Statistica Sinica **24** (2014), 1735-1752
doi:http://dx.doi.org/10.5705/ss.2012.299

ON VARYING-COEFFICIENT INDEPENDENCE SCREENING FOR HIGH-DIMENSIONAL VARYING-COEFFICIENT MODELS

Rui Song[1], Feng Yi[2] and Hui Zou[2]

[1]*North Carolina State University and* [2]*University of Minnesota*

Abstract: Varying coefficient models have been widely used in longitudinal data analysis, nonlinear time series, survival analysis, and so on. They are natural nonparametric extensions of the classical linear models in many contexts, keeping good interpretability and allowing us to explore the dynamic nature of the model. Recently, penalized estimators have been used for fitting varying-coefficient models for high-dimensional data. In this paper, we propose a new computationally attractive algorithm called IVIS for fitting varying-coefficient models in ultra-high dimensions. The algorithm first fits a gSCAD penalized varying-coefficient model using a subset of covariates selected by a new varying-coefficient independence screening (VIS) technique. The sure screening property is established for VIS. The proposed algorithm then iterates between a greedy conditional VIS step and a gSCAD penalized fitting step. Simulation and a real data analysis demonstrate that IVIS has very competitive performance for moderate sample size and high dimension.

Key words and phrases: Penalized regression, sure screening property, varying-coefficient models.

1. Introduction

It is more and more common to confront situation in which when the number of predictor variables p is in the tens of thousands, potentially much larger than the number of observations n. Examples include data from microarrays, proteomics, brain images, etc. Variable selection hence becomes an increasingly important task. There is a vast literature on variable selection for regression problems under linear regression settings. Recent developments mostly focus on penalized methods, including the LASSO (Tibshirani (1996)), SCAD (Fan and Li (2001)), the Dantzig selector (Candes and Tao (2007)) and their variations. These methods have been thoroughly studied for variable selection with high-dimensional data (van de Geer (2008); Bickel, Ritov and Tsybakov (2009); Meinshausen and Yu (2009)). A computationally simpler method that can work well in practice for very high dimensional data is sure independence screening (SIS), demonstrated in Fan and Lv (2008) in the classical regression

context. The sure independence screening recruits the features with best marginal utility, which corresponds to the largest marginal absolute Pearson correlation between the response and predictor in the context of least-squares regression for the linear model. Fan and Lv (2008) showed that SIS has a sure screening property: with probability close to 1, it can retain all important features in the model. After sure screening, the remaining covariates are used to fit a penalized linear regression model. Recent works on sure screening include Fan, Samworth and Wu (2009), Fan and Song (2010), Fan, Feng and Song (2011), Zhu et al. (2011), Li, Zhong and Zhu (2012), Li et al. (2012), among others.

This paper concerns variable selection in the varying coefficient model, an important and useful generalization of the linear regression model. and etc. demanding. It is common to present data as longitudinal observations $\{Y_{ij}, X_i(t_{ij}), t_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots, n_i\}$, where t_{ij} and n_i are the time of the jth measurement and the number of repeated measurement for the ith subject, respectively, Y_{ij} and $\mathbf{X}_i(t_{ij}) = (X_{1i}(t_{ij}), \ldots, X_{pi}(t_{ij}))'$ are the ith subject's observed outcome and covariates at time t_{ij}. Examples include longitudinal data analysis (Hoover et al. (1998)) and functional response models (Rice (2004)) among others. Interest focuses mostly on investigating the time-dependent effects of the covariates on responses measured repeatedly and/or longitudinally. Different regression models are proposed for this type of data, among them the varying-coefficient model has gained a lot of popularity. For variable selection with varying-coefficients models, Wang, Li and Huang (2008) and Wang, Li and Huang (2008) both proposed a group penalization method in the fixed p case, and Wei, Huang and Li (2011) recently extended this work to the case of diverging p. However, for very large p, these penalized methods remain computationally demanding.

In this paper, we consider screening of the important covariates in varying-coefficient models by ranking the magnitude of nonparametric marginal correlations. The magnitude of the proposed screener can preserve the non-sparsity of the varying-coefficient models under some reasonable conditions, even with converging minimum strength of signals. Our work can be regarded as an extension of the SIS procedures proposed in Fan and Lv (2008) and Fan, Feng and Song (2011), with differences and our contributions highlighted as follows. Here the minimum distinguishable signal is related to the stochastic error in estimating the nonparametric components, the approximation errors in modeling nonparametric components, and the number of observations within each subject. Efforts were made to study the influence of the longitudinal observations on the sure screening property. This led to a result on the extent to which the dimensionality can be reduced by varying-coefficient independence screening. The dimensionality of the model is allowed to grow near exponentially with the sample size. We propose

an iterative nonparametric independence screening procedure, IVIS-gSCAD, to reduce the false positive rate and stabilize the computation. Additionally, we use B-spline to approximate the nonparametric coefficients, which is computationally easier than using local polynomial regression.

The outline of the paper is as follows. In Section 2, we propose the varying-coefficient independence screening method based on B-spline to approximation. In Section 3, an iterative varying-coefficient screening (IVIS) method is proposed. In Section 4, simulation studies are Brought up to demonstrate the performance of the proposed method. In addition, a data set is used as an illustration of varying-coefficient regression models. The paper concludes in Section 5 and the Web Appendix contains all technical proofs.

2. Varying-coefficient Independence Screening

Consider the population $\{\mathbf{X}(t), Y(t)\}$ from the time-varying coefficient model

$$Y(t) = \mathbf{X}(t)'\boldsymbol{\alpha}(t) + \epsilon(t),\ t \in \mathcal{T}, \tag{2.1}$$

where $\mathbf{X}(t) = (X_1(t), \ldots, X_p(t))'$ are the covariates, $\boldsymbol{\alpha}(t) = (\alpha_1(t), \ldots, \alpha_p(t))'$ are the time-varying coefficients, $\epsilon(t)$ is a mean zero stochastic process, $Y(t)$ is a mean zero outcome function, and $\mathcal{T}$ is the time interval in which the measurements are taken.

Our purpose is to identify the set $\mathcal{M}_\star = \{l :\ \alpha_l(t) \neq 0\}$. We consider p marginal nonparametric regression problems:

$$\min_{\beta(t) \in L_2(P)} E(Y(t) - X_l(t)\beta(t))^2, \tag{2.2}$$

where P denotes the joint distribution of $\mathbf{X}(t)$ and $Y(t)$, and $L_2(P)$ is the class of square integrable functions under the measure P. The minimizer of (2.2) is $\beta_{l0}(t) = EX_l(t)Y(t)$. The population version of VIS is to screen the time-varying coefficients $\alpha_l(t)$ in (2.1) according to $|EX_l(t)Y(t)|$ to select a small group of covariates via thresholding.

Suppose there is a random sample of n independent subjects $\{\mathbf{X}_i(t), Y_i(t)\}_{i=1}^n$ from model (2.1). Let t_{ij} and n_i be the time of the jth measurement and the number of repeated measurement for the ith subject. $Y_{ij} = Y_i(t_{ij})$ and $\mathbf{X}_i(t_{ij}) = (X_{1i}(t_{ij}), \ldots, X_{pi}(t_{ij}))'$ are the ith subject's observed outcome and covariates at time t_{ij}. Based on longitudinal observations $\{Y_{ij}, \mathbf{X}_i(t_{ij}), t_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots, n_i\}$, the model can be written as:

$$Y_i(t_{ij}) = \mathbf{X}_i(t_{ij})'\boldsymbol{\alpha}(t_{ij}) + \epsilon_i(t_{ij}). \tag{2.3}$$

For $l = 1, \ldots, p$, let $\{B_{lk}(\cdot), k = 1, \ldots, K_l\}$ denote a basis of B-spline functions. Each $\beta_l(t)$ can be approximated by a linear combination of B-spline basis functions. We consider marginal weighted least square estimation based on

B-spline expansion, for $l = 1, \ldots, p$, by minimizing $e_l = \sum_{i=1}^{n} \omega_i \sum_{j=1}^{n_i} \Big(Y_{ij} - \sum_{k=1}^{K_l} X_{lij} B_{lk}(t_{ij}) \gamma_{lk} \Big)^2$ with respect to the γ_{lk}. Choices of ω_i can be 1 or $1/n_i$, equal weight to observations and equal weight to subjects respectively.

Let $\boldsymbol{\gamma}_l = (\gamma_{l1}, \ldots, \gamma_{lK_l})'$. Define $\mathbf{B}_l(t) = (B_{l1}(t), \ldots, B_{lK_l}(t))'$, $\mathbf{U}_{lij} = X_{lij} \mathbf{B}_l(t_{ij})$, $\mathbf{U}_{li} = (\mathbf{U}_{li1}, \ldots, \mathbf{U}_{lin_i})'$ and $\mathbf{W}_i = \mathrm{diag}(w_i, \ldots, w_i)$ with size n_i. Let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$ and $X_{lij} = X_{li}(t_{ij})$ for $j = 1, \ldots, n_i$ and $\mathbf{X}_{li} = (X_{li1}, \ldots, X_{lin_i})'$. We can express e_l as $e_l = e_l(\boldsymbol{\gamma}_l) = \sum_{i=1}^{n} (\mathbf{Y}_i - \mathbf{U}_{li}\boldsymbol{\gamma}_l)' \mathbf{W}_i (\mathbf{Y}_i - \mathbf{U}_{li}\boldsymbol{\gamma}_l)$. Let $\mathbf{U}_l'\mathbf{W}\mathbf{Y} = \sum_i \mathbf{U}_{li}'\mathbf{W}_i\mathbf{Y}_i$ and $\mathbf{U}_l'\mathbf{W}\mathbf{U}_l = \sum_i \mathbf{U}_{li}'\mathbf{W}_i\mathbf{U}_{li}$. Since $\mathbf{U}_l'\mathbf{W}\mathbf{U}_l$ is invertible with probability approaching one (as is established in Lemma 1 in the Web Appendix), the unique minimizer of $e_l(\boldsymbol{\gamma}_l)$ is

$$\hat{\boldsymbol{\gamma}}_l = \Big(\mathbf{U}_l'\mathbf{W}\mathbf{U}_l \Big)^{-1} \mathbf{U}_l'\mathbf{W}\mathbf{Y}. \tag{2.4}$$

Let $\hat{\beta}_l(t) = \mathbf{B}_l'(t)\hat{\boldsymbol{\gamma}}_l = \sum_k \hat{\gamma}_{lk} B_{lk}(t)$. Take

$$\hat{\mathcal{M}}_{\nu_n} = \{ l : \ \omega_l = \frac{1}{|\mathcal{T}|} \int_{\mathcal{T}} \hat{\beta}_l(t)^2 dt \geq \nu_n \}$$

as the selected set, where $|\mathcal{T}|$ is the length of $\mathcal{T}$, and ν_n is a pre-specified threshold. To compute $\int_{\mathcal{T}} \hat{\beta}_l(t)^2 dt/|\mathcal{T}|$, we take N equally spaced time points $t_1 \leq \ldots \leq t_N$ in $\mathcal{T}$, and compute $\omega_{Nl} = (1/N) \sum_{i=1}^{N} \hat{\beta}_l(t_i)^2$. As long as N is large enough, ω_{Nl} can be used as ω_l. In our numerical study we let $N = 10,000$.

We correspondingly define the population version of the marginal least square regression,

$$u_l = u_l(\boldsymbol{\gamma}_l) = E(\mathbf{Y} - \mathbf{U}_l\boldsymbol{\gamma}_l)' \mathbf{W} (\mathbf{Y} - \mathbf{U}_l\boldsymbol{\gamma}_l). \tag{2.5}$$

It can be shown that the unique minimizer of $u_l(\boldsymbol{\gamma}_l)$ is

$$\tilde{\boldsymbol{\gamma}}_l = \Big(E\mathbf{U}_l'\mathbf{W}\mathbf{U}_l \Big)^{-1} E\mathbf{U}_l'\mathbf{W}\mathbf{Y}.$$

Let $\tilde{\beta}_l(t) = \mathbf{B}_l'(t)\tilde{\boldsymbol{\gamma}}_l = \sum_k \tilde{\gamma}_{lk} B_{lk}(t)$. It can be shown that $\tilde{\beta}_l(t)$ is the projection of $\beta_{l0}(t)$ onto the space $\mathcal{G}_l$, a linear space of spline functions on $\mathcal{T}$ with a fixed degree and knot sequence.

Let $\mathbf{X}_{li} = \mathrm{diag}(X_{li1}, \ldots, X_{lin_i})$, and take

$$\mathbf{B}_{li} = \begin{pmatrix} B_{l1}(t_{i1}) & \ldots & B_{l1}(t_{in_i}) \\ \vdots & \ddots & \vdots \\ B_{lK_l}(t_{i1}) & \ldots & B_{lK_l}(t_{in_i}) \end{pmatrix}.$$

It can be seen that $\mathbf{U}_{li} = \mathbf{X}_{li}\mathbf{B}_{li}'$. With some algebra, we can rewrite (2.4) as

$$\hat{\boldsymbol{\gamma}}_l = \Big(\sum_i \mathbf{B}_{li}\mathbf{X}_{li}\mathbf{W}_i\mathbf{X}_{li}\mathbf{B}_{li}' \Big)^{-1} \sum_i \mathbf{B}_{li}\mathbf{X}_{li}\mathbf{W}_i\mathbf{Y}_i.$$

When $n_i = 1$ for $i = 1, \ldots, n$, (2.3) boils down to the linear model. In this case, $\hat{\beta}_l(t)$ is the marginal correlation proposed in Fan and Lv (2008).

3. Theoretical Results

To establish the sure screening property, we decompose $\hat{\beta}_l(t) - \beta_{l0}(t) = \hat{\beta}_l(t) - \tilde{\beta}_l(t) + \tilde{\beta}_l(t) - \beta_{l0}(t)$ corresponding to the estimation error and the approximation error, respectively. Define $\omega = \max_i \omega_i$, $N = \max_i n_i$, $K_s = \min_l K_l$, $K_m = \max_l K_l$, and $\text{dist}(\beta_l, \mathbb{G}_l) = \inf_{g_l \in \mathbb{G}_l} \sup_{t \in \mathcal{T}} |\beta_l(t) - g_l(t)|$ as the L_∞ distance between $\beta_l(\cdot)$ and $\mathbb{G}_l$, where $\mathbb{G}_l$ is a linear space of spline functions on $\mathcal{T}$. Let $\rho_n = \max_l \text{dist}(\beta_{l0}, \mathbb{G}_l)$. The following conditions are needed.

A. The observation times $\{t_{ij}\}$, $j = 1, \ldots, n_i$, $i = 1, \ldots, n$, are chosen independently according to a distribution F_T on a finite interval $\mathcal{T}$. $L_1 \le |\mathcal{T}| \le L_2$. They are independent of the response and covariate processes $(Y_i(t), \mathbf{X}_i(t)), i = 1, \ldots, n$. The Lebesgue density $f_T(t)$ satisfies $M_1 \le f_T(t) \le M_2$ uniformly over $t \in \mathcal{T}$ for some positive constants M_1 and M_2.

B. There is a positive constant M_3 such that $|X_l(t)| \le M_3$ for $t \in \mathcal{T}$ and $l = 1, \ldots, p$.

C. $\min_{l \in \mathcal{M}_\star} (1/|\mathcal{T}|) \int_{\mathcal{T}} (EX_l(t)Y(t))^2 dt \ge c_1 n^{-2\kappa}$ for some $\kappa \in (0, 1/2)$.

A lemma shows that the minimum signal $\{\int_{\mathcal{T}} \tilde{\beta}_l(t)^2 dt/|\mathcal{T}|\}_{j \in \mathcal{M}_\star}$ is at the level of the integrated marginal correlation, provided the approximation error is negligible.

Lemma 1. *Under A–C, we have*

$$\min_{l \in \mathcal{M}_\star} \frac{1}{|\mathcal{T}|} \int_{\mathcal{T}} \tilde{\beta}_l(t)^2 dt \ge c_1 \xi n^{-2\kappa},$$

if $\rho_n^2 \le c_1 M_1 (1 - \xi) n^{-2\kappa} K_m M_2^{-1} L_2^{-1}$ *for some* $\xi \in (0, 1)$.

Now we establish the sure screening properties of the varying-coefficient independence screening (VIS). Let $\tilde{Y}_{ij} = \mathbf{X}_i(t_{ij})' \boldsymbol{\alpha}(t_{ij})$, $\tilde{\mathbf{Y}}_i = (\tilde{Y}_{i1}, \ldots, \tilde{Y}_{in_i})'$, $\tilde{\mathbf{Y}} = (\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_n)'$. We need additional conditions:

D. $\|\tilde{\mathbf{Y}}\|_\infty < B_1$ for some positive constant B_1, $\|\cdot\|_\infty$ the sup norm.

E. The random errors $\{\varepsilon_i(t)\}_{i=1}^n$ are i.i.d. with conditional mean zero and, for any $B_2 > 0$, there exists a positive constant B_3 such that $E[\exp(B_2|\varepsilon_i(t)|) \mid \mathbf{X}_i(t)] < B_3$, for $t \in \mathcal{T}$.

F. There exist a positive constant c_1 and a $\xi \in (0, 1)$ such that $\rho_n^2 \le c_1 M_1 (1 - \xi) n^{-2\kappa} K_m M_2^{-1} L_2^{-1}$.

Theorem 1. *If A-F hold, and* $\nu_n = c_6 n^{-2\kappa}$ *with* $c_6 \le c_1 \xi/2$, *then*

$$P(\mathcal{M}_\star \subset \widehat{\mathcal{M}}_{\nu_n}) \ge 1 - s_n K_m \Big\{ (8 + 2K_m) \exp\Big(-c_3 N^{-2} \omega^{-2} n^{1-4\kappa} K_m^{-3} \Big) + 6 c_5 K_m \exp\Big(-c_4 n K_m^{-1} \Big) \Big\}.$$

This implies we can handle the NP-dimensionality

$$\log p_n = o(N^{-2}\omega^{-2}n^{1-4\kappa}K_m^{-3} + nK_m^{-1}). \tag{3.1}$$

For p_n and n satisfying this condition, $P(\mathcal{M}_\star \subset \widehat{\mathcal{M}}_{\nu_n}) \to 1$. The maximal size of spline basis K_m, the maximal number of observational time points, and the maximal weight affect the order of dimensionality. In (3.1), the larger the minimum signal level, the smaller the number of basis functions, the smaller the weights, or the smaller the number of observational time points, the higher dimensionality the varying-coefficient independence screening (VIS) can handle. The approximation rate ρ_n also affects this dimensionality, through its relation with the choice of K_m as required in Condition F. Since the approximation error cannot be too large, the number of basis functions cannot be too small. When the β_l have bounded second derivatives and the number of observations for each subject is bounded, we have $\rho_n = O(K_m^{-2})$ (Schumaker (1981)), by taking $K_l = n^{1/5}$, the optimal rate for nonparametric regression (Stone (1985)), we have $\log p_n = o(n^{2/5})$. The second term on the right-hand side of (3.1) is improved compared with Fan, Feng and Song (2011).

Controlling false selection rates is also an important criteria. To achieve vanishing false selection rate, we bound the size of the selected set as follows.

Theorem 2. *Suppose A–F hold and* $\mathrm{Var}(\mathbf{Y}) = O(1)$. *Then for any* $\nu_n = c_6K_mn^{-2\kappa}$, *there exist positive constants* c_3, c_4, *and* c_5 *such that*

$$\begin{aligned}&P[|\widehat{\mathcal{M}}_{\nu_n}| \le O\{n^{2\kappa}\lambda_{\max}(\mathbf{\Sigma})\}]\\ &\ge 1 - p_nK_m\Big\{(8+2K_m)\exp\Big(-c_3N^{-2}\omega^{-2}n^{1-4\kappa}K_m^{-3}\Big) + 6c_5K_m\exp\Big(-c_4nK_m^{-1}\Big)\Big\},\end{aligned}$$

where $\mathbf{\Sigma} = E(\mathbf{UWU}')$.

Thus the correlation within the basis functions leads to dimension reduction with varying-coefficient models. When the number of observations for each subject and the weights are bounded, $K_m = 1$, and $\lambda_{\max}(\Sigma) = O(n^\tau)$, the number of selected variables is of order $O(n^{2\kappa+\tau})$. This is the same order as in Fan and Lv (2008) for the i.i.d. case.

4. Iterative VIS Procedure

As the independence screening procedure with marginal utilities uses only the marginal information of the covariates instead of the full model, its sure screening property may fail when the required technical conditions are not satisfied. Fan and Lv (2008) summarize potential problems for SIS with linear models. Similar problems are possible issues for the proposed screening methods as

well: A covariate that is jointly important but marginally unimportant to the response is not picked up by independent screening methods. Unimportant covariates that are highly correlated with the important covariates can have higher priority of selection by independent screening methods than important covariates that are relatively weakly related to the response. This issue does not affect sure screening, but it increases the false positive selection rates.

To address these issues while maintaining the computational expediency, Fan and Lv (2008) proposed iterative screening procedure to jointly employ a large-scale screening and moderate-scale selection strategy for linear models. We adapt the idea and propose an iterative screening procedure for VIS as follows.

1. Initial selection with marginal VIS and moderate-size variable selection. For every $l \in \{1, \cdots, p\}$, apply the independence VIS procedure to choose a set $\mathcal{A}_1$ of indices of size k_1, which can be taken as $\lfloor 2n/(3\log(n)) \rfloor$ to guarantee at least two iterations. Apply some existing penalized algorithm for grouped-variables selection, such as the group lasso in Yuan and Lin (2006), or the group SCAD in Wang, Chen and Li (2007), to the set $\mathcal{A}_1$ to select a subset $\mathcal{M}_1$. Inside the penalized method, the penalty parameter can be selected by Bayes information type of criterion or (generalized) cross validation.
2. Forward large-scale conditional marginal screening. For every $l \in \mathcal{M}_1^c = \{1, \cdots, p\} \backslash \mathcal{M}_1$, compute the conditional marginal least squares with the set of features $\mathcal{M}_1$ in the model.

$$\min \sum_{i=1}^{n} (\mathbf{Y}_i - \sum_{m \in \mathcal{M}_1} \mathbf{U}_{mi}\boldsymbol{\gamma}_m - \mathbf{U}_{li}\boldsymbol{\gamma}_l)' \mathbf{W}_i (\mathbf{Y}_i - \sum_{m \in \mathcal{M}_1} \mathbf{U}_{mi}\boldsymbol{\gamma}_m - \mathbf{U}_{li}\boldsymbol{\gamma}_l).$$

   This regression reflects the additional contribution of the lth covariate conditioning to $\mathcal{M}_1$. After marginally screening, as in the first step, pick a set $\mathcal{A}_2$ of indices of size $k_2 = 1$.
3. Backward moderate-size variable selection. Apply the penalized method used in the first step to the set $\mathcal{M}_1 \bigcup \mathcal{A}_2$ to select a subset $\mathcal{M}_2$.
4. Iteration until stabilization. Iterate Steps 2 and 3 until $|\mathcal{M}_l|$ is beyond a pre-specified number, or $\mathcal{M}_l = \mathcal{M}_{l-1}$.

In the next section, we discuss the performance of IVIS in numerical examples. From our simulation studies, IVIS seems to combine the strength of large scale screening and moderate-scale selection in both variable selection and prediction performance. The rationale is that false negatives can be effectively controlled in the initial screening step and the subsequent conditional screening steps, and the false positives can be effectively decreased in the penalized variable selection steps. Our numerical studies are consistent with this conjecture,

see Table 3. Iterations stabilize the whole process and can lead to good prediction performance of IVIS. Detailed theoretical analysis justification here is beyond the present scope and is to be considered in future research.

5. Numerical Examples

5.1. Simulations

We used three simulation models to examine the finite-sample performance of IVIS and VIS+gSCAD. We fixed the sample at 200 and the dimension at 500 in all examples. For each model we ran 100 independent replicates.

Simulation model 1, Wei, Huang and Li (2011). The response variable was generated as $y_i(t_{ij}) = \sum_{l=1}^{p} x_{li}(t_{ij})\beta_l(t_{ij}) + \epsilon_i(t_{ij})$. The time points t_{ij} were taken from $\{1, 2, 3, \ldots, 30\}$ with probability 0.4; The number of observed time points n_i for different subjects are different. Only the first six variables had nonzero coefficient functions. The coefficient functions were

$$\beta_1(t) = 15 + 20\sin(\frac{\pi t}{15}),\ \beta_2(t) = 15 + 20\cos(\frac{\pi t}{15}),$$
$$\beta_3(t) = 2 - 3\sin(\frac{\pi(t-25)}{15}),\ \beta_4(t) = 2 - 3\cos(\frac{\pi(t-25)}{15}),$$
$$\beta_5(t) = 6 - 0.2t^2,\ \beta_6(t) = -4 + \frac{(20-t)^3}{2,000}.$$

The variables were

$$x_1(t) \sim \text{Unif}[\frac{t}{10}, 2 + \frac{t}{10}],\ \{x_l(t)\}_{l=2}^{5} \sim N(0, \frac{1 + x_1(t)}{2 + x_1(t)}),$$
$$x_6(t) \sim N(3\exp(\frac{t}{30}), 1),\ \{x_l(t)\}_{l=7}^{500} \sim \text{MVN}(\mathbf{0}, \Sigma),$$

where $\Sigma_{t,s} = \text{Cov}\,(x_l(t), x_l(s)) = 4\exp(-|t-s|)$. The random error $\epsilon(t) = Z(t) + E(t)$, where $Z(t)$ had the same distribution as $\{x_l(t)\}_{l=7}^{500}$ and $E(t)$ is $N(0, 4)$.

Simulation model 2. The response variable was again
$y_i(t_{ij}) = \sum_{l=1}^{p} x_{li}(t_{ij})\beta_l(t_{ij}) + \epsilon_i(t_{ij})$, where time points were taken from $\{1, 2, \ldots, 30\}$ with probability 0.5. The variables $(x_1, x_2, \ldots, x_p)$ were simulated as

$$x_l = \frac{W_l + U}{2},\ l = 1, \ldots, p,$$

where $W_1, W_2, \ldots, W_p$ and U were i.i.d. $Unif(0, 1)$. The random error $\epsilon \sim$

Table 1. Variable selection performance of IVIS and VIS+gSCAD.

	IVIS					VIS+gSCAD				
	SA	ES	MS	OS1	OS2+	SA	ES	MS	OS1	OS2+
Model 1	100%	99%	0%	0%	1%	88%	88%	12%	0%	0%
Model 2	100%	78%	0%	17%	5%	0%	0%	100%	0%	0%
Model 3	100%	88%	0%	11%	1%	100%	100%	0%	0%	0%

$N(0,1)$. The coefficient functions were

$$\beta_1(t) = 7\cos^2(\frac{t-10}{7}) + 0.1t,\ \beta_2(t) = -0.5t,$$
$$\beta_3(t) = \frac{(t-15)^2}{20},\ \beta_4(t) = 15\sin(\frac{t+5}{3.5})\exp(-\frac{t}{30}),$$
$$\beta_l(t) = 0,\ l \geq 5.$$

Simulation model 3: The response variable was as before except that the time points were taken from $\{1, 2, \ldots, 30\}$ with probability 0.3. Only the first six coefficient functions were nonzero:

$$\beta_1 = \beta_3 = \beta_5 = 1 \text{ and } \beta_2 = \beta_4 = \beta_6 = -1.$$

Here $\{x_k(t)\}_{l=1}^{450}$ were i.i.d Gaussian processes with mean zero and variance one and

$$x_l(t) = \sum_{j=1}^{6} \frac{x_j(t)(-1)^{(j+1)}}{5} + \sqrt{1-\frac{6}{25}}\epsilon_l(t),\ k = 451, \ldots, 500,$$

where $\{\epsilon_l(t)\}_{k=451}^{500}$ were Gaussian with mean zero and variance three. The random error $\epsilon(t) \sim N(0,1)$. This model is a parametric linear model in order to examine whether nonparametric screening and estimation does much worse than parametric screening and estimation.

As shown in Table 1, we used several quantities to measure the variable selection performance: "SA" is the percentage of occasions on which all the correct variables are included in the selected model; "ES" is the frequency of exactly selecting all true variables and nothing else; "MS" is the percentage of occasions on which some correct variables are missed; "OS1" is the frequency of exactly one false variable selected and "OS2" is the frequency of selecting 2 or more false variables. We see that VIS+gSCAD tends to be too greedy in Models 1 and 2, missing some true variables, but IVIS always selects all true variables. A 0% for "MS" indicates extremely low false negative rates for all three models using IVIS. Small values for "OS1" and "OS2+" show low false positive rates of variable selection. Overall, IVIS has a very good variable selection performance.

Table 2. The number of iterations needed to achieve stabilization in IVIS.

Iterations	2	3	4	5	>5
Model 1	43	40	16	0	1
Model 2	1	23	69	7	0
Model 3	88	11	0	1	0

Computation time depends heavily on the implementation of the group SCAD algorithm, as the screening process is pretty fast. So if IVIS can achieve stability within a few iterations, the computation is reasonably fast. Here we report the number of iterations needed to achieve stability. Two is the minimum number of iterations needed to confirm that variables selected in the current iteration match the previous selection, and we observe that stability is reached almost always within 5 iterations. In Model 3 with six constant coefficients, IVIS converges within 3 steps 99% of the time.

In Table 3 we record the average numbers of false positives and false negatives in the first 4 iterations for the three models. In our simulations of these different models, after the first step (VIS+groupSCAD) usually the majority of true variables were selected with occasionally, a few false variables. As the number of iteration increased, the number of false positives decreased after the screen step, the number of false negatives increased after the gSCAD step, and stability was achieved quickly. When false variables were selected, the accuracy of coefficient estimation for the true variables was not compromised.

Two quantities are used to measure the estimation accuracy of IVIS in Table 4. For each coefficient function estimator $\hat{\beta}_j(t)$, the integrated mean squared error (IMSE) is $\int(\hat{\beta}_j(t)-\beta_j(t))^2dt$. This can be computed by numeric integration. We also report the relative IMSE (RIMSE) which is the ratio of the IMSE of an estimator relative to the IMSE of the oracle estimator, that knows the true variables and only needs to estimate the true coefficient functions. In Models 1 and 2, the oracle estimator uses 5 and 10 B-spline basis functions to estimate each true coefficient function, as does IVIS. The RIMSE of IVIS is close to 1 in these two models, which is expected given the variable selection results in Table 1. In Model 3 the RIMSE of IVIS is larger than 2, although IVIS still does good variable selection. This disparity can be explained by the fact that in Model 3 we actually allow the oracle to use the knowledge that the true coefficient functions are constant so that the oracle estimator directly estimates these constants, and not using 5 B-splines basis functions.

In Table 5 we compare the prediction accuracy of the oracle estimator, IVIS, and VIS+gSCAD. The prediction errors were computed on an independent test dataset. We see that IVIS and the oracle have nearly identical prediction performance in all three models. In Models 1 and 3, VIS+gSCAD performs very

Table 3. The average false positive and false negative rates with standard error in prentases, after screening and after gSCAD for the first four iterations for Models 1, 2, and 3.

	After Screening		After gSCAD	
	false negative	false positive	false negative	false positive
Model 1:				
Iteration 1	0.13	19.13	0.13	0.01
	(0.034)	(0.034)	(0.034)	(0.010)
Iteration 2	0.00	0.88	0.00	0.43
	(0.000)	(0.036)	(0.000)	(0.050)
Iteration 3	0.00	1.43	0.00	0.18
	(0.000)	(0.050)	(0.000)	(0.052)
Iteration 4	0.00	1.54	0.00	0.08
	(0.000)	(0.061)	(0.000)	(0.042)
Model 2:				
Iteration 1	1.96	22.96	1.96	1.13
	(0.049)	(0.037)	(0.037)	(0.103)
Iteration 2	0.96	1.13	0.96	0.90
	(0.037)	(0.103)	(0.037)	(0.033)
Iteration 3	0.05	0.99	0.05	0.06
	(0.022)	(0.017)	(0.022)	(0.028)
Iteration 4	0.00	1.01	0.00	0.03
	(0.103)	(0.017)	(0.000)	(0.017)
Model 3:				
Iteration 1	0.00	19.00	0.00	0.00
	(0.000)	(0.000)	(0.000)	(0.000)
Iteration 2	0.00	1.00	0.00	0.10
	(0.000)	(0.000)	(0.000)	(0.030)
Iteration 3	0.00	1.10	0.00	0.10
	(0.000)	(0.030)	(0.000)	(0.030)
Iteration 4	0.00	1.10	0.00	0.10
	(0.000)	(0.030)	(0.000)	(0.030)

similarly to IVIS and the oracle estimator but has significantly worse prediction in Model 2. This is consistent with its unsatisfactory variable selection performance in Table 1.

In Figure 1−3 we depict the estimated coefficient functions by IVIS compared to the ground truth.

5.2. Data

The experiment of Spellman et al. (1998) recorded genome-wide mRNA levels for 6178 yeast ORFs (open reading frames) simultaneously over approximately two cell cycle periods at 7-minutes intervals for 119 minutes with a total of 18 time points. The cell cycle is an ordered set of events and the cell cycle process is

Table 4. IMSE and relative IMSE for estimating true β's.

	β_1	β_2	β_3	β_4	β_5	β_6
Model 1 (IMSE)	3.29	22.41	1.83	0.69	0.76	0.47
	(1.18)	(0.71)	(0.47)	(0.44)	(0.44)	(0.26)
Model 1 (RIMSE)	1.14	1.00	1.00	1.02	1.01	1.27
	(0.27)	(0.00)	(0.01)	(0.18)	(0.05)	(0.54)
Model 2 (IMSE)	4.27	3.75	3.84	4.24	NA	NA
	(2.29)	(1.70)	(2.28)	(1.69)	NA	NA
Model 2 (RIMSE)	1.19	1.10	1.10	1.05	NA	NA
	(0.64)	(0.35)	(0.39)	(0.18)	NA	NA
Model 3 (IMSE)	0.16	0.21	0.25	0.23	0.17	0.16
	(0.22)	(0.27)	(0.43)	(0.34)	(0.22)	(0.23)
Model 3 (RIMSE)	2.17	2.61	3.14	2.94	2.22	1.96
	(3.15)	(4.84)	(5.54)	(5.24)	(3.70)	(3.69)

Table 5. Prediction error comparison.

	Oracle	IVIS	VIS+gSCAD
Mode 1	8.93	8.95	9.75
	(0.28)	(0.29)	(2.20)
Model 2	1.04	1.05	3.58
	(0.03)	(0.03)	(0.71)
Model 3	1.01	1.04	1.04
	(0.03)	(0.04)	(0.04)

commonly divided into G1-S-G2-M stages, where the G1 stage stands for "GAP 1", the S stage stands for "Synthesis" during which DNA replication occurs, the G2 stage stands for "GAP 2", and the M stage stands for "mitosis" during which nuclear and cytoplasmic division occur. The experiment identified approximately 800 genes that vary in a periodic fashion during the yeast cell cycle; little was known about the regulation of most of these genes. Transcription factors (TFs) play critical roles in gene expression regulation. A transcription factor is a protein that binds to specific DNA sequences, thereby controlling the flow of genetic information from DNA to mRNA.

We applied our IVIS method to investigate the transcription factors (TFs) involved in the yeast cell cycle. We considered 240 genes without missing values, and there were 96 transcriptional factors with at least one nonzero binding probability. Let $y_i(t_j)$ denote the log-expression level for gene i at time point t_j during the cell cycle process; the chromatin immunoprecipitation (ChIP-chip) data of Lee et al. (2002) was used to derive the binding probabilities. This dataset has been analyzed by Wang, Li and Huang (2008) and Wei, Huang and Li (2011) who used a varying coefficient model to link the binding probabilities to the log-gene expression levels:

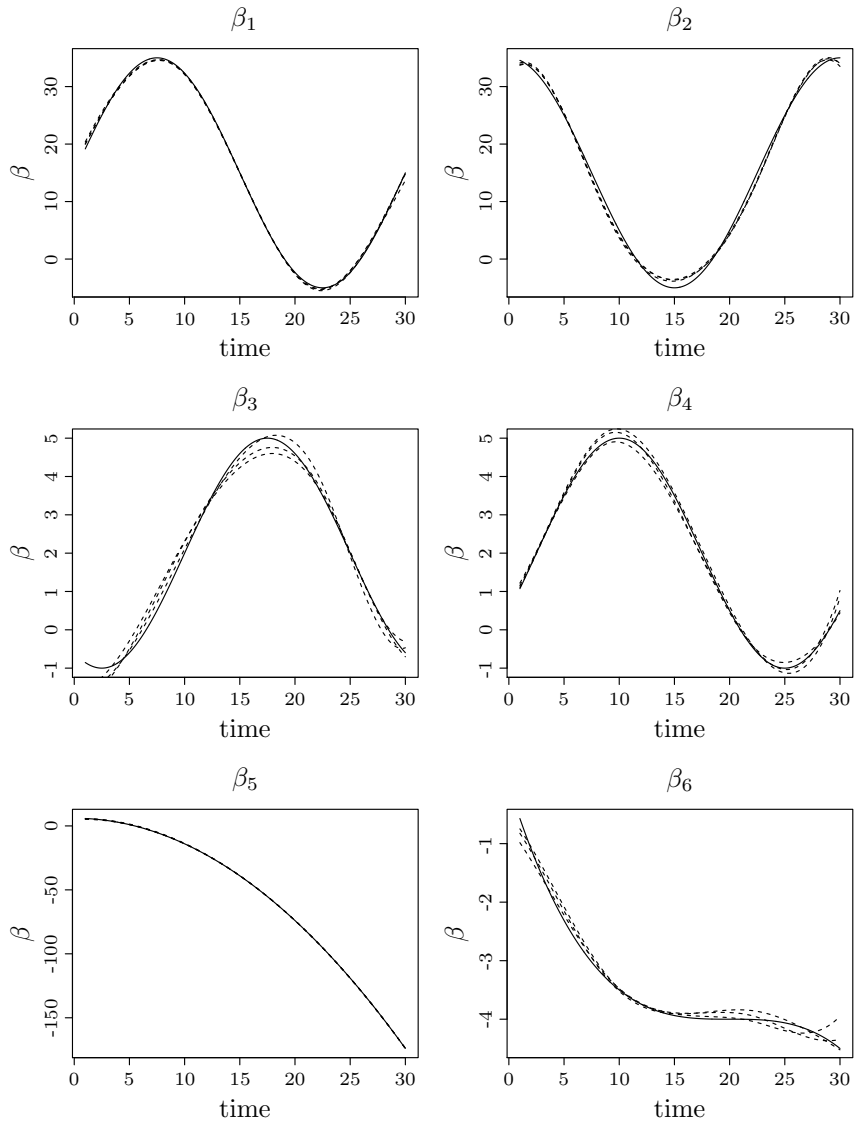

Figure 1. Model 1: the line is the true β curve, while the dashed lines are estimated curves by IVIS in three runs.

$$y_i(t_j) = \mu(t_j) + \sum_{l=1}^{96} x_{i,l}\beta_l(t_j) + \epsilon_i(t_j).$$

This dataset has a moderately high dimension, $p = 96$ with $n = 240$. We first used gSCAD to obtain a sparse estimator of the varying coefficient model. Figure 1 in the Web Appendix, shows the estimated β curves over time for 21 known

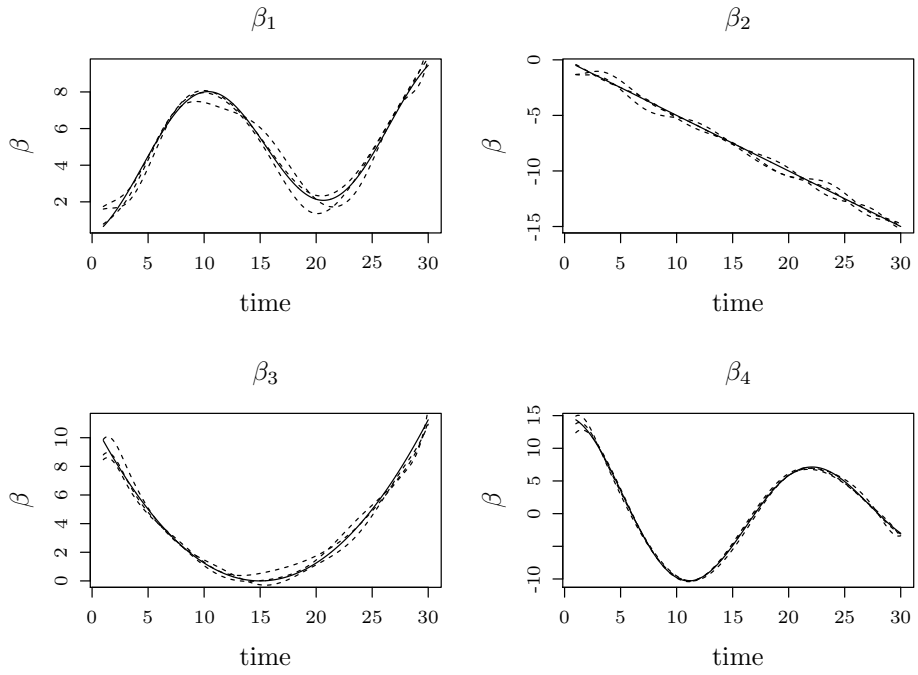

Figure 2. Model 2: the line is the true β curve, while the dashed lines are estimated curves by IVIS in three runs.

yeast TFs.

In order to demonstrate the performance of IVIS in a high dimensional case, we added extra 384 pure noise variables to the original data to have a total 480 variables. We can test IVIS in the high-dimensional setting, as the total number of variables is double the number of subjects. The 384 noise variables for each subject were independently sampled from the standard normal distribution. We applied IVIS to the augmented dataset and repeated the process 100 times. Among the 21 known important TFs, IVIS on average identified 14 TFs with standard deviation 0.84. Figure 2 in Web Appendix shows the estimated β curves of 14 TFs identified by IVIS in one trial. Although the curves are not the same as those in Figure 1 in the Web Appendix, similar patterns are shown for most of the 14 TFs. We compare the estimated transcriptional effects side-by-side for 5 TFs in Figure 4.

We also compared the prediction error of IVIS with estimation for the full model without variable selection. Five-fold cross validation was used to calculate prediction error. We ran 100 replicates for each method. In Table 6, we record the prediction error of SCAD, IVIS, and that of no variable selection with and without adding 384 noise variables. Here IVIS significantly outperforms the

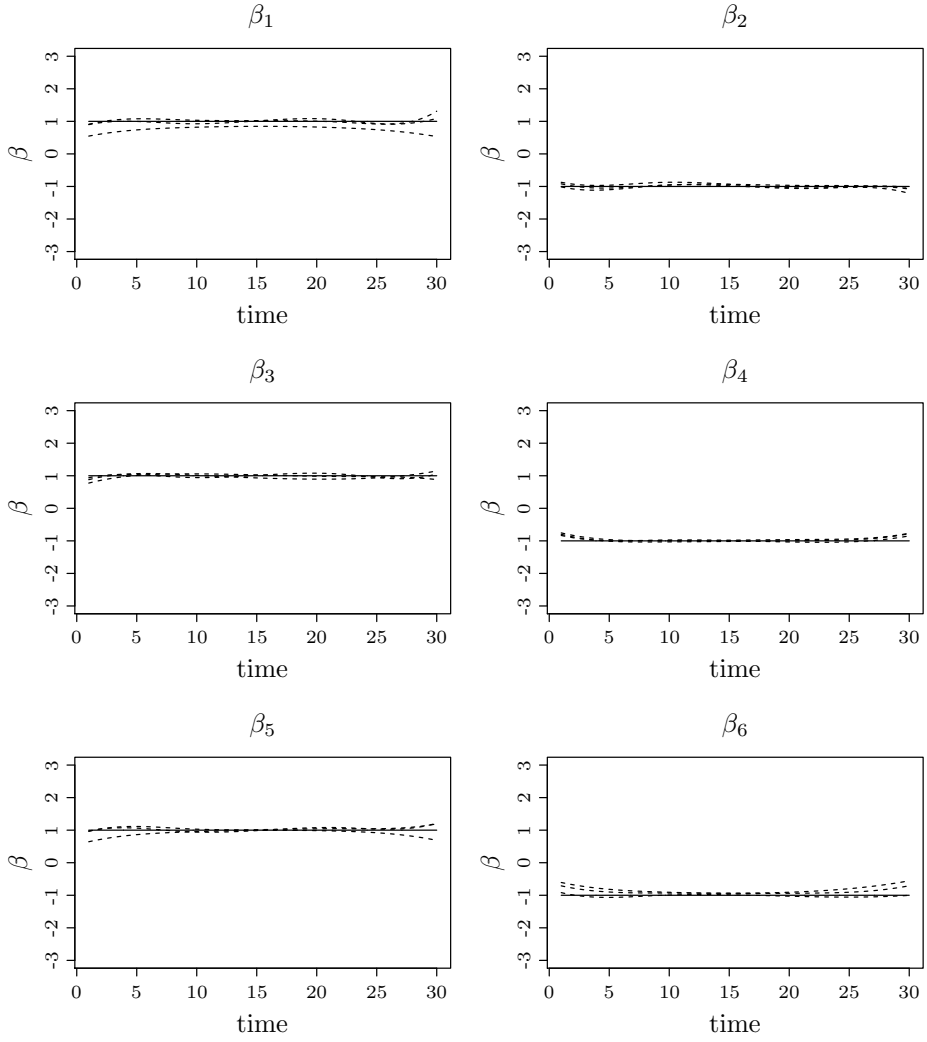

Figure 3. Model 3: the line is the true β curve, while the dashed lines are estimated curves by IVIS in three runs.

estimation without variable selection in terms of prediction when noise variables are added. The performance of IVIS in the high dimensional setting is close to the SCAD in much lower dimensional setting. The prediction error of IVIS is also much lower than that of the full model without variable selection and without noise variables. This suggests the prediction power of IVIS in data with very high dimensional covariates. The R codes for the data analysis are available upon request.

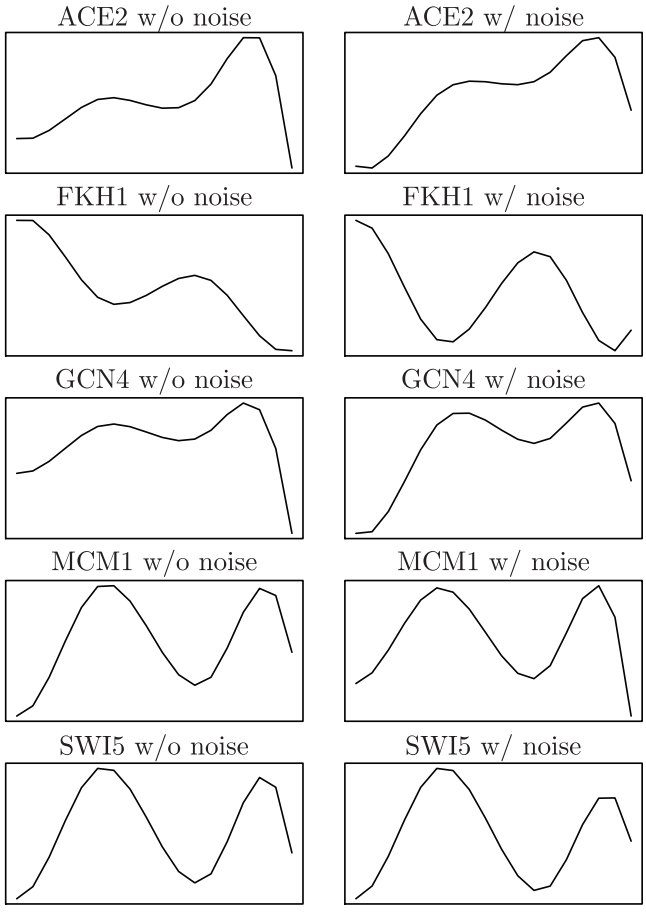

Figure 4. Comparison of estimated time-varying transcriptional effects for 5 TFs identified by SCAD w/o noise (left column) and IVIS w/ 384 noise (right column).

Table 6. Prediction error comparison.

	w/o noise variables		w/ 384 noise variables	
	SCAD	no variable selection	IVIS	no variable selection
prediction error	0.225	0.507	0.294	0.782
(standard deviation)	(0.004)	(0.019)	(0.017)	(0.037)

6. Conclusion

We have studied VIS for variable screening in varying-coefficient models and established its sure screening property. We have proposed IVIS for fitting varying-coefficient models in ultra-high dimensions by iterating between a greedy conditional VIS step and a gSCAD penalized fitting step. The proposed methodology has been supported by numeric examples.

As to VIS and the nonparametric independence screening (NIS) in Fan, Feng and Song (2011), both methods are flexible extensions of the marginal correlation ranking idea in Fan and Lv (2008), and both methods use B-splines to compute their marginal ranking statistics. The marginal ranking statistics is the fundamental quantity in a marginal screening method. The two methods use different marginal statistics as they are designed for different data structures. NIS uses the marginal correlation of the response variable and the estimated marginal nonparametric regression functions. VIS uses $(1/|\mathcal{T}|)\int_{\mathcal{T}}\hat{\beta}_j(t)^2dt$ to rank the jth covariate, where $\hat{\beta}_j(t)$ is the estimated marginal coefficient function of the jth covariate; this can be viewed as the integrated marginal correlation of the time-varying response variable and the jth time-varying covariate projected onto the B-Spline space. Efforts were taken in VIS to analyze the influence of longitudinal observations on the dimensionality that VIS can handle.

Acknowledgements

Rui Song's research was partially supported by National Science Foundation grant DMS-1007698 and National Cancer Institute P01 CA142538. Hui Zou's research was partially supported by NSF grant DMS-0846068. We thank the Editor, an associate editor and two reviewers for their time and constructive comments that helped us to improve the article.

References

Bickel, P., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37**, 1705-1732.

Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n (with discussion). *Ann. Statist.* **35**, 2313-404.

Fan, J., Feng, Y. and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *J. Amer. Statist. Assoc.* **106**, 544-557.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. Roy. Statist. Soc. Ser. B* **70**, 849-911.

Fan, J., Samworth, R. and Wu, Y. (2009). Ultra-dimensional variable selection via independent learning: beyond the linear model. *J. Machine Learn. Res.* **10**, 1829-1853.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with np-dimensionality. *Ann. Statist.* **38**, 3657–3604.

Hoover, D. R., Rice, J. A., Wu, C. O. and Yang, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85**, 809-822.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K, Hannett, N. M., Harbison, C. T., Thompson, C. M. and Simon, I., *et al.* (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* **298**, 799-804.

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

Li, R., Zhong, W. and Zhu, L.-P. (2012). Feature screening via distance correlation learning. *J. Amer. Statist. Assoc.*. In press.

Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for highdimensional data. *Ann. Statist.* **37**, 246-270.

Rice, J. A. (2004). Functional and longitudinal data analysis: perspectives on smoothing. *Statist. Sinica* **14**, 631-647.

Schumaker, L. L. (1981). *Spline Functions: Basic Theory*. Wiley, New York.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. and Futcher, B. (1998). Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. *Molecular Biology of Cell* **9**, 3273-3297.

Stone, C. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13**, 689-705.

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

van de Geer, S. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist.* **36**, 614-645.

Wang, L., Chen, G. and Li, H. (2007). Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics* **23**, 1486-1494.

Wang, H. and Xia, Y. (2008). Shrinkage estimation of the varying coecient model. *J. Amer. Statist. Assoc.* **104**, 747-757.

Wang, L., Li, H. and Huang, J. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *J. Amer. Statist. Assoc.* **103**, 1556-1569.

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

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. Ser. B* **68**, 49-67.

Zhu, L.-P., Li, L., Li, R. and Zhu, L.-X. (2011). Model-free feature screening for ultrahigh dimensional data. *J. Amer. Statist. Assoc.* **106**, 1464-1475.

Department of Statistics, North Carolina State University, 2311 Stinson Drive, Campus Box 8203, Raleigh, NC 27695-8203, USA.

E-mail: rsong@ncsu.edu

School of Statistics, University of Minnesota, 313 Ford Hall, School of Statistics, 224 Church Street S.E., Minneapolis, MN, 55455, USA.

E-mail: fengyi@stat.umn.edu

School of Statistics, University of Minnesota, 313 Ford Hall, School of Statistics, 224 Church Street S.E., Minneapolis, MN, 55455, USA.

E-mail: zouxx019@umn.edu

(Received October 2012; accepted December 2013)