Statistica Sinica Preprint No: SS-2017-0362	
Title	Feature Screening in Ultrahigh Dimensional Generalized Varying-coefficient Models
Manuscript ID	SS-2017-0362
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202017.0362
Complete List of Authors	Guangren Yang Songshan Yang and Runze Li
Corresponding Author	Guangren Yang
E-mail	tygr@jnu.edu.cn
Notice: Accepted version subject to English editing.	

Statistica Sinica

Feature Screening in Ultrahigh Dimensional Generalized Varying-coefficient Models

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

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

Abstract: Generalized varying coefficient models are particularly useful for examining dynamic effects of covariates on a continuous, binary or count response. This paper is concerned with feature screening for generalized varying coefficient models with ultrahigh dimensional covariates. The proposed screening procedure is based on joint quasi-likelihood of all predictors, and therefore is distinguished from marginal screening procedures proposed in the literature. In particular, the proposed procedure can effectively identify active predictors that are jointly dependent but marginally independent of the response. In order to carry out the proposed procedure, we propose an effective algorithm and establish the ascent property of the proposed algorithm. We further 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 existing ones via Monte Carlo simulations, and illustrate the proposed procedure by a real data example.

Key words and phrases: Generalized varying-coefficient models, ultrahigh dimen-

sional data, variable screening.

1. Introduction

Generalized linear models have been well studied in the literature. Variable selection via penalized likelihood has been developed for generalized linear models with large dimensional covariates (Tibshirani,1996; Fan and Li, 2001). Ultrahigh dimensional data have been collected in various research areas such as genome-wide association studies, proteomics studies, finance, tumor classification and biomedical imaging. Variable selection methods based on penalized likelihood may not perform well for ultrahigh dimensional data due to their algorithmic stability, computational cost and statistical accuracy (Fan, et al., 2009). Fan and Lv (2008) advocates a two stage approach: (a) reduce ultrahigh dimensional covariates to large dimensional by filtering out a large number of irrelevant covariates based on a marginal screening procedure, and (b) apply variable selection methods to the reduced model with large dimensional covariates. Fan and Lv (2008) proposed a sure independence screening (SIS) procedure for linear models using Pearson correlation coefficient as the marginal utility and further established the sure screening property of their procedure under Gaussian linear model framework. Hall and Miller (2009) proposed a feature screening procedure for transformation linear model by using generalized correlation

and Li, et al. (2012) advocated using rank correlation for screening to deal with heavy-tailed distribution and the presence of outlier. Fan, et al. (2009) proposed a SIS procedure for generalized linear models based on marginal likelihood estimate. More details about these marginal feature screening procedures can be found at the recent review paper on feature screening by Liu, et al. (2015).

Varying coefficient models (VCM) were proposed to deal with "curse of dimensionality" (Cleveland, et al., 1992; Hastie and Tibshirani, 1993). As a natural extension of linear regression models by allowing coefficients varying over a variable such as age and time, the VCM are particularly useful for exploring dynamic pattern of effects and have been used in various research fields (See, e.g., Zhu, et al., 2011; Tan, et al, 2012; Liu, et al, 2014). Feature screening procedures for VCM with ultrahigh dimensional covariates (referred to as ultrahigh dimensional VCM for short) have been proposed in the literature. Liu, et al. (2014) developed an SIS procedure for ultrahigh dimensional VCM by taking conditional Pearson correlation coefficients as marginal utility for ranking importance of predictors. Fan, et al. (2014) proposed an SIS procedure for ultrahigh dimensional VCM by extending B-spline techniques in Fan, et al. (2011) for additive models. Xia, et al. (2016) further extends the SIS procedure proposed in Fan, et al. (2014)

to generalized varying coefficient models (GVCM). Cheng, et al. (2016) proposed a forward variable selection procedure for ultrahigh dimensional VCM based on techniques related B-splines regression and grouped variable selection. Song, et al. (2014) extended the proposal of Fan, et al. (2014) for longitudinal data without taking into within-subject correlation, while Chu, et al. (2016) proposed an SIS procedure for longitudinal data based on weighted residual sum of squares to use within-subjection correlation to improve accuracy of feature screening. Although feature screening for ultrahigh dimensional VCM is an active research topic in the literature, there is little work on joint feature screening for ultrahigh dimensional GVCM, which is particularly useful to examine dynamic effects of covariates on a binary, count or continuous response. For example, Li and Zhang (2011) proposed a new semiparametric threshold model for censored longitudinal data analysis. Cheng, et al. (2014) offered a new automatic procedure for finding a sparse semivarying coefficient model, which is widely accepted for longitudinal data analysis. This paper intends to fill this gap.

In this paper, we propose a new feature screening procedure for ultrahigh-dimensional GVCM. The proposed procedure is based on joint likelihood of potential active predictors and therefore is distinguished from the existing SIS procedures (Fan, et al., 2014; Liu, et al., 2014; Xia, et al., 2016) in

that the proposed procedure is not a marginal screening procedure. Wang (2009) proposed a forward regression approach to feature screening in ultrahigh dimensional linear models. Cheng, et al. (2016) further extended the forward regression procedure for ultrahigh dimensional VCM based on techniques related B-splines regression and grouped variable selection. Xu and Chen (2014) proposed a feature screening procedure for generalized linear models via the sparsity-restricted maximum likelihood estimator. As demonstrated in Wang (2009), Xu and Chen (2014) and Cheng, et al. (2016), their approaches can perform better than the sure independence screening procedures, and can effectively identify predictors that are jointly dependent but marginal independent of the response. In this paper, we develop a new screening procedure for the ultra-high dimensional GVCM based on joint likelihood of potential active predictors. The proposed procedure can effectively identify active predictors that are jointly dependent but marginal independent of the response without performing an iterative procedure. We develop a computationally effective algorithm to carry out the proposed procedure and establish the ascent property of the proposed algorithm. We further 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. In summary, this work

makes the following major contributions to the literature. (a) We propose a sure joint screening (SJS) procedure for ultrahigh dimensional GVCM. We further propose an effective algorithm to carry out 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 propose a new feature screening for the ultrahigh dimensional GVCM, and develop an effective algorithm for the proposed screening procedure. We further study theoretical properties of the proposed procedure and algorithm. In Section 3, we present numerical comparisons and an empirical analysis of a real data example. Some discussion and conclusion remarks are given in Section 4. Technical proofs are given in the Appendix.

2. New feature screening procedure for generalized varying coefficient models

Let Y be the response variable and $\{\mathbf{x}, U\}$ its associated covariates, where $\mathbf{x} = (X_1, \cdots, X_p)$ and U be 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 un-

specified smooth regression coefficient functions. Here it is assumed that all $\alpha_j(\cdot)$'s are nonparametric functions and the support of U is finite and denoted by $[a, b]$.

Suppose that $\{U_i, \mathbf{x}_i, Y_i\}$, $i = 1, \ldots, n$, constitute an independent and identically distributed 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 $Q(\mu, y) = \int_\mu^y \frac{s-y}{V(s)} ds$, or equivalently $\frac{\partial Q(\mu,y)}{\partial \mu} = \frac{y-\mu}{V(\mu)}$, 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 B-spline regression method. Let $\mathcal{S}_n$ be the space of polynomial splines of degree $l \geq 1$ and $\{\psi_{jk}, k = 1, \ldots, d_{n_j}\}$ 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

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

for some coefficients $\{\beta_{jk}\}_{k=1}^{d_{n_j}}$. Here d_{n_j} increases with n. We allow d_{n_j} to be different for different j since different coefficient functions may have dif-

ferent smoothness. Under some conditions, each nonparametric coefficient function $\alpha_j(U), j = 1, \cdots, p$ can be well approximated by functions in $\mathcal{S}_n$.

Substituting (2.3) into (2.2), the maximum quasi-likelihood estimate of (2.2) is to maximize

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

with respect to $\boldsymbol{\beta}$, where $\mathbf{z}_i = (X_{i1}\boldsymbol{\psi}_1(U_i)^T, \cdots, X_{ip}\boldsymbol{\psi}_p(U_i)^T)^T$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \cdots, \boldsymbol{\beta}_p^T)^T$. With slight abuse 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 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, and the others do not have impact on the response. We next propose a feature screening procedure for model (2.1).

2.1 A new feature screening procedure

Denote $\|\alpha_j(U)\|_2 = [E\alpha_j^2(U)]^{1/2}$, 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, $\tau(s)$ stands for the cardinality of s. Suppose the effect of $\mathbf{x}$ is sparse, and the true value of $\boldsymbol{\alpha}(U)$ is $\boldsymbol{\alpha}^*(U)$, so $\boldsymbol{\beta}$ is corresponding to $\boldsymbol{\beta}^*$. Denote $s^* = \{j : \|\alpha_j(U)\|_2 > 0\}$. By sparsity,

we means 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. Theoretically we may formulate this problem to be an optimization problem as below:

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

for a pre-specified 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:

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

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 $j = 1, \cdots, p$, the maximization problem in (2.6) is equivalent to

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

For high dimensional problems, it becomes almost impossible to solve the constrained maximization problem (2.7) directly. Alternatively, we consider a proxy of the quasi-likelihood function. It follows by the Taylor expansion for the quasi-likelihood function $\ell(\boldsymbol{\gamma})$ at $\boldsymbol{\beta}$ lying within a neighbor 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 large P_t, small n problems (i.e. $P_t \gg n$), $\ell''(\boldsymbol{\beta})$ becomes not invertible. Low computational cost is always desirable for feature screening. To cope with singularity of the Hessian matrix and save 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 $W(\boldsymbol{\beta}) = \text{diag}(W_1(\boldsymbol{\beta}), \cdots, W_p(\boldsymbol{\beta}))$, a block diagonal matrix with $W_j(\boldsymbol{\beta})$ being 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 will use $W_j(\boldsymbol{\beta}) = -\partial^2\ell(\boldsymbol{\beta})/\partial\boldsymbol{\beta}_j\partial\boldsymbol{\beta}_j^T$.

It can be seen that $h(\boldsymbol{\beta}|\boldsymbol{\beta}) = \ell(\boldsymbol{\beta})$, and under some conditions, $h(\boldsymbol{\gamma}|\boldsymbol{\beta}) \leq \ell(\boldsymbol{\beta})$ for all $\boldsymbol{\gamma}$. This ensures the ascent property. See Theorem 1 below for more details. Since $W(\boldsymbol{\beta})$ is a block diagonal matrix, $h(\boldsymbol{\gamma}|\boldsymbol{\beta})$ is an additive function of γ_j for any given $\boldsymbol{\beta}$. The additivity enables us to have a closed form solution for the following maximization problem

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

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 $j =$

$1, \cdots, p$, and $\tilde{\boldsymbol{\gamma}} = (\tilde{\boldsymbol{\gamma}}_1^T, \cdots, \tilde{\boldsymbol{\gamma}}_p^T)^T = \boldsymbol{\beta} + u^{-1}W^{-1}(\boldsymbol{\beta})\ell'(\boldsymbol{\beta})$ 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 $j = 1, \cdots, p$, and sort g_j so that $g_{(1)} \geq g_{(2)} \geq \cdots \geq g_{(p)}$. The solution of maximization problem (2.9) is the hard-thresholding rule defined below

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

This enables us to effectively screen features by using the following algorithm.

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

Step 2. Set $t = 0, 1, 2, \cdots$, iteratively conduct Step 2a and Step 2b below 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 $g_{(1)}^{(t)} \geq g_{(2)}^{(t)} \geq \cdots \geq g_{(p)}^{(t)}$, the order statistics of $g_j^{(t)}$s. Set $S_t = \{j : g_j^{(t)} \geq g_{(m+1)}^{(t)}\}$, the nonzero index set.

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

Remark: Unlike the screening procedures based on marginal partial likelihood methods, our proposed procedure is to iteratively update $\boldsymbol{\beta}$ using Step 2. This enables the proposed screening procedure to incorporate correlation information among the predictors through updating $\ell_p'(\boldsymbol{\beta})$ and $\ell_p''(\boldsymbol{\beta})$. Thus, the proposed procedure is expected to perform better than the marginal screening procedures when there are some predictors that are marginally independent. Meanwhile, since each iteration in Step 2 can avoid large-scale matrix inversion and, therefore, it can be carried out with low computational costs.

Theorem 1. *Let* $\{\boldsymbol{\beta}^{(t)}\}$ *be the sequence defined in Step 2b in 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* $\lambda_{\min}(A)$ *stands for the maximal and the minimal eigenvalues of a matrix* A*, respectively. If* $u_t \geq \rho^{(t)}$*, then*

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

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

Theorem 1 claims the ascent property of the proposed algorithm if u_t is appropriately chosen. 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 last step. This theorem also provides us some insights about choosing u_t in practical implementation. For varying coefficient models: $E(Y|U, \mathbf{x}) = \mathbf{x}^T\boldsymbol{\alpha}(U)$, 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 i-th row being $\mathbf{z}_i^T$. Thus,

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

which does not depend on the step of iteration t. If $\mathbf{z}_i$'s are marginally standardized so that its marginal sample mean and sample standard deviation equal 0 and 1, respectively, then $\text{diag}(\mathbf{Z}^T\mathbf{Z})^{-1/2}(\mathbf{Z}^T\mathbf{Z})\text{diag}(\mathbf{Z}^T\mathbf{Z})^{-1/2}$ is the corresponding sample correlation matrix of $\mathbf{z}_i$'s. 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 notation $\mathbf{x}_s = \{x_j, j \in s\}$ and associated coefficients $\boldsymbol{\alpha}_s(U) = \{\alpha_j(U), j \in s\}$ corresponding 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 $s^* = \{j : E\alpha_j^2(U) > 0, 1 \leq j \leq p\}$ 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 some theoretical justifications for the screening procedure for the GVCM. The sure screening property (Fan and Lv, 2008)) is referred to as

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

To establish this sure screening property for the proposed feature screening method, we introduce some additional notations as follows. 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 $\tau(s^*) = q < m$. So, we define

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

as the collections of the over-fitted models and the under-fitted models. We investigate the asymptotic properties of $\hat{\boldsymbol{\beta}}_m$ under the scenario where 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 only serve to facilitate theoretical understanding of the proposed feature screening 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,

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

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 exists $w_1, w_2 > 0$ and for some non-negative constants τ_1, τ_2 such that $\tau_1 + \tau_2 < 1/2$ with

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

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

(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,

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

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

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

(a) (de Boor, 1978) For $k = 1, \ldots, d_n$, $\psi_{jk}(U) \geq 0$ 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 $C_3 d_n^{-1} \leq E\psi_{jk}^2(U) \leq C_4 d_n^{-1}$.

(b) (Stone, 1982, 1985) If $\{\alpha_j, j = 1, 2, \cdots, p\}$ is a set of functions in $\mathcal{F}$ described in condition (C2), there exists a positive constant C_5 that does not depend on $\alpha_j(U)$ so that the uniform approximation error has the following bound. $\rho = \sup_{U \in [a,b]} \|\alpha_j(U) - \alpha_{nj}(U)\|_2 \leq C_5 d_n^{-\upsilon}, \forall j$, 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 the properties (a), (b) and Condition (C3), we can derive that

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

Condition (C3) states a few requirements for establishing the sure screening property of the proposed procedure. The first one 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 minimal signal condition in the lit-

erature. Minimal signal condition is a commonly-imposed assumption in existing work on marginal feature screening for other model (e.g, Liu, et al., 2014). By (2.12), it 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. Meanwhile, together with (C6), it confines an appropriate order of m that guarantees the identifiability of s^* over s for $\tau(s) \leq m$. For varying coefficient model discussed in Section 2.1, Condition (C6) requires

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

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 to be exponentially light, as shown in the following Lemma 1. As for Condition D corresponds to our Condition (C6), so Condition (C6) can ensure Y bound.

Remark: In particular, our proposed screening procedure is based on joint quasi-likelihood of all predictors. However, Fan, Ma and Dai (2014) investigate marginal nonparametric screening methods to screen variables in sparse ultra-high-dimensional varying-coefficient models. As for conditions (v)-(vi) in Fan, Ma and Dai (2014), conditions (v) and (vi) are requirements

for the tail distribution of each covariate, and the noise, to establish the sure screening property. However, errors need to be independent but not normally distributed. Corresponding to our condition (C6), we only need to assume the minimize and maximum eigenvalues of Hessian matrix are bounded.

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

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

The proof is given in the Appendix. The sure screening property is an appealing property of a screening procedure since it ensures that the true active predictors are retained in the model selected by the screening procedure. We establish the sure screening property under weaker conditions imposed in Fan, et al. (2014) and Xia, et al. (2016).

One has to specify the value of m in practical implementation. As to the choice of m, there are two scenarios. The first one chooses m by a data-driven method that described in Section 2.3. The second one is an ad hoc method. In the literature of 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). Since 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 3.1, 3.2 and 3.3. Although it is an ad hoc choice, it works reasonably well in our numerical examples. With this choice of m, one is ready to further apply existing methods such as the penalized quasi-likelihood method to further remove inactive predictors. To be distinguished from the SIS procedure, the proposed procedure is referred to as sure joint screening (SJS) procedure.

2.3 Choice of m

Feature screening may be used in various contexts. In some contexts, people may treated m as a pre-specified value. For example, due to budget constraint, a biologist may be able to examine up to m genes that potentially associate with a certain phenotype. In other contexts, people may 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 to select m by minimizing the high-dimensional BIC score:

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

where $\hat{\boldsymbol{\beta}}_j = (\hat{\beta}_{j1}, \ldots, \hat{\beta}_{jd_n})$, $j = 1, \ldots, m$, and C_n is a sequence of numbers that diverges to ∞. Wang, et al. (2013) proposed the HBIC for selecting tuning parameter in the penalized least squares method for high dimensional

linear models. Here we modified their proposal for high dimensional generalized varying-coefficient models. In our simulation, we take $C_n = \log\log n$, and compare its performance with AIC and BIC tuning parameter selectors defined in the same manner. It is worth to noting that the proposed tuning parameter HBIC selector requires to search over $m = 1, 2, \cdots, [n/d_n]$. This is distinguished from that the classical AIC and BIC used for subset selection requires to search over subsets. Thus, the tuning parameter selector does not require expensive computational cost.

Recall notation S_+^m and S_-^m defined in (2.11). Theorem 3 below 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 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 3.4, we will 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 procedures and compare it with the existing ones. All simulation are conducted by using R code. Examples 3.1, 3.2 and 3.3 examine the performance of the proposed screening procedures. Following the literature of feature screening (e.g, Fan and Lv, 2008), we set $m = [n/\log(n)]$ in these examples. Example 3.4 examine the performance of the proposed HBIC, and m is determined by minimizing the HBIC score.

3.1 Simulation studies

In our simulation, the covariate 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 all the predictors $X_1, ..., X_p$ are correlated with each other. In our simulation, we consider two scenarios for $\Sigma = (\sigma_{ij})$

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

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

In our numerical studies, we set the number of B-spline basis functions to be $d_{n_j} = 5, j = 1, \cdots, p$ 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 the true predictors among 1000 simulations.

P_j: The proportion of submodels $\hat{\mathcal{M}}$ with size d that contain X_j among 1000 simulations.

Example 3.1. This example is designated to compare the proposed screening procedure with existing SIS procedures for VCM. Since the proposal of Xia, et al. (2016) under the setting of VCM coincides with that in Fan, et al. (2014), which shares the same spirit as that of Liu, et al. (2014), and Song, et al. (2014) and Chu, et al. (2016) were proposed for longitudinal data, we will concentrate on our comparison with CC-SIS proposed by Liu, et al. (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 $\boldsymbol{\varepsilon} \sim N(0,1)$. 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 $p = 1000$ and 2000, and the sample size $n = 200$ and 400. All simulation results are based on 1000 replications. Simulation results are summarized in Tables 14 and 3.

Table 14 shows the values of $\mathcal{P}_1, \cdots, \mathcal{P}_4$ and $\mathcal{P}_a$ for 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 14, when there exists marginal independence, CC-SIS is unable to detect X_4 whose values of $\mathcal{P}_4$ and $\mathcal{P}_a$ are near zero as expected. However, our method can identify X_4 in this setting and the corresponding values of $\mathcal{P}_4$ and $\mathcal{P}_a$ are close to one. Therefore, our new 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, as the detecting probabilities are close to one. However, CC-SIS performs better when the sample size increases and the dimensionality decreases. On the other hand, those factors have less influences on the new procedure than 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 this setting. In summary, when $\Sigma = \Sigma_1$, regardless of whether marginal independence exists, our new procedure outperforms

CC-SIS.

Table 2 shows the values of $\mathcal{P}_j$s and $\mathcal{P}_a$ for continuous response with $\Sigma = \Sigma_2$. There is no predictor that is jointly dependent but marginally independent of Y. Hence both of the CC-SIS and the new procedure perform well, as most of the values of $\mathcal{P}_a$ are greater than 0.9. Table 2 also indicates that when the sample size increases and the dimensionality decreases, both CC-SIS and our new procedure perform better. Furthermore, this table also shows that those factors have less effect on our new 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 new procedure are close to one. Besides, Table 2 shows that the new procedure performs better than CC-SIS in every case, which is consistent with our theoretical analysis since our new procedure has the sure screening property. Hence, our new procedure also outperforms CC-SIS in the setting of $\Sigma = \Sigma_2$.

In addition, comparing the two methods with different ρ's, Tables 14 and 2 show that when ρ increases, the performance of CC-SIS and the new procedure become worse. This is expected because when the predictors are highly correlated, the unimportant predictors may be selected due to their strong correlations with the true predictors.

We also examine the computational efficiency and empirical conver-

gence of the proposed algorithm for VCM. Table 3 shows the medians and median of absolute deviations (MADs) of computing time (seconds), and the number of iterations over 1000 replications. When $p = 1000$, most of the medians of the computing times are below 5 seconds, and the MAD is pretty small; when $p = 2000$, the computing time increases, but the medians are still mostly below 9 seconds and the MADs are also small. In general, the algorithm converges faster as the sample size increases. As shown in Table 3, the algorithm converges after 5 iterations when $n = 400$ and it usually converges after 10 iterations when $n = 200$. All of the facts above show that the proposed algorithm is reasonably efficient.

Example 3.2. This example is designated to examine the performance of the proposed procedures for binary response. Given $\{U, \mathbf{x}\}$, we generate a binary response with the probability of $Y = 1$ being $p(U, \mathbf{x})$ defined below:

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

where $\text{logit}(t) = \log\{t/(1-t)\}$, 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 be the same as those in Example 3.1.

In this example, we consider $p = 1000$ and 2000, and the sample size $n = 300$ and 500. All simulation results are based on 1000 replications, and

are summarized in Tables 4 and 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 is able to identify the predictor that is jointly important but marginally independent of the response. In general, $\mathcal{P}_4$ is the largest and this is 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 be identified. If there is no marginal independence, the values of $\mathcal{P}_j$s 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, which is also consistent to the sure screening property of the new method.

Furthermore, comparing the performances of the new procedure under different ρ's, Table 4 shows that the new procedure performs better as the value of ρ decreases. This is the same as the pattern for Example 3.1.

Table 5 presents the medians and MADs of computing time (seconds) and the number of iterations for binary response over 1000 simulations. In general, the computing time increases as the sample size and the dimension of predictors increases. The algorithm converges in 5 iterations and it is not

influenced by the sample sizes and the dimension of the predictors. This implies that the proposed algorithm works well for GVCM with binary response.

Example 3.3. This example is designated to examine the performance of the proposed procedures for GVCM with count response. Given $\{U, \mathbf{x}\}$, we generate a count response from a Poisson distribution with mean $\lambda(U, \mathbf{x})$ defined below.

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

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 $\alpha_4(u) = -0.75\rho * \alpha_1(u)$.

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

That is, we re-scale the $\alpha(\cdot)$s in Example 3.1 so that their ranges lie between -1 and 1 since the mean function $\lambda(U, \mathbf{x})$ is in the exponential scale of $\alpha(\cdot)$s.

In this example, we consider $p = 1000$ and 2000, and the sample size $n = 300$, and 500. All the simulation results are based on 1000 replications, and are summarized in Tables 6 and 7.

Table 6 shows the values of $\mathcal{P}_j$s and $\mathcal{P}_a$ for the count responses. In most cases, the values of $\mathcal{P}_j$s and $\mathcal{P}_a$ are very close to one, regardless of whether there exists the marginal independence. In general, the proposed procedure performs better as the sample size increases and the dimensionality decreases. Similar to those in Examples 3.1 and 3.2, the proposed procedure has a better performance with smaller ρ's.

Computing time and the number of iterations of the proposed algorithm are summarized in Table 7. Compared with those in Example 3.2 for binary response, the computing time for count response is relatively shorter. In general, the computing times also become larger as n and p increases. The algorithm converges in fewer steps than the binary case.

Example 3.4. This example is designed to examine the performance of HBIC tuning parameter selector. We set $n = 500$, $p = 1000, 2000$, $\Sigma = \Sigma_2$ with $\rho = 0.5$ and $\boldsymbol{\alpha} = \boldsymbol{\alpha}_2$ is the coefficient functions. We set $C_n = \log(\log n)$ in HBIC, and compare the performance of HBIC with those of the AIC and BIC tuning parameter selectors. The following three criteria are used to evaluate the performances:

1. P: the probability that the true model is selected;

2. C: the number of correctly selected predictors from four active pre-

dictors;

3. I: the number of predictors incorrectly selected as active ones from 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 can reduce model complexity significantly, while retain all active predictors. As seen from Table 8, the HBIC performs much better than the AIC and theBIC in terms of controlling the false positives in linear varying coefficient model. 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 logistic model and Poisson model, the HBIC performs much better than the AIC and the BIC in terms of selecting the true model. The BIC also works well for logistic model and Poisson model, since 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 an empirical analysis of a subset of data collected the Framingham Heart Study (FHS, for short). See Dawber, et al. (1951) and Jaquish (2007) for details about FHS. The

data subset consists of data for 977 subjects. Of interest is to investigate the impact of dynamic genetic effects on obesity. In our analysis, we focus on nonrare SNPs. Here, nonrate SNPs are referred to those SNP whose the minor allele frequency of a SNP is great than 0.05. In our analysis, we include 4395 nonrare SNPs with missing rates being 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, we define the response variable to be 1 if this subject's BMI is greater than 25 and 0 otherwise. The response variable indeed stands for the status of overweight or obese. The goal is to identify the SNPs strongly associated with the response. To examine the dynamic (age-dependent) effect of SNPs and gender on the response. We consider a logistic varying coefficient models with u being age, and 8791 covariates since for each SNP, both dominant effect and additive effect are considered, in addition to include gender as a covariate in our analysis. This leads to high-dimensional logistic varying coefficient model with the sample size $n = 977$.

We first apply the proposed screening procedure to the logistic varying coefficient model with the number of knots being $d_n = 6 \approx 1.5n^{1/5}$. Note that the gender variable is not subject to screening. Thus, there are total 29 variables after screening.

We further apply group lasso to the model obtained from the screening procedure. HBIC is used to select the tuning parameter. The lasso-HBIC selects a model with 20 SNPs. Figure 1 depicts the plots of the estimated coefficient functions along with their pointwise confidence intervals for the model selected by lasso-HBIC. From Figure 1, it can be seen that the intercept function changes over age, and coefficient functions of some SNP are also changing over age too, although they hover around zero.

4. Discussions

In this work, we proposed a SJS feature screening procedure for GVCM with ultrahigh dimensional covariates. The proposed SJS is distinguished from the existing SIS in that the SJS is based on the joint likelihood of potential candidate features. We proposed an effective algorithm to carry out the feature screening procedure, and show that the proposed algorithm possesses an ascent property. We study the sample property of SJS, and establish the sure screening property for SJS. We also conduct numerical study to access the empirical performance of the proposed procedure. The numerical results implies that the proposed algorithm converges quickly and computing time is reasonable.

Supplementary Materials

Supplementary materials include Proofs of Theorem 1-3 in Section 2, Table

1-8 in Sections 3 and Figure 1 in Section 3.

Acknowledgements

Guangren Yang's research was supported by the National Nature Science Foundation of China grant 11471086, the National Social Science Foundation of China grant 16BTJ032, the Fundamental Research Funds for the Central Universities 15JNQM019, the National Statistical Scientific Research Center Projects 2015LD02, Education Bureau of Guangdong Province 2016WTSCX007 and Science and Technology Program of Guangzhou 2016201604030074. Songshan Yang's research was supported by a NIDA, NIH grant P50 DA039838 and a NSF grant DMS 1512422 and Li's research was supported by was supported by NIDA, NIH grants P50 DA039838, P50 DA036107 and a 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

Chen, J. and Chen, Z. (2008). Extended Bayesian information criterion for model selection with large model space. *Biometrika*, **94**, 759-771.

Chen, J. and Chen, Z. (2012). Extended BIC for Small-n-Large-p Sparse

GLM. *Statistics Sinica*, **22**, 555-574.

Cheng, M., Honda, T., Li, J. and Peng, H. (2014). Nonparametric independence screening and structure identification for ultra-high dimensional longitudinal data. *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. (eds, J. M. Chambers and T. J. Hastie), Pacific grove CA: Wadswoth & Brooks/Cole, 309 - 376.

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, Series B*, **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. *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,* **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. *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, Second 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. *Annals of Statistics*, **10**, 1040-1053.

Stone, C. J. (1985). Additive Regression and Other Nonparametric Models. *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,* **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. *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 Statis-*

tical Association, **106**, 1464-1475.

Guangren Yang

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

Email: tygr@jnu.edu.cn

Songshan Yang

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

Email: szy125@psu.edu

Runze Li

Department of Statistics and The Methodology Center,

The Pennsylvania State University, University Park, PA 16802.

E-mail: rzli@psu.edu