-1/2}(\mathbf{Z}^T\mathbf{Z})\mathrm{diag}(\mathbf{Z}^T\mathbf{Z})^{-1/2}$ is the corresponding sample correlation matrix of $\mathbf{z}_i$. Thus, ρ is the largest eigenvalue of the sample correlation matrix.

2.2. Sure screening property

For a subset s of $\{1,\ldots,p\}$ with size $\tau(s)$, recall that $\mathbf{x}_s = \{x_j, j \in s\}$ and its associated coefficients $\boldsymbol{\alpha}_s(U) = \{\alpha_j(U), j \in s\}$ correspond to $\boldsymbol{\beta}_s = \{\boldsymbol{\beta}_j, j \in s\}$, with $\boldsymbol{\beta}_j = (\beta_{j1},\ldots,\beta_{jd_{nj}})$. We denote the true model by $s^* = \{j : E\alpha_j^2(U) > 0, 1 \le j \le p\}$, with $\tau(s^*) = q$. The objective of feature selection is to obtain a subset $\hat{s}$, such that $s^* \subset \hat{s}$ with very high probability.

We now provide theoretical justifications for the screening procedure for the GVCM. The sure screening property (Fan and Lv (2008)) is defined as

$$Pr(s^* \subset \hat{s}) \longrightarrow 1, \quad as \quad n \to \infty. \tag{2.10}$$

To establish this property for the proposed feature screening method, we introduce the following additional notation. For any model s, let $\ell'(\boldsymbol{\beta}_s) = \partial\ell(\boldsymbol{\beta}_s)/\partial\boldsymbol{\beta}_s$ and $\ell''(\boldsymbol{\beta}_s) = \partial^2\ell(\boldsymbol{\beta}_s)/\partial\boldsymbol{\beta}_s\partial\boldsymbol{\beta}_s^T$ be the score function and the Hessian matrix of $\ell(\cdot)$ as a function of $\boldsymbol{\beta}_s$, respectively. Assume that a screening procedure retains m out of p features, such that $\tau(s^*) = q < m$. Therefore, we define

$$S_+^m = \{s : s^* \subset s; \|s\|_0 \le m\} \quad \text{and} \quad S_-^m = \{s : s^* \not\subset s; \|s\|_0 \le m\} \tag{2.11}$$

as collections of over-fitted and under-fitted models, respectively. We investigate the asymptotic properties of $\hat{\boldsymbol{\beta}}_m$ when p, q, m, and $\boldsymbol{\beta}^*$ are allowed to depend on the sample size n. We impose the following conditions, some of which are purely technical and serve only to facilitate a theoretical understanding of the proposed procedure.

(C1) The support of U is bounded and is assumed to be $[a, b]$.

(C2) The functions $\{\alpha_j(U)\}_{j=1}^p$ belong to a class of functions $\mathcal{F}$, whose rth derivative $\alpha_j^{(r)}$ exists and is Lipschitz of order η,

$$\mathcal{F} = \left\{\alpha_j(\cdot) : |\alpha_j^{(r)}(s) - \alpha_j^{(r)}(t)| \le K|s-t|^\eta \text{ for } s,t \in [a,b]\right\},$$

for some positive constant K, where r is a nonnegative integer and $\eta \in (0,1]$, such that $\upsilon = r + \eta > 0.5$.

(C3) There exist $w_1, w_2 > 0$ and nonnegative constants τ_1 and τ_2, such that $\tau_1 + \tau_2 < 1/2$, with

$$\min_{j\in s^*} \|\alpha_j(U)\|_2 \ge w_1 n^{-\tau_1} \quad \text{and} \quad q < m \le w_2 n^{\tau_2}.$$

(C4) $\log p = O(n^\kappa)$, for some $0 \le \kappa < 1 - 2(\tau_1 + \tau_2)$.

(C5) $\mu'(\cdot)/V(\cdot)$ is bounded by some constant $M > 0$.

(C6) There exist constants $C_1, C_2 > 0$, $\delta > 0$, such that, for sufficiently large n,

$$C_1 d_n^{-1} \le \lambda_{\min}[-n^{-1}\ell''(\boldsymbol{\beta}_s)] \le \lambda_{\max}[-n^{-1}\ell''(\boldsymbol{\beta}_s)] \le C_2 d_n^{-1},$$

for $\boldsymbol{\beta}_s \in \{\boldsymbol{\beta} : \|\boldsymbol{\beta}_s - \boldsymbol{\beta}_s^*\|_2 \le \delta\}$ and $s \in S_+^{2m}$, where $\lambda_{\min}[\cdot]$ and $\lambda_{\max}[\cdot]$ denote the smallest and largest eigenvalues of a matrix, respectively.

Under Conditions (C1) and (C2), the following two properties of B-splines are valid.

(a) (de Boor (1978)) For $k = 1, \ldots, d_n$, $\psi_{jk}(U) \ge 0$ and $\sum_{k=1}^{d_n} \psi_{jk}(U) = 1$, $U \in [a, b]$. In addition, there exist positive constants C_3 and C_4, such that $C_3 d_n^{-1} \le E\psi_{jk}^2(U) \le C_4 d_n^{-1}$.

(b) (Stone (1982, 1985)) If $\{\alpha_j, j = 1, 2, \ldots, p\}$ is a set of functions in $\mathcal{F}$ described in condition (C2), there exists a positive constant C_5 that does not depend on $\alpha_j(U)$, such that the uniform approximation error has the following bound: $\rho = \sup_{U\in[a,b]} \|\alpha_j(U) - \alpha_{nj}(U)\|_2 \le C_5 d_n^{-\upsilon}, \forall j$, as $d_n \to \infty$.

Conditions (C1) and (C2) ensure properties (a) and (b), which are required for the B-spline approximation and establishing the sure screening properties.

Note that $\|\alpha_{nj}(U)\|_2^2 = \boldsymbol{\beta}_j^T E\{\boldsymbol{\psi}_j(U)\boldsymbol{\psi}_j(U)^T\}\boldsymbol{\beta}_j$. Based on properties (a) and (b) and Condition (C3), we can derive that

$$\min_{j\in s^*} \|\boldsymbol{\beta}_j\|_2 \ge w_1 d_n n^{-\tau_1}. \tag{2.12}$$

Condition (C3) states a few requirements for establishing the sure screening property of the proposed procedure. The first is the sparsity of $\boldsymbol{\beta}^*$, which makes the sure screening possible with $\tau(\hat{s}) = m > q$. Condition (C3) requires that the signal of the active components ($\|\alpha_j(U)\|_2, j \in s^*$) does not vanish. This is referred to as the minimal signal condition in the literature. A minimal signal condition is a commonly imposed assumption in existing works on marginal feature screening for other models (e.g., Liu, Li and Wu (2014)). From (2.12), the condition is equivalent to requiring that the minimal component in $\boldsymbol{\beta}^*$ does not degenerate too fast, so that the signal is detectable in the asymptotic sequence. Condition (C4) has p diverge with n at up to an exponential rate. At the same time, together with (C6), it confines an appropriate order of m that guarantees the identifiability of s^* over s, for $\tau(s) \le m$. For the VCM discussed in Section 2.1, Condition (C6) requires

$$C_1 d_n^{-1} \le \lambda_{\min}[n^{-1}\mathbf{Z}_s^T\mathbf{Z}_s] \le \lambda_{\max}[n^{-1}\mathbf{Z}_s^T\mathbf{Z}_s] \le C_2 d_n^{-1},$$

where $\mathbf{Z}_s$ is the corresponding design matrix of model s. We establish the sure screening property of the quasi-likelihood estimation by the following theorem.

In Fan and Song (2010), Condition D ensures the tail of the response variable Y is exponentially light, as shown in the following Lemma 1. Furthermore, Condition D corresponds to our Condition (C6); thus Condition (C6) ensures that Y is bound.

Remark 2. In particular, our proposed screening procedure is based on the joint quasi-likelihood of all predictors. However, Fan, Ma and Dai (2014) investigate marginal nonparametric methods for screening variables in sparse ultrahigh-dimensional VCMs. Then, in Fan, Ma and Dai (2014), conditions (v) and (vi) are requirements related to the tail distribution of each covariate and the noise, respectively, which are used to establish the sure screening property. However, errors need to be independent, but not normally distributed. Corresponding to our condition (C6), we need only assume that the minimum and maximum eigenvalues of the Hessian matrix are bounded.

Theorem 2. *Suppose we have n independent observations, with p candidate features, from model* (2.1)*, and that conditions* (C1)-(C7) *are satisfied. Let $\hat{s}$ be the features obtained by* (2.5) *of size m. Then, we have*

$$Pr(s^* \subset \hat{s}) \to 1, \ \text{as } n \to \infty.$$

The proof is given in the online Supplementary Material. The sure screening property is an appealing property of a screening procedure because it ensures that the true active predictors are retained in the model selected by the procedure. We establish the sure screening property under weaker conditions than those imposed in Fan, Ma and Dai (2014) and Xia, Yang and Li (2016).

One has to specify the value of m in a practical implementation. Here, there are two scenarios. In the first, we choose m using the data-driven method described in Section 2.3. The second is an ad hoc method. In the literature on feature screening, it is typical to set $m = [n/\log(n)]$ for a parametric model, where $[a]$ indicates the integer part of a (Fan and Lv (2008)). Because we use a linear combination of d_n B-spline bases in our proposed screening procedure for the GVCM, we set $m = [(n/d_n)/\log(n/d_n)]$ throughout in Examples 1, 2, and 3. Despite being an ad hoc choice, it works reasonably well in our numerical examples. Given this choice of m, we are ready to apply existing methods, such as the penalized quasi-likelihood method, to further remove inactive predictors. Note that to distinguish it from the SIS procedure, we refer to the proposed procedure as sure joint screening (SJS) procedure.

2.3. Choice of m

Feature screening may be used in various contexts. In some, we may treat m as a prespecified value. For example, owing to a budget constraint, a biologist can examine up to m genes that potentially associate with a certain phenotype. In other contexts, we might treat m as a tuning parameter to control model complexity. In such cases, it is desirable to develop an automatic data-driven method to determine m. We propose selecting m by minimizing the following high-dimensional BIC score:

$$HBIC(m) = -2\ell(\hat{\boldsymbol{\beta}}_m) + d_n m \frac{C_n \log(d_n p)}{n},$$

where $\hat{\boldsymbol{\beta}}_j = (\hat{\beta}_{j1}, \ldots, \hat{\beta}_{jd_n})$, for $j = 1, \ldots, m$, and C_n is a sequence of numbers that diverges to ∞. Wang, Kim and Li (2013) proposed the HBIC for selecting the tuning parameter in the penalized least squares method for high-dimensional linear models. Here, we modify their proposal for high-dimensional GVCMs. In our simulation, we take $C_n = \log(\log n)$, and compare its performance with that of the AIC and BIC tuning parameter selectors, defined in the same manner. Note that the proposed HBIC selector for the tuning parameter requires searching over $m = 1, 2, \ldots, [n/d_n]$. In contrast, the classical AIC and BIC used for subset selection require searching over subsets. Thus, the tuning parameter selector does not incur a high computational cost.

Recall the notation S_+^m and S_-^m defined in (2.11). Theorem 3 shows that the HBIC selects the right model size, almost surely.

Theorem 3. *Suppose we have n independent observations with p candidate features from model* (2.1)*, and that conditions* (C3)-(C6) *are satisfied. Let $\hat{s}$ be the features obtained by* (2.4) *and* (2.7) *of size m. Then, we have*

$$Pr\left\{\min_{s \in S_-^m} HBIC(\tau(s)) \le HBIC(q)\right\} \longrightarrow 0, \tag{2.13}$$

where $q = \tau(s^*)$, and

$$Pr\left\{\min_{s \in S_+^m, s \neq s^*} HBIC(\tau(s)) \le HBIC(q)\right\} \longrightarrow 0. \tag{2.14}$$

In Example 4, we examine the performance of the proposed HBIC tuning parameter selector.

3. Numerical Studies

In this section, we conduct numerical studies to examine the finite-sample performance of the proposed procedure, which we then compare with that of existing procedures. All simulations are conducted using R code. Examples 1, 2, and 3 examine the performance of the proposed screening procedures. Following the literature on feature screening (e.g., Fan and Lv (2008)), we set $m = [n/\log(n)]$ in these examples. Example 4 examines the performance of the proposed HBIC, and m is determined by minimizing the HBIC score.

3.1. Simulation studies

In our simulation, the covariates u and $\mathbf{x}$ are generated as follows. First, draw $(U^*, \mathbf{x})^T$ from a $p+1$-dimensional normal distribution $N(0, \Sigma)$. Then, set $U = \Phi(U^*)$, where $\Phi(\cdot)$ is the cumulative distribution function of $N(0,1)$. Thus, U follows a uniform distribution $U(0,1)$ and is correlated with $\mathbf{x}$, and the predictors $X_1, \ldots, X_p$ are all correlated with each other. In our simulation, we consider two scenarios for $\Sigma = (\sigma_{ij})$:

Σ_1: A compound symmetric correlation structure: $\sigma_{ij} = 1$ if $i = j$, and ρ otherwise.

Σ_2: An AR(1) correlation structure: $\sigma_{ij} = \rho^{|i-j|}$.

In our numerical studies, we set the number of B-spline basis functions as $d_{n_j} = 5$, for $j = 1, \ldots, p$, for each coefficient function. We use the following two criteria to assess the performance of the proposed procedure:

P_a: The proportion of submodels $\hat{\mathcal{M}}$ with size d that contain all true predictors among 1,000 simulations.

P_j: The proportion of submodels $\hat{\mathcal{M}}$ with size d that contain X_j among 1,000 simulations.

Example 1. This example compares the proposed screening procedure to existing SIS procedures for VCMs. The proposal of Xia, Yang and Li (2016) under the setting of a VCM coincides with that in Fan, Ma and Dai (2014), which follows the spirit of Liu, Li and Wu (2014). Furthermore, Song, Yi and Zou (2014) and Chu, Li and Reimherr (2016) proposed methods for longitudinal data. Therefore, we concentrate on our comparison with CC-SIS, as proposed by Liu, Li and Wu (2014). Given $\{U, \mathbf{x}\}$, we generate a continuous response from

$$Y = \alpha_1(U)X_1 + \alpha_2(U)X_2 + \alpha_3(U)X_3 + \alpha_4(U)X_4 + \boldsymbol{\varepsilon}, \tag{3.1}$$

where $\varepsilon \sim N(0,1)$. Model (3.1) implies that $\alpha_j(\cdot) = 0$ for $j > 4$ and $\mathcal{M}_* = \{1,2,3,4\}$. We consider two sets of coefficient functions:

$\boldsymbol{\alpha}_1$: Let $\alpha_1(u) = \alpha_2(u) = \alpha_3(u) = 2 + 2\sin^2(2\pi u)$ and $\alpha_4(u) = -3\rho * \alpha_1(u)$.

$\boldsymbol{\alpha}_2$: $\alpha_1(u) = -(3 + 2\cos^2(\frac{\pi}{2}u))$, $\alpha_2(u) = -(3+3u)$, $\alpha_3(u) = (2-u)^2 + 2$, $\alpha_4(u) = 3 + 2\sin^2(\frac{\pi}{2}u)$.

In this example, we consider $p =$ 1,000 and 2,000, with the sample sizes $n = 200$ and 400. All simulation results are based on 1,000 replications. The simulation results are summarized in Tables 1-3.

Table 1 shows the values of $\mathcal{P}_1, \ldots, \mathcal{P}_4$, and $\mathcal{P}_a$ for a continuous response, with $\Sigma = \Sigma_1$. Under the design of $\boldsymbol{\alpha}_1$, X_4 is jointly dependent, but marginally independent of Y. In this setting, the marginal screening procedure fails to identify X_4. As shown in Table 1, when there exists marginal independence, CC-SIS is unable to detect X_4, which has values of $\mathcal{P}_4$ and $\mathcal{P}_a$ near zero, as expected. However, our method does identify X_4 in this setting, and the corresponding values of $\mathcal{P}_4$ and $\mathcal{P}_a$ are close to one. Therefore, our procedure outperforms CC-SIS in the presence of marginal independence. Under the design of $\boldsymbol{\alpha}_2$, there is no predictor that is jointly dependent, but marginally independent of Y. Both CC-SIS and the proposed procedure perform very well, with detection probabilities close to one. CC-SIS performs better when the sample size increases and the dimensionality decreases. However, these factors have less of an effect on the new procedure than they do on CC-SIS. Furthermore, the corresponding values of $\mathcal{P}_j$s and $\mathcal{P}_a$ of our new procedure are closer to one in every case. In summary, when $\Sigma = \Sigma_1$, regardless of whether marginal independence exists, the proposed procedure outperforms CC-SIS.

Table 2 shows the values of $\mathcal{P}_j$s and $\mathcal{P}_a$ for a continuous response, with $\Sigma = \Sigma_2$. There is no predictor that is jointly dependent, but marginally independent of Y. Hence, both procedures perform well, with most values of $\mathcal{P}_a$ greater than 0.9. Table 2 also indicates that when the sample size increases and the dimensionality decreases, both CC-SIS and our procedure perform better. Furthermore, this table shows that these factors have less of an effect on the proposed procedure. For instance, when $n = 200$, some values of $\mathcal{P}_a$ obtained by CC-SIS are less than 0.8, but the corresponding values of $\mathcal{P}_a$ of the proposed procedure are close to one. In addition, Table 2 shows that our procedure outperforms CC-SIS in every case, which is consistent with our theoretical analysis because our procedure exhibits the sure screening property. Hence, our procedure also outperforms CC-SIS in the setting of $\Sigma = \Sigma_2$.

In addition, comparing the two methods for different ρ, Tables 1-2 show that when ρ increases, the performance of both procedures deteriorates. This is expected because when the predictors are highly correlated, unimportant predictors may be selected owing to their strong correlations with the true predictors.

We also examine the computational efficiency and empirical convergence of the proposed algorithm for VCMs. Table 3 shows the medians and median of absolute deviations (MADs) of the computing time (seconds), as well as the number of iterations over 1,000 replications. When $p = 1{,}000$, most of the medians of the computing times are below 5 seconds, and the MAD is relatively small; when $p = 2{,}000$, the computing time increases, but the medians are still mostly below 9 seconds and the MADs are again small. In general, the algorithm converges faster as the sample size increases. As shown in Table 3, the algorithm converges after five iterations when $n = 400$, and it usually converges after 10 iterations when $n = 200$. These results show that the proposed algorithm is reasonably efficient.

Example 2. This example examines the performance of the proposed procedure for a binary response. Given $\{U, \mathbf{x}\}$, we generate a binary response, with the probability of $Y = 1$ being $p(U, \mathbf{x})$, as follows:

$$\text{logit}\{p(U, \mathbf{x})\} = \alpha_1(U)X_1 + \alpha_2(U)X_2 + \alpha_3(U)X_3 + \alpha_4(U)X_4, \tag{3.2}$$

where $\text{logit}(t) = \log\{t/(1-t)\}$, which is the logit link in the logistic regression. Model (3.2) implies that $\alpha_j(\cdot) = 0$ for $j > 4$ and $\mathcal{M}_* = \{1, 2, 3, 4\}$. In this example, the coefficients are set to the same values as those in Example 1.

In this example, we consider $p = 1{,}000$ and 2,000, and sample sizes $n = 300$ and 500. All simulation results are based on 1,000 replications, and are summarized in Tables 4-5.

Table 4 shows the values of $\mathcal{P}_j$s and $\mathcal{P}_a$ for the binary responses. Under the design of Σ_1 and $\boldsymbol{\alpha}_1$, X_4 is jointly dependent, but marginally independent of Y. As shown in Table 4, the values of $\mathcal{P}_4$ and $\mathcal{P}_a$ are very close to one, which means our method identifies the predictor that is jointly important, but marginally independent of the response. In general, $\mathcal{P}_4$ is the largest, because the absolute value of $\alpha_4(U)$ is no less than those of the other three coefficient functions, which makes X_4 much easier to identify. If there is no marginal independence, the values of $\mathcal{P}_j$ and $\mathcal{P}_a$ are very close to one. From the table, we see that the values of $\mathcal{P}_a$ are mostly greater than 0.9. In addition, our procedure performs better as the sample size increases, and the dimensionality decreases, consistent with the sure screening property of the method.

Furthermore, comparing the performance of the proposed procedure under different ρ, Table 4 shows that the proposed procedure performs better as the value of ρ decreases, as in Example 1.

Table 5 presents the medians and MADs of the computing time (seconds) and the number of iterations for the binary response over 1,000 simulations. In general, the computing time increases as the sample size and the dimension of the predictors increase. The algorithm converges in five iterations, and is not influenced by the sample size or the dimension of the predictors. This implies that the proposed algorithm works well for a GVCM with a binary response.

Example 3. This example examines the performance of the proposed procedure for a GVCM with a count response. Given $\{U, \mathbf{x}\}$, we generate a count response from a Poisson distribution with mean $\lambda(U, \mathbf{x})$, as follows:

$$\log\{\lambda(U, \mathbf{x})\} = \alpha_1(U)X_1 + \alpha_2(U)X_2 + \alpha_3(U)X_3 + \alpha_4(U)X_4. \tag{3.3}$$

Model (3.3) implies that $\alpha_j(\cdot) = 0$ for $j > 4$ and $\mathcal{M}_* = \{1, 2, 3, 4\}$. In this example, we consider two sets of coefficient functions:

$\boldsymbol{\alpha}_1$: Let $\alpha_1(u) = \alpha_2(u) = \alpha_3(u) = \{2 + 2\sin^2(2\pi u)\}/4$ and $\alpha_4(u) = -0.75\rho * \alpha_1(u)$.

$\boldsymbol{\alpha}_2$: $\alpha_1(u) = -\{3+2\cos^2((\pi/2)u)\}/6$, $\alpha_2(u) = -(3+3u)/6$, $\alpha_3(u) = \{(2-u)^2+2\}/6$, $\alpha_4(u) = \{3 + 2\sin^2((\pi/2)u)\}/6$.

That is, we rescale the $\alpha(\cdot)$ in Example 1 so that its ranges lies between -1 and 1, because the mean function $\lambda(U, \mathbf{x})$ is in the exponential scale of $\alpha(\cdot)$.

In this example, we consider p = 1,000 and 2,000, and sample sizes n = 300 and 500. All simulation results are based on 1,000 replications, and are summarized in Tables 6-7.

Table 6 shows the values of $\mathcal{P}_j$ and $\mathcal{P}_a$ for the count responses. In most cases, the values of $\mathcal{P}_j$ and $\mathcal{P}_a$ are very close to one, regardless of the presence of marginal independence. In general, the proposed procedure performs better as the sample size increases and the dimensionality decreases. Similarly to Examples 1 and 2, the proposed procedure performs better as ρ decreases.

The computing time and the number of iterations of the proposed algorithm are summarized in Table 7. Compared with those in Example 2 for the binary response, the computing time for the count response is relatively shorter. In general, the computing times also increases with n and p. The algorithm converges in fewer steps than in the binary case.

Example 4. This example examines the performance of the HBIC tuning parameter selector. We set $n = 500$, $p = 1{,}000, 2{,}000$, $\Sigma = \Sigma_2$ with $\rho = 0.5$, and $\boldsymbol{\alpha} = \boldsymbol{\alpha}_2$ as the coefficient functions. We set $C_n = \log(\log n)$ for the HBIC, and compare its performance with that of the AIC and BIC tuning parameter selectors. The following three criteria are used to evaluate the performance:

1. P: the probability that the true model is selected;
2. C: the number of predictors selected correctly from four active predictors;
3. I: the number of predictors selected incorrectly as active from among all inactive predictors.

The simulation results based on 200 replications are summarized in Table 8.

Table 8 shows that the AIC, BIC, and HBIC tuning parameter selectors reduce the model complexity significantly, while retaining all active predictors. The HBIC performs much better than the AIC and BIC in terms of controlling the false positives in a linear VCM. For the HBIC, the probability of obtaining the true model is close to one, and the number of false positives is close to zero. For the logistic and Poisson models, the HBIC performs much better than the AIC and the BIC in terms of selecting the true model. The BIC also works well for the logistic and Poisson models, because the probabilities of obtaining the true model are very close to those of the HBIC.

3.2. An application

We illustrate the proposed methodology by means of an empirical analysis of a subset of data collected as part of the Framingham Heart Study (FHS). See Dawber, Meadors and Moore (1951) and Jaquish (2007) for details about the FHS. The subset consists of data on 977 subjects. Here, we wish to investigate the impact of dynamic genetic effects on obesity. In our analysis, we focus on nonrare SNPs, which are those with a minor allele frequency greater than 0.05. In our analysis, we include 4,395 nonrare SNPs with missing rates less than 0.02. According to Wikipedia, a BMI equal to or greater than 25 is considered overweight, and above 30 is considered obese. Thus, the response variable takes the value one if the subject's BMI is greater than 25, and zero otherwise. The response variable denotes a status of overweight or obese. The goal is to identify those SNPs strongly associated with the response in order to examine the dynamic (age-dependent) effect of SNPs and gender on the response. We consider a logistic VCM with u denoting age and 8,791 covariates. For each SNP, both the

dominant effect and the additive effect are considered, and we include gender as a covariate in our analysis. This leads to a high-dimensional logistic VCM with the sample size $n = 977$.

We first apply the proposed screening procedure to the logistic VCM with the number of knots equal to $d_n = 6 \approx 1.5n^{1/5}$. Note that the gender variable is not subject to screening. Thus, we have a total of 29 variables after screening.

We further apply a group Lasso to the model obtained from the screening procedure. The HBIC is used to select the tuning parameter. The Lasso-HBIC selects a model with 20 SNPs. Figure 1 depicts plots of the estimated coefficient functions, along with their pointwise confidence intervals for the selected model. Figure 1 shows that the intercept function changes over age. In addition, the coefficient functions of some SNPs change over age too, although they hover around zero.

4. Conclusion

In this work, we proposed an SJS feature screening procedure for a GVCM with ultrahigh-dimensional covariates. The proposed SJS method differs from the existing SIS method, because the SJS method is based on the joint likelihood of the potential candidate features. We have also proposed an effective algorithm for implementing the feature screening procedure, and show that the proposed algorithm possesses the ascent property. In addition, we establish the sure screening property for the SJS method. We also conduct a numerical study to assess the empirical performance of the proposed procedure. The numerical results imply that the proposed algorithm converges quickly and that the computing time is reasonable.

Supplementary Material

The online Supplementary Material provides for proofs of Theorems 1-3 in Section 2, as well as Tables 1-8 and Figure 1 in Section 3.

Acknowledgements

Guangren Yang's research was supported by the National Nature Science Foundation of China grants 11871173 and 71974076, the National Social Science Foundation of China grant 16BTJ032, the Guangdong Province Nature Science Foundation of China grants 2019A1515010721 and the Fundamental Research Funds for the Central University 19JNYH08. Songshan Yang's research was

supported by NIDA, NIH grant P50 DA039838 and NSF grant DMS 1512422, and Li's research was supported by NIDA, NIH grants P50 DA039838, and P50 DA036107 and NSF grant DMS 1512422. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIDA or the NIH.

References

Cheng, M., Honda, T., Li, J. and Peng, H. (2014). Nonparametric independence screening and structure identification for ultra-high dimensional longitudinal data. *The Annals of Statistics* **42**, 1819–1849.

Cheng, M.-Y., Honda, T. and Zhang, J.-T. (2016). Forward variable selection for sparse ultra-high dimensional varying-coefficient models. *Journal of American Statistical Association* **111**, 1209–1221.

Chu, W., Li, R. and Reimherr, M. (2016). Feature screening for time varying-coefficient models with ultra-high dimensional longitudinal data. *Annals of Applied Statistics* **10**, 596–617.

Cleveland, W. S., Grasse, E. and Shyu, W. M. (1992). Local regression models. In *Statistical Models in S* (Edited by s, J. M. Chambers and T. J. Hastie), 309–376. Wadswoth & Brooks/Cole, Pacific grove CA.

Dawber, T. R., Meadors, G. F. and Moore, F. E., Jr. (1951). Epidemiological approaches to heart disease: the framingham study, *American Journal of Public Health* **41**, 279–286.

de Boor, C. (1978). *A Practical Guide to Splines*, Springer Verlag, Berlin.

Fan, J., Feng, Y. and Song, R. (2011). Nonparametric independence screening in sparse ultra-high dimensional additive models. *Journal of the American Statistical Association* **116**, 544–557.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *Journal of the Royal Statistical Society, (Statistical Methodology)* **70**, 849–911.

Fan, J., Samworth, R. and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *Journal of Machine Learning Research* **10**, 1829–1853.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics* **38**, 3567–3604.

Fan, J., Ma, Y. and Dai, W. (2014). Nonparametric independence screening in sparse ultra-high dimensional varying-coefficient models. *Journal of the American Statistical Association* **109**, 1270–1284.

Hall, P. and Miller, H. (2009). Using generalized correlation to effect variable selection in very high dimensional problems, *Journal of Computational and Graphical Statistics* **18**, 533–550.

Hastie, T. J. and Tibshirani, R. J. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **55**, 757–796.

Jaquish, C. (2007). The framingham heart study, on its way to becoming the gold standard for cardiovascular genetic epidemiology, *BMC Medical Genetics* **8**, 63.

Li, J. and Zhang, W. (2011). A semiparametric threshold model for censored longitudinal data analysis. *Journal of the American Statistical Association* **106**, 685–696.

Li, G., Peng, H., Zhang, J. and Zhu, L.-X. (2012). Robust rank correlation based screening. *The Annals of Statistics* **40**, 1846 - 1877.

Lin, Z. Y. and Bai, Z. D. (2009). *Probability Inequalities*. Science Press, Beijing.

Liu, J., Li, R. and Wu, R. (2014). Feature selection for varying-coefficient models with ultrahigh dimensional covariates. *Journal of American Statistical Association* **109**, 266–274.

Liu, J., Zhong, W. and Li, R. (2015). A selective overview of feature screening for ultra-high dimensional data. *Science in China Series A: Mathematics* **58**, 2033–2054.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd Edition.

Song, R., Yi, F. and Zou, H. (2014). On varying-coefficient independence screening for high dimensional varying-coefficient models. *Statistica Sinica* **24**, 1735–1752.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics* **10**, 1040–1053.

Stone, C. J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics* **13**, 689–705.

Tan, X., Shiyko, M., Li, R., Li, Y. and Dierker, L. (2012). A Time-varying effect model for intensive longitudinal data. *Psychological Methods* **17**, 61–77.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **58**, 267–288.

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association* **104**, 1512–1524.

Wang, L., Kim, Y. and Li, R. (2013). Calibrating nonconvex penalized regression in ultrahigh dimension. *The Annals of Statistics* **41**, 2505–2536.

Wei, F., Huang, J. and Li, H. (2011). Variable selection and estimation in high dimensional varying-coefficient models. *Statistica Sinica* **21**, 1515–1540.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-newton method. *Biometrika* **61**, 439–447.

Xia, X., Yang, H. and Li, J. (2016). Feature screening for generalized varying-coefficient models with application to dichotomous response. *Computational Statistics & Data Analysis* **102**, 85–97.

Xu, C. and Chen, J. (2014). The sparse MLE for ultra-high dimensional feature screening. *Journal of the American Statistical Association* **109**, 1257–1269.

Zhu, L.-P., Li, L., Li, R. and Zhu, L.-X. (2011). Model-free feature screening for ultra-high dimensional data. *Journal of the American Statistical Association* **106**, 1464–1475.

Department of Statistics, School of Economics, Jinan University, Guangzhou, P.R. China 510632.

E-mail: tygr@jnu.edu.cn

Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA.

E-mail: szy125@psu.edu

Department of Statistics and The Methodology Center, The Pennsylvania State University, University Park, PA 16802, USA.

E-mail: rzli@psu.edu

(Received December 2016; accepted July 2018)