Statistica Sinica Preprint No: SS-2017-0083	
Title	Forward Additive Regression for Ultrahigh Dimensional Nonparametric Additive Models
Manuscript ID	SS-2017-0083
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202017.0083
Complete List of Authors	Wei Zhong Sunpeng Duan and Liping Zhu
Corresponding Author	Liping Zhu
E-mail	zhulp1@hotmail.com

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0083

Forward Additive Regression for Ultrahigh-Dimensional Nonparametric Additive Models

Wei Zhong[1], Sunpeng Duan[1] and Liping Zhu[2]

Xiamen University[1] *and Renmin University of China*[2]

Abstract: Ultrahigh-dimensional data are collected in many scientific fields where the predictor dimension is often much higher than the sample size. To effectively reduce the ultrahigh-dimensionality, many marginal screening approaches have been developed. However, existing screening methods may miss important predictors that are marginally independent of the response, or may select unimportant predictors owing to their high correlations with important predictors. Iterative screening procedures have been proposed to address this issue. However, studying their theoretical properties is not straightforward. Penalized regressions are not computationally efficient or numerically stable when the predictors are ultrahigh-dimensional. To overcome these drawbacks, a forward regression approach has been developed for linear models. However, nonlinear dependence between predictors and the response is often present in ultrahigh-dimensional problems. In this study, we extend the FR to develop a forward additive regression (FAR) method for selecting significant predictors in ultrahigh-dimensional nonparametric additive models. We establish the screening consistency for the FAR method and examine its finite-sample performance using Monte Carlo sim-

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0083

ulations. Our simulations indicate that, compared with marginal screenings, the FAR is much more effective in terms of identifying important predictors for additive models. When the predictors are highly correlated, the FAR even outperforms iterative marginal screenings, such as the iterative nonparametric independence screening. We also apply the FAR method to a real-data analysis in genetic studies.

Key words and phrases: Additive models, forward regression, screening consistency, ultrahigh-dimensionality, variable selection.

1. Introduction

Advances in modern information technology allow researchers in various scientific fields to collect high-dimensional data, where the number of predictors is greater than the sample size. Under the sparsity assumption that only a small subset of predictors truly contribute to the response, penalized regression methods have been studied intensively for various parametric and nonparametric models. These methods include, but are not limited to, the LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), elastic net (Zou and Hastie, 2005), adaptive LASSO (Zou, 2006; Huang, Ma, and Zhang, 2008), grouped LASSO (Yuan and Lin, 2006), and Dantzig selector (Candes and Tao, 2007). These methods are able to select significant variables and estimate parameters simultaneously, enhancing both model interpretability and predictability.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0083

When the predictor dimension is much greater than the sample size, the aforementioned penalized approaches may suffer from computational complexity, algorithmic instability, or statistical inaccuracy (Fan, Samworth, and Wu, 2009). Since the seminal work of Fan and Lv (2008), various marginal screening procedures have been proposed to reduce ultrahigh-dimensionality. Screening ranks all predictors using a marginal utility that measures the importance of each predictor. For example, Fan and Lv (2008) developed a sure independence screening (SIS) using a Pearson correlation ranking procedure for Gaussian linear regressions. Hall and Miller (2009) considered a generalized correlation based on polynomial transformations of the predictors. For additional examples, see Fan and Song (2010) for generalized linear models, Zhu et al. (2011) for multi-index models, Fan, Feng, and Song (2011) for nonparametric additive models, He, Wang, and Hong (2013) for heterogeneous nonparametric models, and Liu, Li, and Wu (2014) and Fan, Ma, and Dai (2014) for varying-coefficient models, among others. Without imposing a specific regression model structure, some dependence/independence measures have been used as marginal utilities to develop model-free variable screenings. These measures include the distance correlation (Li, Zhong, and Zhu, 2012), Kendall's τ-rank correlation (Li et al., 2012), Kolmogorov−Smirnov test statistic (Mai and Zou, 2013), mar-

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0083

tingale difference correlation (Shao and Zhang, 2014), Pearson Chi-square test statistic (Huang, Li, and Wang, 2014), and mean variance index (Cui, Li, and Zhong, 2015).

Despite being computationally efficient and possessing the sure-screening property, existing marginal screening methods may fail to detect predictors that truly contribute to, but are marginally independent of the response variable. Another problem is that marginal methods tend to select some unimportant predictors, owing to their high correlations with the important predictors. To overcome these drawbacks, iterative marginal variable screening procedures have been developed. For example, the iterative sure-independence screening (ISIS) proposed by Fan and Lv (2008) is conducted in the following way. In the first step, we select an initial set of predictors using the SIS, and then regress the response over the selected predictors. In the second step, we treat the residuals as the new responses, and then apply the SIS again for the remaining predictors to obtain another subset. The procedure is performed iteratively and the union of the selected subsets is the final set of predictors. Additional examples can be found in the aforementioned references. However, iterative screening methods lack necessary theoretical justifications. Another alternative solution to the problems with marginal variable screenings is to use a forward regression (FR).

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0083

In an important work, Wang (2009) developed an FR for variable screening in ultrahigh linear regression models. Wang (2009) also demonstrated theoretically and numerically that the FR is able to identify all relevant predictors consistently. Cheng, Honda, and Zhang (2016) further extended the FR to ultrahigh-dimensional varying-coefficient models. Cheng et al. (2015) proposed a groupwise FR for linear models that incorporates multiple predictors in each step.

It is well known that nonlinear dependence between predictors and the response variable is often present in ultrahigh-dimensional data. In this case, traditional linear models may be not adequate to fit the data. On the other hand, fully nonparametric models may suffer from the "curse of dimensionality" problem. In this study, we consider a nonparametric additive model for ultrahigh-dimensional data. This approach increases the flexibility of ordinary linear models and allows a nonlinear transformation of each predictor to be added to the regression model, where the unknown transformed functions are estimated in a nonparametric manner. In the literature, penalized regression methods have been well studied for nonparametric additive models. See Lin and Zhang (2006), Meier, Geer, and Bühlmann (2009) and Huang, Horowitz, and Wei (2010). For sparse ultrahigh-dimensional additive models, Fan, Feng, and Song (2011)

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0083

designed a nonparametric independence screening (NIS) that fits marginal nonparametric regressions of the response against each predictor individually, and then ranks their importance according to the magnitudes of the estimated nonparametric components. The iterative version of the NIS (INIS) was also introduced to remedy the aforementioned drawbacks.

In this study, motivated by the appealing theoretical properties and outstanding numerical performance of the FR of Wang (2009), we propose a forward additive regression (FAR) procedure for ultrahigh-dimensional nonparametric additive models. The FAR procedure works as follows. In the first step, we fit the marginal regression models using B-spline smoothing, compute the residual sum squares (RSS) for each model, and select the predictor that corresponds to the minimum RSS. This step is identical to the marginal NIS procedure. In the second step, we keep the selected predictor in the model, and then add a new one from the remaining predictors, one at a time. Next, we fit the augmented models and then add the predictor with the minimum RSS to the selected subset. We repeat the second step until a certain stopping rule is reached. The FAR enjoys several advantages from both theoretical and practical viewpoints. First, we rigorously establish the screening consistency for the FAR method under some mild conditions. This justifies that the model selected by the FAR

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0083

contains the truly important set of predictors with probability approaching one. Note that the FAR method is essentially a special case of the INIS procedure of Fan, Feng, and Song (2011) when the INIS adds one predictor per step. The main contribution of this work is that the theorems of the FAR remedy the lack of a theoretical justification for the INIS for nonparametric additive models. Second, the FAR addresses the drawbacks of marginal variable screenings. Specifically, the FAR selects important covariates that are marginally independent of the response, and prevents adding unimportant covariates that may be selected by marginal methods, owing to their high correlations with the important variables. Third, the implementation is easy and the sequential procedure provides a clear solution path. This path is straightforward to interpret in the sense that the importance of the predictors can be ranked according to the selection order.

The rest of this article is organized as follows. In Section 2, we develop the FAR procedure. Section 3 derives the screening consistency of the FAR algorithm. Simulation studies and a real-data analysis are presented in Section 4. Section 5 concludes the paper.

2. FAR

In this section, we introduce the model setup for the FAR approach for ultrahigh-dimensional nonparametric additive models and present the

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0083

details of the FAR algorithm.

2.1 Model Setup

Let $\{(\mathbf{x}_i, Y_i), i = 1, \ldots, n\}$ be a random sample of size n from the population $(\mathbf{x}, Y)$, where $Y_i \in \mathbb{R}^1$ is the response variable and $\mathbf{x}_i = (X_{i1}, \cdots, X_{ip})^{\mathrm{T}} \in \mathbb{R}^p$ is the predictor vector, with $p \gg n$, for the ith observation. Without loss of generality, we assume that the mean of the response is zero. In practice, we can centralize the response first. To study the relationship between the predictors and the response, we assume that the observations satisfy the following nonparametric additive model:

$$Y_i = \sum_{j=1}^{p} f_j(X_{ij}) + \varepsilon_i, \tag{1}$$

where f_j denotes an unknown function and ε_i is an unobserved random variable with mean zero and finite variance σ^2. For identifiability, we assume that all $f_j(\cdot)$ are centered (i.e., $E\{f_j(X_j)\} = 0, 1 \le j \le p$). We refer to X_j as a relevant/important predictor if $E\{f_j(X_j)^2\} > 0$, and an irrelevant/unimportant predictor if $E\{f_j(X_j)^2\} = 0$. Let $\mathcal{A} = \{1, \cdots, p\}$ be the full model and $\mathcal{T} = \{j : E\{f_j(X_j)^2\} > 0\}$ be the true model that contains all relevant predictors. Under the sparsity assumption, we assume that p_0 predictors can truly contribute to the response, and that the model

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0083

size $|\mathcal{T}| = p_0 \ll p$. For convenience, we also use the generic notation $\mathcal{M} = \{j_1, \cdots, j_{d^*}\} \subseteq \mathcal{A}$ to denote an arbitrary model corresponding to $X_{j_1}, \cdots, X_{j_{d^*}}$.

To estimate the nonparametric components, we use B-spline basis. Let $\mathcal{S}_n$ be the space of polynomial splines of degree $l \geq 1$, and let $\{\psi_{jk}, k = 1, \cdots, m_n\}$ denote a normalized B-spline basis for the jth predictor with $\|\psi_{jk}\|_\infty \leq 1$, where $\|\cdot\|_\infty$ is the sup norm and m_n is the sum of the polynomial degree and the number of knots used to create the B-spline basis. In theory, we may choose $m_n = O(n^{1/(2d+1)})$, as per Stone (1985) and Huang, Horowitz, and Wei (2010), which allows m_n to increase at a relatively slow rate with the sample size, where $d > 1$ is specified in Section 3. We can represent any $f_{nj} \in \mathcal{S}_n$ by the linear combination of normalized B-spline basis functions. That is,

$$f_{nj}(x) = \sum_{k=1}^{m_n} \gamma_{jk}\psi_{jk}(x), \text{ for } 1 \leq j \leq p. \tag{2}$$

Thus,

$$Y_i = \sum_{j=1}^{p}\sum_{k=1}^{m_n} \gamma_{jk}\psi_{jk}(X_{ij}) + \xi_i, \tag{3}$$

where $\xi_i = \sum_{j=1}^{p}\{f_j(X_{ij}) - f_{nj}(X_{ij})\} + \varepsilon_i$. Here, we implicitly assume that

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0083

$f_j(X_{ij})$ can be well approximated by $f_{nj}(X_{ij}) \in \mathcal{S}_n$ by choosing some suitable coefficients $\{\gamma_{j1}, \ldots, \gamma_{jm_n}\}$ under some smoothness conditions (Stone, 1985). Specifically, Huang, Horowitz, and Wei (2010) showed that $\|f_n - f\|_2 = O_p(m_n^{-d})$, where $\|\cdot\|_2$ is the L_2-norm. When p is fixed and small, the ordinary least squares estimators can be obtained for (3). When p is moderately high, penalized regression methods with grouped penalties have been well studied for (3), for example, by Lin and Zhang (2006), Meier, Geer, and Bühlmann (2009), and Huang, Horowitz, and Wei (2010). When p is much higher than the sample size, Fan, Feng, and Song (2011) proposed an NIS that reduces the dimensionality efficiently. In the next subsection, we will propose a new FAR algorithm to select important variables for (3).

First, we introduce some notation. For simplicity, we write $\psi_k(X_{ij}) = \psi_{jk}(X_{ij})$. Let $U_{ij} = \{\psi_1(X_{ij}), \cdots, \psi_{m_n}(X_{ij})\}^{\mathrm{T}} \in \mathbb{R}^{m_n}$, such that U_{ij} consists of values of the centered basis functions for the ith observation of the jth predictor. Let $\mathbf{U}_j = (U_{1j}, \cdots, U_{nj})^{\mathrm{T}} \in \mathbb{R}^{n \times m_n}$ be the "design" matrix corresponding to the jth predictor. Hence, the total "design" matrix is $\mathbf{U} = (\mathbf{U}_1, \cdots, \mathbf{U}_p) \in \mathbb{R}^{n \times pm_n}$. Moreover, define $\boldsymbol{\gamma}_j = (\gamma_{j1}, \cdots, \gamma_{jm_n})^{\mathrm{T}}$ as an $m_n \times 1$ vector of parameters corresponding to the jth predictor in the model, and denote $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^{\mathrm{T}}, \cdots, \boldsymbol{\gamma}_p^{\mathrm{T}})^{\mathrm{T}}$ as a $pm_n \times 1$ vector of parameters. For an arbitrary candidate model $\mathcal{M}$, we use the notation

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0083

$\mathbf{U}_{i(\mathcal{M})} = \{U_{ij} : j \in \mathcal{M}\}$ to denote the subvector of $\mathbf{U}_i$ corresponding to $\mathcal{M}$. Similarly, $\mathbf{U}_{(\mathcal{M})}$ is the "subdesign" matrix corresponding to $\mathcal{M}$. Lastly, let $\mathbf{Y} = (Y_1, \cdots, Y_n)^{\mathrm{T}} \in \mathbb{R}^n$ be the response vector.

2.2 FAR Algorithm

Under the assumption that $\mathcal{T}$ exists, our main objective is to discover all relevant predictors consistently. To this end, we propose the following FAR algorithm for Model (1) when the dimension p is ultrahigh.

Algorithm 1 Forward Additive Regression Algorithm

Step 1. (Initialization). Set $\mathcal{S}^{(0)} = \emptyset$. Let the step index $\ell = 0$.

Step 2. (FAR Updating).

(2.1) *Evaluation.* In the ℓth step ($\ell \geq 1$), given $\mathcal{S}^{(\ell-1)}$, we construct a candidate model $\mathcal{M}_j^{(\ell-1)} = \mathcal{S}^{(\ell-1)} \cup \{j\}$ for every $j \in \mathcal{A}\backslash\mathcal{S}^{(\ell-1)}$. Then, we compute the residual sum squares $\mathrm{RSS}_j^{(\ell-1)} = \mathbf{Y}^{\mathrm{T}}\{\mathbf{I}_n - \tilde{\mathbf{H}}_j^{(\ell-1)}\}\mathbf{Y}$ for $\mathcal{M}_j^{(\ell-1)}$, where $\tilde{\mathbf{H}}_j^{(\ell-1)} = \mathbf{U}_{(\mathcal{M}_j^{(\ell-1)})}\{\mathbf{U}^{\mathrm{T}}_{(\mathcal{M}_j^{(\ell-1)})}\mathbf{U}_{(\mathcal{M}_j^{(\ell-1)})}\}^{-1}\mathbf{U}^{\mathrm{T}}_{(\mathcal{M}_j^{(\ell-1)})}$ is a projection matrix and $\mathbf{I}_n \in \mathbb{R}^{n\times n}$ is the identity matrix.

(2.2) *Screening.* We obtain $a_\ell = \underset{j\in\mathcal{A}\backslash\mathcal{S}^{(\ell-1)}}{\arg\min}\, \mathrm{RSS}_j^{(\ell-1)}$, and update $\mathcal{S}^{(\ell)} = \mathcal{S}^{(\ell-1)} \cup \{a_\ell\}$.

Step 3. (Solution Path). Iterate Step 2 $[n/m_n]$ times, yielding a total of $[n/m_n]$ nested candidate models, where $[c]$ denotes the largest integer no larger than the value c. We then collect those models a solution path $\mathbb{S} = \{\mathcal{S}^{(\ell)} : 1 \leq \ell \leq [n/m_n]\}$, with $\mathcal{S}^{(\ell)} = \{a_1, \cdots, a_\ell\}$.

The FAR algorithm extends the FR for linear models proposed by Wang (2009) to nonparametric additive models using B-spline smoothing tech-

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0083

niques. The algorithm is computationally efficient and easy to implement. The first step of the FAR is similar to that of the NIS of Fan, Feng, and Song (2011). That is, they both select a predictor that achieves the minimum RSS from all marginal regression models. In the remaining steps, and in contrast to existing screening methods, which treat all predictors independently, the FAR keeps all preselected predictors in the models and evaluates the conditional contributions of new predictors to the response. Thus, the FAR selects the important predictors that are marginally independent of the response. At the same time, it avoids adding unimportant predictors that have high correlations with the important variables. Note that this FAR procedure is similar to the work of Cheng et al. (2015), who proposed a groupwise forward selection procedure for linear models. However, there are some conceptual differences between the two. First, the working models are different. Cheng et al. (2015) focused on a linear model and suggested a groupwise forward regression, whereas we study a nonparametric additive model. This difference is the same as that between Yuan and Lin (2006), who use a group LASSO to select variables in a groupwise manner in linear models, and Huang, Horowitz, and Wei (2010), who studied variable selection in nonparametric additive models using an adaptive group LASSO. Second, in the additive model, we need to estimate

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0083

the unknown nonparametric components using m_n B-spline basis functions. Here, m_n is allowed to go to infinity, in theory. In the theoretical proofs, we deal with the difference between the estimated $f_{nj}(\cdot)$ and the true function $f_j(\cdot)$, which adds further challenges.

Remark: although the FAR procedure is computationally easy and efficient, it incurs a greater computational burden than that of marginal variable screening methods, such as the NIS. The computational complexity of each step of the FAR is similar to that of the NIS. Because the computational complexity of the NIS is $O(nm_np)$, the computational complexity of the FAR is $O(Knm_np)$, where K is the user-specified number of steps. If we choose $K = [n/m_n]$, the computational complexity becomes $O(n^2p)$, which is still linearly related to the dimensionality of the predictors.

3. Theoretical Properties

In this section, we present the regularity assumptions for the FAR algorithm and establish its screening consistency property. Despite the iterative marginal screening methods being able to address the drawbacks of simple marginal screenings in practice, they lack some theoretical justification. The following theorems fill this gap by studying the theoretical properties of the FAR algorithm.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0083

The following technical assumptions are required to establish the screening consistency of the FAR.

(A1) The nonparametric components $\{f_j, j = 1, \ldots, p\}$ belong to a class of functions $\mathcal{F}$, the rth derivative of which, $f^{(r)}$, exists and satisfies a Lipschitz condition of order α. That is,

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

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

(A2) The support of each predictor, X_j, is $[a, b]$, where a and b are finite real numbers. The marginal density function g_j of X_j satisfies $0 < K_1 \leq g_j(X_j) \leq K_2 < \infty$ on $[a, b]$, for $j = 1, \cdots, p$, for some constants K_1 and K_2.

(A3) The number of nonzero components p_0 is fixed and there is a constant $c_f > 0$, such that $\min_{j \in \mathcal{T}} \|f_j\|_2 \geq c_f$.

(A4) The random errors $\{\varepsilon_i, i = 1, \ldots, n\}$ are independent and identically distributed (i.i.d.) with conditional mean zero. For any $K_3 > 0$, there exists a positive constant K_4, such that $E\{\exp(K_2|\varepsilon_i|)|\mathbf{x}_i\} < K_4$.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0083

(A5) There exist constants $\log p = O(n^{c_p})$ with $0 < c_p < 2d/(2d+1)$.

These technical assumptions are standard conditions for high-dimensional nonparametric regression models. See Huang, Horowitz, and Wei (2010); Fan, Feng, and Song (2011), and Fan and Zhong (2016). In particular, (A1) is the Lipschitz condition, which is commonly assumed in the nonparametric literature to require that the function is sufficiently smooth. (A2) is the same as Condition (A4) in Huang, Horowitz, and Wei (2010) and Condition (B) in Fan, Feng, and Song (2011). (A3) is the same as (A1) in Huang, Horowitz, and Wei (2010), and (A4) is identical to Condition (E) in Fan, Feng, and Song (2011). (A5) allows us to deal with the ultrahigh-dimensionality with $p = O\{\exp(n^{c_p})\}$.

Next, we establish the screening consistency property of the FAR method in the following theorem.

Theorem 1. (Screening Consistency Property) *Suppose conditions (A1)–(A5) hold and* $K_0 > c_2 \mathrm{var}(Y)/c_1^2 c_3 c_f^2$, *where* c_1, c_2, *and* c_3 *are the positive constants defined in the proofs. Then we have that*

$$P(\mathcal{T} \subset \mathcal{S}^{(p_0 K_0)}) \to 1.$$

Theorem 1 states that the FAR algorithm can detect all relevant predictors

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0083

within p_0K_0 steps, with probability tending to one. The theorem remedies the lack of a theoretical justification of the INIS for nonparametric additive models. Note that we implicitly require that $p_0K_0 < [n/m_n]$ in the practical implementation, because the FAR algorithm can run for at most $[n/m_n]$ steps.

Furthermore, we follow Wang (2009) to select the best model using the extended BIC (Chen and Chen, 2008), which is defined as

$$\mathrm{BIC}(\mathcal{M}) = \log\hat{\sigma}^2(\mathcal{M}) + n^{-1}m_n|\mathcal{M}|(\log n + 2\log p m_n), \tag{1}$$

where $\mathcal{M}$ is an arbitrary candidate model, with $|\mathcal{M}| < [n/m_n]$, and $\hat{\sigma}^2(\mathcal{M}) = (n - |\mathcal{M}|)^{-1}\mathbf{Y}^{\mathrm{T}}\{\mathbf{I}_n - \mathbf{H}_{(\mathcal{M})}\}\mathbf{Y}$, where $\mathbf{H}_{(\mathcal{M})} = \mathbf{U}_{(\mathcal{M})}\{\mathbf{U}_{(\mathcal{M})}^{\mathrm{T}}\mathbf{U}_{(\mathcal{M})}\}^{-1}\mathbf{U}_{(\mathcal{M})}^{\mathrm{T}}$. We define $\hat{m} = \arg\min_{1\le m\le[n/m_n]} \mathrm{BIC}(\mathcal{S}^{(m)})$ and $\widehat{\mathcal{S}} = \mathcal{S}^{(\hat{m})}$. In the following theorem, we show theoretically that the FAR algorithm using the extended BIC also enjoys the sure-screening property. That is, the set of truly relevant predictors can be contained in the selected model $\widehat{\mathcal{S}}$.

Theorem 2. (BIC) *Under model 1, suppose conditions (A1)−(A5) hold. Then, as* $n \to 1$,

$$P\left(\mathcal{T} \subseteq \widehat{\mathcal{S}}\right) \to 1.$$

Although the FAR algorithm with the extended BIC rule satisfies sure-

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0083

screening consistency, in practice, we recommend applying a sophisticated regularization method for nonparametric additive models after the screening step. Examples included the COSSO of Lin and Zhang (2006) and adaptive grouped LASSO of Huang, Horowitz, and Wei (2010). This helps to refine the selection of relevant predictors and achieve better theoretical properties, such as the oracle property and selection consistency.

4. Numerical Studies

4.1 Monte Carlo Simulation

In this section, Monte Carlo simulations are carried out to investigate the finite-sample performance of the FAR approach and to compare it with existing screening procedures such as the SIS and ISIS (Fan and Lv, 2008), DC-SIS (Li, Zhong, and Zhu, 2012), DC-ISIS (Zhong and Zhu, 2014), and NIS and INIS (Fan, Feng, and Song, 2011). To implement the FAR, NIS, and INIS, we set $m_n = [n^{1/5}] + 2 = 5$.

In the simulation, we choose $n = 200$ and $p = 1000$. The FAR can choose at most $[n/m_n] = 40$ covariates. Following Fan and Lv (2008), we set the selected model size $[n/\log n] = 37$. To make the comparison as fair as possible, we stop the FAR algorithm when it selects 37 predictors. For all iterative screening methods, we execute the screening procedure once

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0083

only. That is, we select the first $[n/(2\log n)] = 18$ predictors using the marginal screening method in the first step, and then choose the remaining 19 covariates in the following iteration step.

Each experiment is repeated 100 times. We follow Li, Zhong, and Zhu (2012) to evaluate the finite-sample performance using the following two criteria: (1) the proportion of which a single relevant predictor is selected from 100 replications, denoted by $\mathcal{P}_s$; and (2) the proportion of which all true predictors are selected from 100 replications, denoted by $\mathcal{P}_a$. We claim that larger $\mathcal{P}_s$ and $\mathcal{P}_a$ lead to better performance. Ideally, both $\mathcal{P}_s$ and $\mathcal{P}_a$ are equal to one, which means that all truly relevant predictors are added to the reduced model.

Example 1. In this example, we generate data from the following model:

$$Y = 3f_1(X_1) + f_2(X_2) - 1.5f_3(X_3) + f_4(X_4) + \varepsilon, \tag{1}$$

where $f_1(x) = -\sin(2x)$, $f_2(x) = x^2 - 25/12$, $f_3(x) = x$, and $f_4(x) = \exp(x) - 2/5 \cdot \sinh(5/2)$, $\varepsilon \sim N(0,1)$. We generate the covariates $\mathbf{x} = (X_1, \cdots, X_p)^{\mathrm{T}}$ from a multivariate normal distribution $\mathrm{MVN}(\mathbf{0}, \mathbf{\Sigma})$. Here, we consider two matrices $\mathbf{\Sigma} = (\sigma_{ij})_{p\times p}$: (1) an AR(1) structure, $\sigma_{ij} = \rho^{|i-j|}$; and (2) a compound symmetry (CS) structure, $\sigma_{ij} = \rho$, for $i \neq j$. We also

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0083

consider three levels of correlations, $\rho = 0.2, 0.5$ and 0.8.

The empirical results are shown in Table 1. For the AR(1) structure with large ρ, the four adjacent truly relevant predictors are highly correlated. Thus, both screening methods and the FAR are able to select these predictors with a large probability. When $\rho = 0.2$, we ignore the performance of the SIS and DC-SIS. On the other hand, for the CS structure with large ρ, all pairs of variables have a high correlation, which makes any marginal screening methods perform worse. This is because marginal screening methods tend to add some unimportant variables that are highly correlated with large true signals. The iterative screenings can refine the selection. However, the FAR method outperforms the other methods in all cases, especially when the correlations are high.

Note that the computational complexity of the FAR is $O(Knm_np)$, where K is the number of steps. In this example, the computational time for each run of the NIS is 1.9 seconds on average, based on 100 simulations on a personal computer (64-bit windows 10 system, Intel (R) i7-6650U CPU, 2.21 GHz, 16GB RAM). The computational time for the FAR is 157.3 seconds, on average, which is less than three minutes. If we choose $K = 10$, the computational time for the FAR decreases to 22.6 seconds. Note that even when $K = 10$ is smaller, the FAR selects all relevant predictors in this

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0083

example. As expected, the computational cost of the FAR is heavier than that of the marginal variable screening, but is acceptable for better variable selection performance, in practice.

Table 1: The proportions $\mathcal{P}_s$ and $\mathcal{P}_a$ in Example 1. The selected model size is 37.

		(1) AR Structure					(2) CS Structure				
		$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$
ρ	Method	X_1	X_2	X_3	X_4		X_1	X_2	X_3	X_4	
	FAR	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	SIS	0.87	0.51	1.00	1.00	0.46	0.77	0.09	1.00	1.00	0.08
	ISIS	0.88	0.42	1.00	1.00	0.39	0.68	0.11	1.00	1.00	0.07
0.2	DC-SIS	1.00	0.95	1.00	1.00	0.95	1.00	0.31	1.00	1.00	0.31
	DC-ISIS	1.00	0.89	1.00	1.00	0.89	1.00	0.51	1.00	0.51	1.00
	NIS	1.00	0.96	1.00	1.00	1.00	1.00	0.79	0.97	1.00	0.77
	INIS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	FAR	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	SIS	1.00	1.00	1.00	1.00	1.00	0.64	0.04	0.99	1.00	0.04
	ISIS	1.00	1.00	1.00	1.00	1.00	0.53	0.04	0.98	1.00	0.00
0.5	DC-SIS	1.00	1.00	1.00	1.00	1.00	1.00	0.18	0.99	1.00	0.18
	DC-ISIS	1.00	1.00	1.00	1.00	1.00	1.00	0.42	0.97	0.97	0.38
	NIS	1.00	1.00	1.00	1.00	1.00	0.95	0.59	0.81	1.00	0.49
	INIS	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.98	1.00	0.97
	FAR	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	SIS	1.00	1.00	1.00	1.00	1.00	0.36	0.06	0.83	0.91	0.00
	ISIS	1.00	1.00	1.00	1.00	1.00	0.27	0.08	0.69	0.84	0.01
0.8	DC-SIS	1.00	1.00	1.00	1.00	1.00	1.00	0.14	0.83	0.88	0.11
	DC-ISIS	1.00	1.00	1.00	1.00	1.00	0.98	0.34	0.71	0.78	0.18
	NIS	1.00	1.00	1.00	1.00	1.00	0.94	0.55	0.71	1.00	0.37
	INIS	1.00	1.00	1.00	1.00	1.00	1.00	0.96	0.93	1.00	0.91

Example 2. Following Meier, Geer, and Bühlmann (2009) and Fan, Feng,

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0083

Table 2: The proportions $\mathcal{P}_s$ and $\mathcal{P}_a$ in Example 2. The selected model size is 37.

	$t=0$					$t=1$				
	$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$
Method	X_1	X_2	X_3	X_4		X_1	X_2	X_3	X_4	
FAR	1.00	1.00	1.00	1.00	1.00	0.98	0.81	1.00	1.00	0.79
SIS	1.00	0.03	1.00	1.00	0.03	0.00	0.02	1.00	1.00	0.00
ISIS	1.00	0.05	1.00	1.00	0.05	0.00	0.00	1.00	1.00	0.00
DC-SIS	1.00	0.31	1.00	1.00	0.31	0.00	0.05	1.00	1.00	0.00
DC-ISIS	1.00	0.28	1.00	1.00	0.28	0.86	0.05	1.00	1.00	0.04
NIS	1.00	0.81	1.00	1.00	0.81	0.00	0.28	1.00	1.00	0.00
INIS	1.00	1.00	1.00	1.00	1.00	0.97	0.49	1.00	1.00	0.48

and Song (2011), we generate the data from the following additive model:

$$Y = 5f_1(X_1) + 3f_2(X_2) + 4f_3(X_3) + 6f_4(X_4) + \sqrt{1.74}\varepsilon, \tag{2}$$

where $f_1(x) = x$, $f_2(x) = (2x-1)^2$, $f_3(x) = \sin(2\pi x)/\{2-\sin(2\pi x)\}$, and $f_4(x) = 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + \sin(2\pi x)^2 + 0.4\cos(2\pi x)^3 + 0.5\sin(2\pi x)^3$. The covariates are simulated according to the random-effect model

$$X_j = \frac{W_j + tU}{1+t}, \qquad j = 1, \cdots, p,$$

where $W_1, \cdots, W_p$ and U are i.i.d. Uniform(0,1), and $\varepsilon \sim N(0,1)$. When $t = 0$, the covariates are all independent. In this case, both the INIS

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0083

and FAR exhibit the same satisfactory performance in terms of adding all important covariates. Other methods have some difficulty selecting a variable that is quadratically correlated with the response. When $t = 1$, the pairwise correlation of covariates is 0.5, in which case, marginal screening methods fail to detect the first two variables. Both the FAR and the INIS can perform reasonably well.

Example 3. Following Fan and Lv (2008), we consider the following linear model, which is actually a special case of the additive model:

$$Y = cX_1 + cX_2 + cX_3 - 3c\sqrt{\rho}X_4 + \varepsilon, \tag{3}$$

where $\varepsilon \sim N(0, 1)$ and c is a constant used to control the signal-to-noise ratio (SNR). Here, we consider two kinds of SNRs: $c = 5$ and $c = 2.5$. The covariates are simulated from a multivariate normal distribution. All X_k except X_4 are equally correlated with a Pearson correlation coefficient ρ, whereas X_4 has a Pearson correlation $\sqrt{\rho}$ with all other $p - 1$ variables. This makes X_4 marginally independent of the response, although it is truly relevant for the response in the linear model. In the simulation, we set $\rho = 0.2,\ 0.5,$ and 0.8. Because X_4 is marginally independent of Y, no marginal screenings detect X_4. Although both the INIS and DC-ISIS perform very

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0083

well in all cases, the empirical performance of the FAR is even better when the correlation is high, such as $\rho = 0.8$.

Table 3: The proportions $\mathcal{P}_s$ and $\mathcal{P}_a$ in Example 3. The selected model size is 37.

		c=5					c=2.5				
		$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$
ρ	Method	X_1	X_2	X_3	X_4		X_1	X_2	X_3	X_4	
	FAR	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	SIS	1.00	1.00	1.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00
	ISIS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.95	0.95
0.2	DC-SIS	1.00	1.00	1.00	0.01	0.01	1.00	1.00	1.00	0.00	0.00
	DC-ISIS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	NIS	1.00	1.00	1.00	0.01	0.01	1.00	1.00	1.00	0.00	0.00
	INIS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	FAR	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	SIS	1.00	1.00	1.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00
	ISIS	1.00	1.00	1.00	0.01	0.01	1.00	1.00	1.00	0.00	0.00
0.5	DC-SIS	1.00	1.00	1.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00
	DC-ISIS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	NIS	0.99	0.99	1.00	0.00	0.00	1.00	0.99	1.00	0.01	0.01
	INIS	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	0.99	0.99
	FAR	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	SIS	0.96	0.93	0.95	0.00	0.00	0.93	0.92	0.92	0.00	0.00
	ISIS	1.00	1.00	1.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00
0.8	DC-SIS	0.95	0.92	0.94	0.00	0.00	0.91	0.89	0.93	0.00	0.00
	DC-ISIS	1.00	1.00	1.00	0.97	0.97	1.00	1.00	1.00	0.94	0.94
	NIS	0.90	0.85	0.93	0.00	0.00	0.76	0.80	0.82	0.00	0.00
	INIS	1.00	1.00	1.00	0.88	0.88	1.00	0.99	1.00	0.80	0.79

Example 4 In this example, following Zhong and Zhu (2014), we generate

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0083

Table 4: The proportions $\mathcal{P}_s$ and $\mathcal{P}_a$ in Example 4. The selected model size is 37.

	$\rho = 0.5$					$\rho = 0.8$				
	$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$
Method	X_1	X_{101}	X_{201}	X_{202}		X_1	X_{101}	X_{201}	X_{202}	
FAR	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
SIS	1.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
ISIS	1.00	1.00	0.02	0.03	0.00	1.00	1.00	0.02	0.01	0.00
DC-SIS	1.00	1.00	0.05	0.04	0.00	1.00	1.00	0.28	0.26	0.17
DC-ISIS	1.00	1.00	0.79	0.86	0.67	1.00	1.00	0.94	0.90	0.84
NIS	1.00	1.00	0.61	0.46	0.33	1.00	1.00	0.85	0.77	0.71
INIS	1.00	1.00	0.99	1.00	0.99	1.00	1.00	1.00	0.99	0.99

our data from the following additive model:

$$Y = 2f_1(X_1) + \sqrt{6}f_2(X_{101}) + 3f_3(X_{201}) - 0.6f_4(X_{202}) + \varepsilon, \tag{4}$$

where ε is i.i.d. N(0,1), $f_1(x) = \exp(2x/3)$, $f_3(x) = \sin(3\pi x/4 + 3/2)/\{2 - \sin(3\pi x/4 + 3/2)\}$, $f_4(x) = \log(x^2)$, and

$$f_2(x) = \begin{cases} x - 4, & \text{if } x < -2, \\ |x| & \text{if } |x| < 2, \\ 4 - x & \text{if } x > 2. \end{cases}$$

In this example, we first generate p-dimensional predictors $\mathbf{x}$ from MVN($\mathbf{0}$,$\mathbf{\Sigma}$), where $\mathbf{\Sigma} = (\sigma_{ij})_{p\times p}$ and $\sigma_{ij} = \rho^{|i-j|}$, for $i, j = 1, \cdots, p$. Then, we re-

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0083

place each X_k with $X_k^* = 0.8X_1 + \xi_k$, with ξ_k from i.i.d. N(0,1) for $k = 2, 3, \cdots, 100$. Hence, $(X_2, \cdots, X_{100})$ are highly correlated with X_1. We want to check whether these screening methods and the FAR can identify X_{101}, X_{201}, and X_{202}.

The results are summarized in Table 4. We can see that the marginal screening methods (SIS, NIS, DC-SIS) fail to pick out the active predictors X_{101}, X_{201}, and X_{202}. This is because the 99 unimportant covariates that present a strong signal because of their high correlation with X_1 hide the marginal signals of the other three active predictors. However, the FAR and iterative screening procedures can remove the relatively strong signals from $X_2, X_3, \cdots, X_{100}$, and then reveal the truly marginal signals of those active predictors. The FAR method can be considered a conditional screening method. Once X_1 is selected and kept in the model, other spurious variables $X_2, X_3, \cdots, X_{100}$, have a lower chance of being selected than do truly active variables. In summary, the FAR method performs very well in terms of variable screening for ultrahigh-dimensional nonparametric additive models.

4.2 Cardiomyopathy Microarray Data

In this section, we use cardiomyopathy microarray data with $(n, p) = (30, 6319)$ to examine the empirical performance of the FAR method and

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0083

to compare it with other existing methods. This data set was reported by Segal, Dahlquist, and Conklin (2003), Hall and Miller (2009), and Li, Zhong, and Zhu (2012). The goal is to identify the most important genes for the overexpression of a G protein-coupled receptor (Ro1) in mice. In this example, we use the Ro1 expression as the response variable Y, and other gene expression levels as the covariates $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$.

We first standardize the data. Then, we apply the FAR, SIS, ISIS, NIS, INIS, DC-SIS, and DC-ISIS. For the nonparametric B-spline estimation, we choose $m_n = 3$, which is the same as the cubic splines used in Hall and Miller (2009). The FAR can choose at most $[n/m_n] = 10$ covariates. As in our simulations, we set the selected model size $d = [n/\log n] = 8$. To compare all models fairly, the FAR also chooses eight covariates. For the iterative methods, we choose four covariates in the first step, and four covariates in the second step. The selection results are reported in Figure 4.1.

We can see that all methods rank the gene labeled $Mas.2877.0$ at the top. Both Hall and Miller (2009) and Li, Zhong, and Zhu (2012) claimed that gene $Mas.2877.0$ is the most predictive for the response. To make the problem more challenging and to compare the performance of the different methods, we create 20 artificial covariates that are highly correlated with

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0083

gene $Mas.2877.0$, as follows:

$$X_{Art.i} = \frac{4}{3} X_{Mas.2877.0} + \varepsilon_i, \qquad i = 1, \cdots, 20, \tag{5}$$

where ε_i are i.i.d. N(0,1). The results show that the Pearson correlation between each artificial gene and gene $Mas.2877.0$ is 0.8. Then, we apply the above methods again to the real data and the 20 artificial genes. The selection results are reported in Figure 4.1, where a row corresponding to a method name with the label $*$, such as FAR*, shows the result from this method for the data with artificial genes. The squares in black and red denote genes selected by a particular method, and those in blue denote genes that are not selected. We observe that gene $Mas.2877.0$ is selected by all seven methods. However, when we add noise to the data set, all methods except the FAR select at least one artificial covariate. The FAR performs robustly in the presence of noise, making it a useful alternative approach to variable screening in ultrahigh-dimensional nonparametric additive models.

5. Conclusion

We propose a forward additive regression (FAR) for ultrahigh-dimensional nonparametric additive models. The FAR method estimates the nonparametric components and selects important predictors iteratively to deter-

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0083

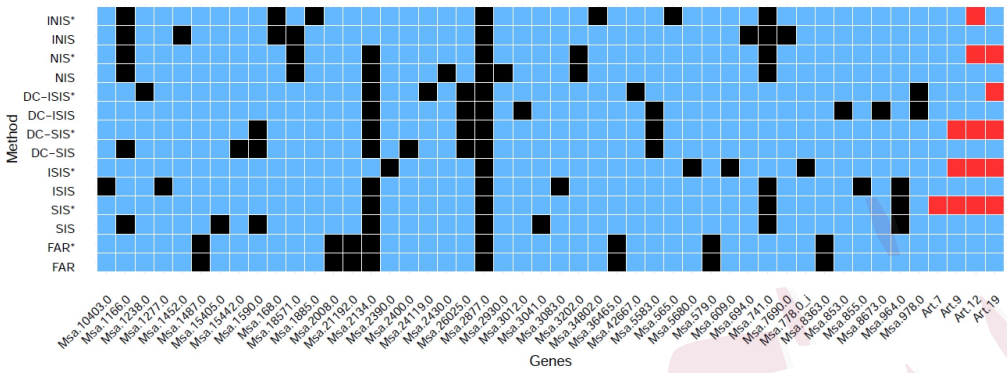

Figure 1: Gene selection results in the cardiomyopathy microarray data example.

mine a solution path. Compared with the penalized regression, our proposed method is computationally more efficient and, hence, is useful for ultrahigh-dimensional predictors. Compared with the screening methods, our proposed method identifies important predictors that are marginally independent of the response variables. Compared with the iterative screening methods, our method exhibits the desirable sure-screening property. Comprehensive numerical studies using simulations and real data confirm the effectiveness of the proposed method.

Supplementary Material

All technical proofs are included in the online Supplemental Material.

Acknowledgment

Zhong's work is supported by the National Natural Science Foundation

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0083

of P. R. China (NNSFC) 11671334, 11301435, University Distinguished Young Researchers Program in Fujian Province's and the Fundamental Research Funds for the Central Universities 20720181004. Zhu is the corresponding author and his work is supported by NNSFC 11731011, Project of Key Research Institute of Humanities and Social Sciences at Universities (16JJD910002) and National Youth Top-notch Talent Support Program, P. R. China.

References

Belloni, A., Chen, D., Chernozhukov V., and Hansen, C. (2012), "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain," *Econometrica*, **80**, 2369–2429.

Candes, E. and Tao, T. (2007), "The Dantzig Selector: Statistical Estimation When p Is Much Larger than n (with discussion)," *Annals of Statistics*, **35**, 2313–2404.

Chen,J.,and Chen, Z. (2008), "Extended Bayesian Information Criterion for Model Selection With Large Model Space," *Biometrika*, **95**, 759–771.

Cheng, M.Y., Honda, H., and Zhang, J. (2016), "Forward Variable Selection for Sparse Ultra-High Dimensional Varying Coefficient Models," *Journal of the American Statistical Association*, **111**, 1209–1221.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0083

Cheng, M.Y., Feng, S., Li, G. and Lian, H. (2015), "Greedy Forward Regression for Variable Screening," manuscript, arXiv:1511.01124.

Cui, H., Li, R., and Zhong, W. (2015), "Model-free Feature Screening for Ultrahigh Dimensional Discriminant Analysis," *Journal of the American Statistical Association,* **110,** 630-641.

Fan, J., Feng, Y., and Song, R. (2011), "Nonparametric Independence Screening in Sparse Ultra-high Dimensional Additive Models," *Journal of the American Statistical Association*, **106**, 544–557.

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and It Oracle Properties," *Journal of the American Statistical Association*, **96**, 1348–1360.

Fan, J. and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space (with discussion)," *Journal of the Royal Statistical Society, Series B*, **70**, 849–911.

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.

Fan, J., Samworth, R., and Wu, Y. (2009), "Ultrahigh Dimensional Feature Selection: Beyond the Linear Model," *Journal of Machine Learning Research*, **10**, 1829–1853.

Fan, J. and Song, R. (2010), "Sure Independence Screening in Generalized Linear Models with NP-dimensionality," *The Annals of Statistics*, **38**, 3567–3604.

Fan, Q. and Zhong, W. (2016), "Nonparametric Additive Instrumental Variable Estimator: a

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0083

Group Shrinkage Estimation Perspective," *Journal of Business and Economic Statistics*, forthcoming.

He, X., Wang, L., and Hong, H. (2013), "Quantile-Adaptive Model-Free Variable Screening for High-Dimensional Heterogeneous Data", *The Annals of Statistics*, **41**, 342–369.

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**, 553–550.

Li, R., Zhong, W. and Zhu, L. (2012), "Feature screening via distance correlation learning," *Journal of American Statistical Association*, **107**, 1129–1139.

Huang, D., Li, R. and Wang, H. (2014), "Feature Screening for Ultrahigh Dimensional Categorical Data with Applications," *Journal of Business and Economic Statistics*, **32**, 237–244.

Huang, J., Horowitz, J. and Wei, F. (2010), "Variable Selection in Nonparametric Additive Models," *The Annals of Statistics,* **38**, 2282–2313.

Huang, J., Ma, S. G., and Zhang, C. H. (2008), "Adaptive Lasso for Sparse High-dimensional Regression Models," *Statistica Sinica*, **18**, 1603–1618.

Li, G., Peng. H., Zhang J., and Zhu, L. (2012), "Robust Rank Correlation Based Screening," *The Annals of Statistics*, **40**, 1846–1877.

Li, R., Zhong, W., and Zhu, L. (2012), "Additive Regression and Other Nonparametric Models," *Journal of the American Statistical Association*, **107**, 1129–1139.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0083

Lin, Y., and Zhang, H. H. (2006), "Component Selection and Smoothing in Multivariate Nonparametric Regression," *The Annals of Statistics,* **34**, 2272–2297.

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.

Mai, Q. and Zou, H. (2013), "The Kolmogorov Filter for Variable Screening in High-Dimensional Binary Classification," *Biometrika*, **100**, 229–234.

Meier, L., Geer, V., and Bühlmann, P. (2009), "High-Dimensional Additive Modeling," *The Annals of Statistics*, **37**, 3779–3821.

Segal, M. R., Dahlquist, K. D., and Conklin, B. R. (2003), "Regression Approach for Microarray Data Analysis," *JJournal of Computational Biology*, **10**, 961–980.

Shao, X. and Zhang, J. (2014), "Martingale Difference Correlation and Its Use in High-dimensional Variable Selection," *Journal of the American Statistical Association,* **109,** 1302-1318.

Stone, C. (1985), "Additive Regression and Other Nonparametric Models," *The Annals of Statistics*, **13**, 689–705.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via LASSO," *Journal of the Royal Statistical Society, Series B*, **58**, 267–288.

Wang, H. (2009), "Forward Regression for Ultra-High Dimensional Variable Screening," *Journal*

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0083

of the American Statistical Association, **104**, 1512–1524.

Yuan, M. and Lin, Y. (2006), "Model Selection and Estimation in Regression with Grouped Variables," *Journal of the Royal Statistical Society, Series B*, **68**, 49–67.

Zhong, W., and Zhu, L. (2014), "An Iterative Approach to Distance Correlation-based Sure Independence Screening," *Journal of Statistical Computation and Simulation*, **80**, 1–15.

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

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, **101**, 1418–1429.

Zou, H. and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society, Series B*, **67**, 301–320.

Zou, H. and Li, R. (2008), "One-step Sparse Estimates in Nonconcave Penalized Likelihood Models," *Annals of Statistics*, **36**, 1509–1533.

Wei Zhong, Wang Yanan Institute for Studies in Economics, Department of Statistics, School of Economics, Fujian Key Laboratory of Statistical Science, Xiamen University, Xiamen 361005, China.

E-mail: wzhong@xmu.edu.cn

Sunpeng Duan, Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen 361005, China.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0083

E-mail: fredduan.dsp@gmail.com

Liping Zhu, Center for Applied Statistics, Institute of Statistics and Big Data, Renmin University of China, Beijing 100872, China.

E-mail: zhu.liping@ruc.edu.cn