Statistica Sinica **24** (2014), 1365-1387
doi:http://dx.doi.org/10.5705/ss.2012-316

VARIABLE SELECTION FOR SPARSE HIGH-DIMENSIONAL NONLINEAR REGRESSION MODELS BY COMBINING NONNEGATIVE GARROTE AND SURE INDEPENDENCE SCREENING

Shuang Wu[1], Hongqi Xue[1], Yichao Wu[2] and Hulin Wu[1]

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

Abstract: In many regression problems, the relations between the covariates and the response may be nonlinear. Motivated by the application of reconstructing a gene regulatory network, we consider a sparse high-dimensional additive model with the additive components being some known nonlinear functions with unknown parameters. To identify the subset of important covariates, we propose a method for simultaneous variable selection and parameter estimation by iteratively combining a large-scale variable screening (the nonlinear independence screening, NLIS) and a moderate-scale model selection (the nonnegative garrote, NNG) for the nonlinear additive regressions. We have shown that the NLIS procedure possesses the sure screening property and is able to handle problems with non-polynomial dimensionality; for finite dimension problems, the NNG for the nonlinear additive regressions has selection consistency for the unimportant covariates and estimation consistency for the parameter estimates of the important covariates. The proposed method is applied to simulated data and to real data for identifying gene regulations to illustrate its numerical performance.

Key words and phrases: Gene regulations, independence learning, nonlinear regressions, nonnegative garrote, sigmoid function, sure screening.

1. Introduction

With the rapid advancement of modern technologies, high-dimensional data now arise frequently in such areas as microarrays, RNA-seq, proteomics, biomedical imaging, signal processing, and finance. In high-dimensional statistical modeling, it is a fundamental problem to identify important explanatory variables. For linear regression models, many penalization methods have been proposed to conduct variable selection and estimation, and much effort has gone into their statistical properties in high-dimensional settings. These methods include bridge regression (Frank and Friedman (1993); Huang, Horowitz, and Ma (2008)), least absolute shrinkage and selection operator or Lasso (Tibshirani (1996); Zhao and Yu (2006); van de Geer (2008); Zhang and Huang (2008)), the nonnegative garrote (Breiman (1995); Yuan and Lin (2007)), the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li (2001); Fan and Peng (2004); Kim, Choi, and Oh (2008)), the adaptive Lasso (Zou (2006)), and the Dantzig selector (Candes and Tao (2007)).

In many applications, there is little prior information to justify the assumption that the effects of covariates on the response take a linear form. A widely adopted approach to address nonlinearity is to extend the linear regression to a nonparametric additive model. The nonlinear effect of a covariate is modeled through a nonparametric function that is usually approximated by a linear combination of some basis functions. The selection of important variables then corresponds to the selection of groups of basis functions. Methods for selecting grouped variables have been studied in Antoniadis and Fan (2001) and Yuan and Lin (2006). Lin and Zhang (2006) proposed the component selection and smoothing operator (COSSO) method for model selection in smoothing spline ANOVA with a fixed number of covariates. Recently, Ravikumar et al. (2009), Meier, Geer, and Bühlmann (2009), and Huang, Horowitz, and Wei (2010) considered variable selection in high-dimensional nonparametric additive models where the number of additive components is larger than the sample size.

Nonparametric additive models are flexible enough to account for a variety of nonlinear relationships, and are essentially linear once the nonparametric components are expressed as linear combinations of basis functions. Consequently, computation is relatively easy. Still alternative nonlinear models with known functional forms are necessary for some applications. In particular, a given nonlinear model can be more interpretable, and it avoids the selection of nonparametric smoothing parameters such as the number of bases, whose impact on variable selection has not been well understood. We focus on exploring variable selection for nonlinear additive models.

Our work is motivated by the application of reconstructing a gene regulatory network (GRN), in which the regulatory function of each gene is a sigmoid function. The sigmoid function model for describing a GRN has been discussed by Mestl, Plahte, and Omholt (1995) and Chen et al. (2004), among others. A simple sigmoid function is

$$f(x) = \frac{1}{1+e^{-\alpha x}}. \tag{1.1}$$

It is usually used in modeling nonlinear systems that exhibit "saturation", the parameter $\alpha > 0$ indicating the transition rate. Since in biological systems it is common that most nodes are only directly connected to a small number of other nodes (Jeong et al. (2000)), it is often assumed that the GRN is a sparse network. The identification of a GRN then requires the selection of the important regulators for each target gene using variable selection techniques.

We consider the high-dimensional nonlinear additive model

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j f_j(X_j, \boldsymbol{\alpha}_j) + \epsilon, \tag{1.2}$$

where Y is the response variable, $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ is the vector of covariates, $\boldsymbol{f}(\boldsymbol{X}, \boldsymbol{\alpha}) = \{f_1(X_1, \boldsymbol{\alpha}_1), \ldots, f_p(X_p, \boldsymbol{\alpha}_p)\}^T$ are known nonlinear functions, and ϵ is the random error term with mean 0 and variance $0 < \sigma^2 < \infty$. The parameters $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_p^T)^T$ usually control the shape of $\boldsymbol{f}$. We let the dimension p of the covariates increase with the sample size n, even larger than n, and the coefficient vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is to be sparse in the sense that most of its elements are zeros. Given the i.i.d. observed data $\{(\boldsymbol{x}_i, y_i), i = 1, \ldots, n\}$ with $\boldsymbol{x}_i = (x_{i1}, x_{i2} ..., x_{ip})^T$, we are interested in identifying the subset of important covariates $\mathcal{M}_* = \{1 \leq j \leq p : \beta_j \neq 0\}$ and also in estimating the parameters $(\beta_0, \beta_j, \boldsymbol{\alpha}_j)$, $j \in \mathcal{M}_*$.

Compared with the linear and generalized linear models, there is little statistical research on statistical inference, and especially variable selection, for nonlinear regression and generalized nonlinear models. Jennrich (1969), Malinvaud (1970), and Wu (1981) proposed nonlinear least squares (NLS) estimation for nonlinear regression models and derived their asymptotic properties. A good review can be found in Seber and Wild (2003). Wei (1998), Kosmidis and Firth (2009), and Biedermann and Woods (2011) studied likelihood-based estimation and statistical inference for generalized nonlinear models. Xue, Miao, and Wu (2010) considered NLS estimation for nonlinear ordinary differential equation models with measurement errors and established their large sample theories. Jiang, Jiang, and Song (2011) proposed a weighted composite quantile regression (WCQR) estimation and studied the model selection with the adaptive LASSO and SCAD penalties for nonlinear models with a diverging number of parameters.

For given estimates $\hat{\boldsymbol{\alpha}}_j$, (1.2) can be treated as a linear regression model and variable selection methods developed for linear regressions can be then applied to estimate β_j. We propose to use the nonnegative garrote (NNG), which requires a consistent initial estimate. Intuitively, the variable selection procedure works better if the initial estimates are close to the true values and, for a number of covariates, this is difficult. Here we adopt the idea of independence screening introduced by Fan and Lv (2008). The sure independence screening (SIS) method performs efficient dimension reduction via marginal correlation learning for ultra-high dimensional feature selection. It has been shown that, with probability tending to 1, the independence screening technique retains all of the important features in the model. The SIS was later extended to high-dimensional generalized linear models (Fan and Song (2010)) and nonparametric additive models (Fan, Feng, and Song (2011)). The idea of using marginal information to deal with high dimensionality was also adopted in other works. Hall, Titterington, and Xue (2009) considered a marginal utility derived from an empirical likelihood point of view. Hall and Miller (2009) proposed a generalized correlation ranking that allows nonlinear relationships. More recently, Wang (2012) studied sure independence screening using a factor profiling approach; Xue and Zou (2011) considered sure independence screening for sparse signal recovery; Gorst-Rasmussen and Scheike (2012) investigated independent screening for single-index hazard rate models with right censored data; see Zhu et al. (2012) and Li, Zhong, and Zhu (2012) for recent developments of model-free approaches for feature screening.

In this paper, we propose a nonlinear independence screening method based on the residual sum of squares of the marginal modeling, and establish its sure screening property for model (1.2). We fit p marginal nonlinear regressions of the response Y against each covariate X_j separately, and rank their importance to the joint model according to the residual sum of squares of the marginal models. The covariates whose residual sum of squares are smaller than a threshold are selected and then a more refined variable selection technique such as the nonnegative garrote can be applied to the nonlinear additive model conditional on these selected covariates. This procedure can significantly reduce the dimension of the covariates and, more importantly, its sure screening property ensures that all the important covariates are retained with probability tending to 1. Our method is a non-trivial extension of Fan and Lv (2008) and Fan and Song (2010). For the nonlinear model (1.2), the minimum distinguishable signal of the marginal screening is closely related to the stochastic and numerical errors in the optimization of the nonlinear parameters. In addition, the objective function derived from the nonlinear model often has multiple local minima, making the estimation and marginal screening more challenging. We extend the sure independent screening approach from linear models to nonlinear models, making some local assumptions, such as the local convexity. We show in Section 3 that our nonlinear independence screening approach can handle problems with non-polynomial or ultra-high dimensionality.

The remainder of the paper is organized as follows. Section 2 describes the nonnegative garrote for variable selection in nonlinear additive models with finite predictors and its asymptotic properties. We elaborate on independence screening for the nonlinear additive models with ultra-high dimensional covariates and establish its sure independence screening property in Section 3. We present the results of simulation studies in Section 4 and provide an illustrative application in Section 5. Section 6 includes concluding remarks and some discussion. Proofs of the asymptotic results in Sections 2 and 3 can be found in the web-appendix.

2. Nonnegative Garrote for Nonlinear Additive Models

2.1. Method

In this section, we consider the case of $p < n$. For the multiple linear regression model

$$y_i = \sum_{j=1}^{p} \beta_j z_{ij} + \varepsilon_i,\ i = 1, \ldots, n, \tag{2.1}$$

the nonnegative garrote (Breiman (1995)) finds a set of nonnegative scaling factors c_j to minimize

$$\frac{1}{2}\sum_{i=1}^{n}\Big[y_i - \sum_{j=1}^{p} c_j \hat{\beta}_j z_{ij}\Big]^2 + n\lambda_n \sum_{j=1}^{p} c_j,\ \text{subject to } c_j \geq 0, \tag{2.2}$$

with an initial estimates $\hat{\beta}_j$ for model (2.1). The garrote estimates are then given by $\tilde{\beta}_j = \hat{c}_j \hat{\beta}_j$. An appropriately chosen λ_n can shrink some $\hat{c}_j$ to exactly 0 and thus produces a sparse model. We omit the subscript n in λ_n when no confusion occurs. The selection consistency of the nonnegative garrote was first proved by Zou (2006), and Yuan and Lin (2007) showed that, as long as the initial estimate is consistent in terms of estimation, the nonnegative garrote estimate is consistent in terms of both estimation and model selection given that the tuning parameter λ is appropriately chosen. The ordinary least squares estimator is often chosen as the initial estimation.

We extend the idea of the nonnegative garrote (NNG) to the additive nonlinear regression (1.2) with finite p predictors. To simplify the presentation, the response variable Y and the nonlinear functions f_j are assumed to be centered with $\mathrm{E}(Y) = 0$ and $\mathrm{E}[f_j(X_j, \boldsymbol{\alpha}_j)] = 0$ for all $j = 1, \ldots, p$, so the intercept $\beta_0 = 0$. The method can be easily adapted to include an intercept. Given initial estimates $\hat{\boldsymbol{\alpha}}_j$, (1.2) reduces to (2.1) with $z_{ij} = f_j(x_{ij}, \hat{\boldsymbol{\alpha}}_j)$, $j = 1, \ldots, p$, as the predictors. One can then proceed with the nonnegative garrote with the initial estimates $\hat{\beta}_j$ for variable selection and estimation. Suppose the nonnegative garrote selects covariates X_j, $j \in \hat{\mathcal{S}} = \{1 \leq j \leq p : \hat{c}_j \neq 0\}$. The parameters $\boldsymbol{\alpha}_j$ are then updated conditional on $\hat{\beta}_j$, $j \in \hat{\mathcal{S}}$. These steps iterate until some convergence criterion is met, for instance, the residual sum of squares of the fitted model does not change up to a tolerance. The initial estimates need to be carefully chosen, because the solution path consistency of the nonnegative garrote depends on the consistency of the initial estimator. We use the nonlinear least square estimates of (1.2) as the initial estimates when the model dimension of (1.2) is not very high.

The algorithm can be viewed as the solution of a separable nonlinear least squares problem (Ruhe and Wedin (1980); Golub and Pereyra (2003)). The parameters are separated into two sets, where $\{\beta_j\}$ are linear parameters and $\{\boldsymbol{\alpha}_j\}$

are nonlinear parameters. In each iteration, the optimization with respect to $\{\beta_j\}$ is performed first, and the correction to $\{\boldsymbol{\alpha}_j\}$ follows after that. It has been shown that eliminating one set of parameters can result in faster convergence of the optimization problem. We refer to Ruhe and Wedin (1980) and Golub and Pereyra (2003), and references therein, for detailed descriptions of separable nonlinear least squares problems and the convergence properties of related algorithms.

2.2. Asymptotic properties

Let $\boldsymbol{\gamma} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T$ with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_p^T)^T$, with true parameter $\boldsymbol{\gamma}_0 = (\boldsymbol{\beta}_0^T, \boldsymbol{\alpha}_0^T)^T$. Take $\ell(\boldsymbol{\gamma}, \boldsymbol{X}, Y) = [Y - F(\boldsymbol{X}, \boldsymbol{\gamma})]^2$ with $F(\boldsymbol{X}, \boldsymbol{\gamma}) = \sum_{j=1}^p \beta_j f_j(X_j, \boldsymbol{\alpha}_j)$, and $\mathrm{P}_n \ell(\boldsymbol{\gamma}, \boldsymbol{x}, \boldsymbol{y}) = n^{-1} \sum_{i=1}^n \ell(\boldsymbol{\gamma}, \boldsymbol{x}_i, y_i) = n^{-1} \sum_{i=1}^n [y_i - \sum_{j=1}^p \beta_j f_j(x_{ij}, \boldsymbol{\alpha}_j)]^2$. We need the following conditions.

(A1) For all $j = 1, \ldots, p$, $\beta_j \in \mathcal{B}_j \subset R$ and $\boldsymbol{\alpha}_j \in \mathcal{H}_j \subset R^{d_j}$, where $\mathcal{B}_j$ and $\mathcal{H}_j$ are compact and convex with finite diameters A_1 and A_2, respectively. For $\Gamma = \prod_{j=1}^p \mathcal{B}_j \times \mathcal{H}_j \subset R^{p+\sum_{j=1}^p d_j}$, $\boldsymbol{\gamma}_0$ is an interior point of Γ.

(A2) For any $\boldsymbol{\gamma} \in \Gamma$, $\mathrm{E}[F(\boldsymbol{X}, \boldsymbol{\gamma}) - F(\boldsymbol{X}, \boldsymbol{\gamma}_0)]^2 = 0$ if and only if $\beta_j = \beta_{0j}$ and $\boldsymbol{\alpha}_j = \boldsymbol{\alpha}_{0j}$ for all $j = 1, \ldots, p$.

(A3) If $\Omega_{1n} = \{\boldsymbol{X} : \|\boldsymbol{X}\|_\infty \le K_n\}$ and $I_{1n}(\boldsymbol{X}) = I(\boldsymbol{X} \in \Omega_{1n})$ for some sufficiently large positive constants K_n, $\|\cdot\|_\infty$ the supremum norm, then $F(\boldsymbol{X}, \boldsymbol{\gamma})$, $\frac{\partial}{\partial \boldsymbol{\gamma}} F(\boldsymbol{X}, \boldsymbol{\gamma})$, and $\frac{\partial^2}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T} F(\boldsymbol{X}, \boldsymbol{\gamma})$ exist and are continuous and uniformly bounded for $\boldsymbol{X} \in \Omega_{1n}$ and $\boldsymbol{\gamma} \in \Gamma$, with bounds k_1, k_2, and k_3, respectively.

(A4) The covariate vector $\boldsymbol{X}$ has a continuous density and there exist constants D_1 and D_2 such that the density function g_j of X_j satisfies $0 < D_1 \le g_j(x) \le D_2 < \infty$ on the support $[a, b]$ of X_j, $1 \le j \le p$.

(A5) There exists a sequence of random variables $\{\hat{\boldsymbol{\gamma}}\}$ with $\frac{\partial P_n \ell(\boldsymbol{\gamma}, \boldsymbol{x}, \boldsymbol{y})}{\partial \boldsymbol{\gamma}}|_{\boldsymbol{\gamma}=\hat{\boldsymbol{\gamma}}} = 0$ such that $P_n \ell(\boldsymbol{\gamma}, \boldsymbol{x}, \boldsymbol{y})$ is locally convex in some neighborhood of $\hat{\boldsymbol{\gamma}}$, say $\mathcal{A}(\delta) = \{\boldsymbol{\gamma} \in \Gamma, \|\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}\| \le \delta\}$ for some $\delta > 0$ and $\boldsymbol{\gamma}_0 \in \mathcal{A}(\delta)$.

(A6) The Fisher information

$$\mathrm{I}(\boldsymbol{\gamma}) = \mathrm{E}\left[\frac{\partial F(\boldsymbol{X}, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} \frac{\partial F(\boldsymbol{X}, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}^T}\right]$$

is finite and positive definite at $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$.

(A7) If $\Omega_n = \{(\boldsymbol{X}, Y) : \|\boldsymbol{X}\|_\infty \le K_n, |Y| \le K_n^*\}$ and $I_n(\boldsymbol{X}, Y) = I\{(\boldsymbol{X}, Y) \in \Omega_n\}$ for some sufficiently large positive constants K_n^*, there exists a constant C_1 such that with $b_n = C_1 \max\{k_1^2, A_1 k_1 k_2\} V_1^{-1} (p/n)^{1/2}$ and V_1 given in (S.6) in the web-appendix,

$$\sup_{\boldsymbol{\gamma} \in \Gamma, \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\| \le b_n} \left| E[\ell(\boldsymbol{\gamma}, \boldsymbol{X}, Y) - \ell(\boldsymbol{\gamma}_0, \boldsymbol{X}, Y)](1 - I_n(\boldsymbol{X}, Y)) \right| \le o\left(\frac{p}{n}\right).$$

Conditions (A2) and (A6) ensure identifiability and the existence of the nonnegative garrote. We use (A5) to replace Condition C in Fan and Song (2010), because our objective function may have multiple local minima. The proposed nonnegative garrote for nonlinear additive models requires an initial estimate and a natural choice is the NLS estimate. From the standard NLS estimation theories developed by Jennrich (1969), Malinvaud (1970), and Wu (1981), under (A1)−(A6), the NLS estimator $\hat{\boldsymbol{\gamma}}$ defined in (A5) is consistent and asymptotically normal. Let $\hat{c}_j(\lambda)$ minimize (2.2). The nonnegative garrote estimates are $\tilde{\beta}_j = \hat{c}_j(\lambda)\hat{\beta}_j$ and $\tilde{\boldsymbol{\alpha}}_j = \hat{\boldsymbol{\alpha}}_j$ for $j = 1, \ldots, p$.

Theorem 1.

(i) *Assume var*$\{\boldsymbol{f}(\boldsymbol{X}, \boldsymbol{\alpha}_0)\}$ *is positive definite and the initial estimate* $\hat{\boldsymbol{\gamma}}$ *satisfies* $\max_{1\le j\le p}(|\hat{\beta}_j - \beta_{0j}| + \|\hat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}_{0j}\|) = O_p(\delta_n)$ *for some* $\delta_n \to 0$. *Under* (A1)−(A3), *for* $\lambda \to 0$ *satisfying* $\delta_n = o(\lambda)$, *we have* $P\{\hat{c}_j(\lambda) = 0\} \to 1$ *for any* j *such that* $\beta_{0j} = 0$, *and* $\hat{c}_j(\lambda) = 1 + O_p(\lambda)$ *for any* j *such that* $\beta_{0j} \neq 0$.

(ii) *With the nonlinear least squares estimate as the initial estimate* $\hat{\boldsymbol{\gamma}}$, (A1)−(A6), *and* $\lambda \to 0$ *satisfying* $n^{-1/2} = o(\lambda)$, *we have* $P\{\tilde{\beta}_j = 0\} \to 1$ *for any* j *such that* $\beta_{0j} = 0$, *and estimation consistency for the estimates corresponding to important predictors almost surely.*

We establish an exponential bound for the tail probability of the NLS estimator $\hat{\boldsymbol{\gamma}}$, parallel to Theorem 1 in Fan and Song (2010), that is to be used in the next section.

Theorem 2. *For the NLS estimator* $\hat{\boldsymbol{\gamma}}$ *at* (A5), *if* (A1)−(A7) *hold, then for any* $t > 0$, *we have* $P\left(\sqrt{n}\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| \ge 16k_n^* \max\{k_1, A_1k_2\}(1+t)/V_1\right) \le \exp(-2t^2) + nP(\Omega_n^c)$, *with* $k_n^* = 2(k_1 + K_n^*)$.

Next we show that the required δ_n consistency of the initial estimates can be achieved with $\delta_n = 1/\sqrt{n}$ by using the nonlinear least squares estimate.

3. Independence Screening for High Dimensional Nonlinear Regressions

As the dimension p of covariates in (1.2) grows with the sample size n, and especially when $p \gg n$, the iterative algorithm of Section 2 may not work well. The nonlinear least squares for models with a large number of parameters and redundant covariates may not converge suitably to the local minimum, and the nonnegative garrote cannot be applied when the sample size is smaller than the number of covariates. We take the idea of sure independence screening for linear or generalized linear regression models proposed by Fan and Lv (2008) and Fan and Song (2010), and extend it to nonlinear additive regression models.

Rewrite p as p_n and fit p_n marginal nonlinear regressions of the response Y against each covariate X_j separately, then rank their importance to the joint model according to a measure of the goodness of fit of their marginal models. In Fan and Lv (2008) and Fan and Song (2010), the covariates are standardized, $\mathrm{E}X_j^2 = 1$, $j = 1, \ldots, p_n$, so the magnitude of the coefficient estimate of the marginal model can preserve the nonsparsity of the joint model. To extend this idea to (1.2), we take $\hat{\beta}'_j = \hat{\beta}_j^M \sqrt{n^{-1}\sum_{i=1}^n f_j^2(x_{ij}, \hat{\boldsymbol{\alpha}}_j^M)}$, where $\hat{\beta}_j^M$ and $\hat{\boldsymbol{\alpha}}_j^M$, $j = 1, \ldots, p_n$, are the NLS estimates of the marginal models. Such $\hat{\beta}'_j$ can be considered as the marginal NLS estimates obtained by standardizing $f_j(X_j, \boldsymbol{\alpha}_j)$ and we can use the magnitude of $\hat{\beta}'_j$ as the marginal utility for the independence screening. We refer to this strategy as marginal least square estimate or MLSE screening when we select the set of variables with the magnitude of $\hat{\beta}'_j$ greater than a prespecified threshold ζ_n:

$$\hat{\mathcal{N}} = \{1 \le j \le p_n : |\hat{\beta}'_j| \ge \zeta_n\}. \tag{3.1}$$

The MLSE screening involves $\hat{\beta}_j^M$ and $\hat{\boldsymbol{\alpha}}_j^M$, because both estimates contribute to determine the relationship between Y and X_j. Another way to incorporate the information of both $\hat{\beta}_j^M$ and $\hat{\boldsymbol{\alpha}}_j^M$ is through the residual sum of squares (RSS) of the component-wise nonlinear regressions. We propose to rank the covariates X_j's according to the RSS of the marginal models and select the set of variables:

$$\hat{\mathcal{M}} = \{1 \le j \le p_n : \mathrm{RSS}_j \le \xi_n\}, \tag{3.2}$$

with $\mathrm{RSS}_j = \min_{\beta_j, \boldsymbol{\alpha}_j} \sum_{i=1}^n [y_i - \beta_j f_j(x_{ij}, \boldsymbol{\alpha}_j)]^2$ as the residual sum of squares of the jth marginal fit, and ξ_n as a prespecified threshold value. This strategy is analogous to the likelihood ratio screening proposed in Fan, Samworth, and Wu (2009) and we refer to it as RSS screening. We show in the following section that it is asymptotically equivalent to the MLSE screening in preserving the nonsparsity of the joint model. In practical computations, the sets $\hat{\mathcal{N}}$ and $\hat{\mathcal{M}}$ may not be the same because the nonlinear optimization problem can have several local minima. Moreover, RSS screening is easier to implement as one does not need to standardize the nonlinear functions $f_j(X_j, \hat{\boldsymbol{\alpha}}_j)$. We have found in our numerical analyses that RSS screening is more robust than MLSE screening, so it is adopted in our simulation studies and the data application. To determine a data-driven threshold ξ_n for RSS screening, one can use random permutation to create null models as in Fan, Feng, and Song (2011). An alternative thresholding scheme is to choose d covariates with the smallest marginal residual sum of squares. In case of large p_n, this thresholding approach can save a lot of computing effort by avoiding the random permutations.

The nonlinear independence screening (NLIS) procedure reduces the dimensionality of covariates from p_n to a possibly much smaller space with model size $|\hat{\mathcal{M}}|$. Following variable screening, we can apply a more refined model selection technique to choose important covariates in the nonlinear additive model conditional on the selected variable set $|\hat{\mathcal{M}}|$, and for this purpose, we use the nonnegative garrote for nonlinear additive models described in Section 2. Based on the philosophy in Fan and Lv (2008), independence screening should be applied when p_n is very large; while for moderately high p_n, one should use penalized methods for variable selection. In our case, the penalized model selection via the nonnegative garrote for nonlinear additive models requires good initial parameter estimates through NLS optimization and this is very difficult when p_n is large. For variable selection of (1.2), screening is necessary when p_n is only relatively high.

The crucial point for us is that the nonlinear independence screening procedure does not mistakenly miss important covariates. We show in the following section that our procedure has a sure screening property, as defined by Fan and Lv (2008), so that all important covariates are retained with probability tending to 1. This is an extension of Fan and Lv (2008) and Fan and Song (2010) because, in our case, the minimum distinguishable signal in selecting $\hat{\mathcal{M}}$ is closely related to the stochastic and numerical errors in estimating the nonlinear parameters. In addition, fitting marginal models to a joint regression can be considered as a type of model misspecification (White (1982)) as most of the covariates are dropped from the model fitting. The NLS objective function of a misspecified nonlinear model can be unstable, so our implementation is more difficult than that of linear models. The theoretical development requires assumptions that are appropriate to the nonlinear problem setting, such as local convexity.

3.1. Asymptotic properties

Let $\mathcal{M}_* = \{1 \leq j \leq p_n : \beta_{0j} \neq 0\}$ be the true sparse model with size $\nu_n = |\mathcal{M}_*|$, where $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p_n})^T$ and $\boldsymbol{\alpha}_0 = (\boldsymbol{\alpha}_{01}^T, \ldots, \boldsymbol{\alpha}_{0p_n}^T)^T$ are the true parameter values. Let $\boldsymbol{\gamma}_j = (\beta_j, \boldsymbol{\alpha}_j^T)^T$, $F_j(X_j, \boldsymbol{\gamma}_j) = \beta_j f_j(X_j, \boldsymbol{\alpha}_j)$, $\ell(\boldsymbol{\gamma}_j, X_j, Y) = [Y - F_j(X_j, \boldsymbol{\gamma}_j)]^2$, and $P_n \ell(\boldsymbol{\gamma}_j, \boldsymbol{x}_j, \boldsymbol{y}) = n^{-1} \sum_{i=1}^n [y_i - \beta_j f_j(x_{ij}, \boldsymbol{\alpha}_j)]^2$. Take

$$\boldsymbol{\gamma}_j^M = \{\beta_j^M, (\boldsymbol{\alpha}_j^M)^T\}^T = \arg\min_{\boldsymbol{\gamma}_j} \mathrm{E}\ell(\boldsymbol{\gamma}_j, X_j, Y), \tag{3.3}$$

for $j = 1, \ldots, p_n$, where E is expectation under the true model.

The following conditions are needed for the asymptotic properties of the nonlinear independence screening procedure.

(B1) There exist positive constants k_4 and k_5 such that $k_5 \leq \mathrm{E} f_j^2(X_j, \boldsymbol{\alpha}_j) \leq k_4$ for all $j = 1, \ldots, p_n$.

(B2) The marginal Fisher information:

$$\mathrm{I}_j(\boldsymbol{\gamma}_j) = \mathrm{E}\left[\frac{\partial F_j(X_j,\boldsymbol{\gamma}_j)}{\partial \boldsymbol{\gamma}_j}\frac{\partial F_j(X_j,\boldsymbol{\gamma}_j)}{\partial \boldsymbol{\gamma}_j^T}\right]$$

is finite and positive definite at $\boldsymbol{\gamma} = \boldsymbol{\gamma}_j^M$, for $j = 1,\ldots,p_n$.

(B3) There exists a constant $C_2 > 0$ such that for all $j = 1,\ldots,p_n$,

$$\sup_{\boldsymbol{\gamma}\in\Gamma,\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_j^M\|\le C_2}\left|\mathrm{E}\beta_j^2 f_j^2(X_j,\boldsymbol{\alpha}_j)I(|X_j|\ge K_n)\right| \le o(n^{-1}).$$

(B4) There exist some positive constants m_0, m_1, s_0, s_1, and $a > 1$, such that for any real t,

$$\mathrm{E}\exp[tf_j(X_j,\boldsymbol{\alpha}_j)] \le s_0\exp(m_0t^a), \quad j = 1,\ldots,p_n,$$
$$\mathrm{E}\exp(t\epsilon_i) \le s_1\exp(m_1t^a), \quad i = 1,\ldots,n.$$

(B5) For $j = 1,\ldots,p_n$, there exists a marginal least squares estimator (MLSE) $\hat{\boldsymbol{\gamma}}_j^M$ with $\frac{\partial P_n\ell(\boldsymbol{\gamma}_j,\boldsymbol{x}_j,\boldsymbol{y})}{\partial\boldsymbol{\gamma}_j}|_{\boldsymbol{\gamma}_j=\hat{\boldsymbol{\gamma}}_j^M} = 0$ such that $P_n\ell(\boldsymbol{\gamma}_j,\boldsymbol{x}_j,\boldsymbol{y})$ is locally convex in some neighborhood of $\hat{\boldsymbol{\gamma}}_j^M$, say $\mathcal{A}_{1\delta} = \{\boldsymbol{\gamma}_j\in\Gamma, \|\boldsymbol{\gamma}_j - \hat{\boldsymbol{\gamma}}_j^M\| \le \delta\}$ for some $\delta > 0$ and $\boldsymbol{\gamma}_j^M \in \mathcal{A}_{1\delta}$.

(B6) There exists a positive constant V_2 such that for all $\boldsymbol{\gamma}_j \in \mathcal{A}_{1\delta}$, we have $\mathrm{E}\ell(\boldsymbol{\gamma}_j, X_j, Y) \ge V_2\|\boldsymbol{\gamma}_j - \boldsymbol{\gamma}_j^M\|^2$, $j = 1,\ldots,p_n$.

(B7) If $\Sigma_{\boldsymbol{\alpha}} = \mathrm{cov}[\boldsymbol{f}(\boldsymbol{X},\boldsymbol{\alpha}_0), \boldsymbol{f}(\boldsymbol{X},\boldsymbol{\alpha}^M)]$ and $\lambda_{\max}(\Sigma_{\boldsymbol{\alpha}})$ is its maximum eigenvalue, there exist positive constants D_3 and D_4 such that $0 < D_3 \le \boldsymbol{\beta}_0^T\Sigma_{\boldsymbol{\alpha}}\boldsymbol{\beta}_0 \le D_4 < \infty$.

Condition (B1) is imposed to bound the nonlinear effects of the covariates so that their scales are of the same order. Condition (B4) is similar to Condition D in Fan and Song (2010) and is satisfied if $f_j(X_j,\boldsymbol{\alpha}_j)$ has a subgaussian distribution, i.e., $\mathrm{P}\exp[tf_j(X_j,\boldsymbol{\alpha}_j)] \le \exp(\tau^2t^2/2)$ for all real t, or equivalently, $\mathrm{P}\{|f_j(X_j,\boldsymbol{\alpha}_j)| \ge t\} \le c^*\exp\{-t^2/(2\tau^2)\}$ for some constant c^*. bounded symmetric distribution. Condition (B5) is the marginal version of Condition (A5). Our Theorem 3 gives a uniform convergence result for the MLSEs defined in (B5) and the sure screening property of MLSE screening, and Theorem 4 provides an asymptotic rate on the MLSE model size. These theorems are extensions of Theorems 4 and 5 in Fan and Song (2010) to the nonlinear additive model setting (1.2). Condition (B7) is parallel to Condition F in Fan and Song (2010). For Lemma 3 referenced in the following, see the supplementary materials.

Theorem 3. *Suppose* (A1), (A3), (A4), *and* (B1)−(B6) *hold.*

(i) *If* $n^{1-2\kappa}/(k_n^*)^2 \to \infty$ *for some constant* $0 < \kappa < 1/2$ *given in Lemma 3, then for any* $c_3 > 0$*, there exists a positive constant* c_4 *such that*

$$\begin{aligned} &P\Big(\max_{1\le j\le p_n} \|\hat{\boldsymbol\gamma}_j^M - \boldsymbol\gamma_j^M\| \ge c_3 n^{-\kappa}\Big) \\ &\quad\le p_n\Big\{\exp\Big[-\frac{c_4 n^{1-2\kappa}}{(k_n^*)^2}\Big] + n s_2 \exp[-m_2(K_n^*)^{a/(a-1)}]\Big\}, \end{aligned}$$

with $s_2 = s_0^{\nu_n} s_1$*.*

(ii) *Let* $\hat{\mathcal{N}}_{\zeta_n} = \{1 \le j \le p_n : |\hat\beta_j^M| \ge \zeta_n\}$*, where* ζ_n *is a predefined threshold value. If, in addition, conditions in Lemma 3 hold, then by taking* $\zeta_n = c_5 n^{-\kappa}$ *with* $c_5 \le c_2/2$ (c_2 *is given in Lemma* 3)*, we have*

$$P(\mathcal{M}_* \subset \hat{\mathcal{N}}_{\zeta_n}) \ge 1 - \nu_n\Big\{\exp\Big[-\frac{c_4 n^{1-2\kappa}}{(k_n^*)^2}\Big] + n s_2 \exp[-m_2(K_n^*)^{a/(a-1)}]\Big\}.$$

Theorem 4. *Under Conditions* (A1), (A3), (A4) *and* (B1)−(B7)*, for any* $\zeta_n = c_5 n^{-2\kappa}$*,*

$$\begin{aligned} &P[|\hat{\mathcal{N}}_{\zeta_n}| \le O\{n^{2\kappa}\lambda_{\max}(\Sigma_{\boldsymbol\alpha})\}] \\ &\quad\ge 1 - p_n\Big\{\exp\Big[-\frac{c_4 n^{1-2\kappa}}{(k_n^*)^2}\Big] + n s_2 \exp[-m_2(K_n^*)^{a/(a-1)}]\Big\}. \end{aligned}$$

Remark 1. Under Condition (B1), after some derivations, we can get that $\hat\beta_j^M$ and $\hat\beta_j'$ have the same order, with probability tending to 1 exponentially fast. Thus Theorem 3 and Theorem 4 also hold for $\hat\beta_j'$.

Remark 2. In order to obtain the optimal order of K_n^*, we balance the two terms in the upper bound of Theorem 3(i) and get $K_n^* = n^{(1-2\kappa)(a-1)/(3a-2)}$. It follows that

$$P\Big(\max_{1\le j\le p_n} \|\hat{\boldsymbol\gamma}_j^M - \boldsymbol\gamma_j^M\| \ge c_3 n^{-\kappa}\Big) = O\left\{p_n \exp\left(-c_4 n^{(1-2\kappa)a/(3a-2)}\right)\right\}.$$

Furthermore, if Y is bounded, then K_n^* can be taken as a finite constant. So

$$P\Big(\max_{1\le j\le p_n} \|\hat{\boldsymbol\gamma}_j^M - \boldsymbol\gamma_j^M\| \ge c_3 n^{-\kappa}\Big) = O\left\{p_n \exp\left(-c_4 n^{1-2\kappa}\right)\right\}.$$

In both cases, the tail probability in Theorem 3 has the exponential rate. Thus, similar to Fan and Song (2010), we can handle non-polynomial dimensionality with $\log p_n = o\left\{n^{(1-2\kappa)a/(3a-2)}\right\}$ or $\log p_n = o(n^{1-2\kappa})$.

Next we consider the sure screening property of the RSS screening (3.2). Let $\mathbf{R}^* = (\mathrm{R}_1^*, \ldots, \mathrm{R}_{p_n}^*)^T$, where $\mathrm{R}_j^* = \mathrm{E}\ell(\boldsymbol\gamma_0^M, X_j, Y) - \mathrm{E}\ell(\boldsymbol\gamma_j^M, X_j, Y)$ and

$\boldsymbol{\gamma}_0^M = \{0, (\boldsymbol{\alpha}_j^M)^T\}^T$. The empirical values of $\mathbf{R}^*$ can be written as $\mathbf{R}_n = (\mathrm{R}_{1,n}, \ldots, \mathrm{R}_{p_n,n})^T$, where $\mathrm{R}_{j,n} = P_n \ell(\hat{\gamma}_0^M, \boldsymbol{x}_j, \boldsymbol{y}) - P_n \ell(\hat{\boldsymbol{\gamma}}_j^M, \boldsymbol{x}_j, \boldsymbol{y})$ and $\hat{\gamma}_0^M = \{0, (\hat{\boldsymbol{\alpha}}_j^M)^T\}^T$, $j = 1, \ldots, p_n$. Obviously, we have $\ell(\boldsymbol{\gamma}_0^M, X_j, Y) = Y^2 \equiv [Y - 0 \cdot f_j(X_j, \boldsymbol{\alpha}_j^M)]^2$ and $\ell(\hat{\gamma}_0^M, x_j, y) = y^2 \equiv [y - 0 \cdot f_j(x_j, \hat{\boldsymbol{\alpha}}_j^M)]^2$. We sort the vector $\mathbf{R}_n$ in a descending order and select a set of variables:

$$\hat{\mathcal{M}}_{\xi_n} = \{1 \le j \le p_n : \mathrm{R}_{j,n} \ge \xi_n\},$$

where ξ_n is a predefined threshold value. We show that this RSS screening is equivalent to the MLSE screening in the sense that they both possess the sure screening property, and the numbers of selected variables of these two screening methods are of the same order.

Theorem 5. *Suppose that Conditions* (A1), (A3), (A4), (B1)−(B7) *and conditions in Lemma 3 hold. By taking* $\xi_n = c_7 n^{-2\kappa}$ *for a sufficiently small* $c_7 > 0$*, there exists a* c_8 *such that, with* $s_2 = s_0^{\nu_n} s_1$*,*

$$P(\mathcal{M}_* \subset \hat{\mathcal{M}}_{\xi_n}) \ge 1 - \nu_n \Big\{ \exp\Big[- \frac{c_8 n^{1-2\kappa}}{(k_n^*)^2} \Big] + n s_2 \exp[-m_2 (K_n^*)^{a/(a-1)}] \Big\}.$$

Theorem 6. *Under the same conditions as in Theorem* 5*, we have*

$$\begin{aligned} P[|\hat{\mathcal{M}}_{\xi_n}| &\le O\{n^{2\kappa} \lambda_{\max}(\Sigma_{\boldsymbol{\alpha}})\}] \\ &\ge 1 - p_n \Big\{ \exp\Big[- \frac{c_8 n^{1-2\kappa}}{(k_n^*)^2} \Big] + n s_2 \exp[-m_2 (K_n^*)^{a/(a-1)}] \Big\}. \end{aligned}$$

3.2. Iterative NLIS-NNG algorithm

Following Fan and Lv (2008) and Fan, Samworth, and Wu (2009), we propose a strategy to iteratively combine a large-scale variable screening and a moderate-scale model selection for the nonlinear additive regressions to further enhance the performance of the method in terms of false selection errors. The complete algorithm works as follows:

1. For every $j \in \{1, \ldots, p_n\}$, compute the marginal fit by solving

$$\min_{\beta_j, \boldsymbol{\alpha}_j} \sum_{i=1}^{n} [y_i - \beta_j f_j(x_{ij}, \boldsymbol{\alpha}_j)]^2. \tag{3.4}$$

Rank the covariates according to the marginal residual sum of squares

$$\mathrm{RSS}_j = \sum_{i=1}^{n} [y_i - \hat{\beta}_j^M f_j(x_{ij}, \hat{\boldsymbol{\alpha}}_j^M)]^2.$$

Select the top d covariates with the smallest RSS_j, or covariates with RSS_j smaller than a threshold ξ_n estimated from the random permutation. The set of selected covariates is denoted by $\mathcal{S}_1$.

2. Apply the NNG for the nonlinear additive model introduced in Section 2 on the set $\mathcal{S}_1$ to select a subset $\mathcal{M}_1$. The BIC score of the model with covariates in $\mathcal{M}_1$ is computed, denoted as BIC(1).
3. For every $j \in \mathcal{M}_1^c = \{1, \dots, p_n\} \setminus \mathcal{M}_1$, minimize

$$\sum_{i=1}^{n}[y_i - \sum_{l \in \mathcal{M}_1} \beta_l f_l(x_{il}, \boldsymbol{\alpha}_l) - \beta_j f_j(x_{ij}, \boldsymbol{\alpha}_j)]^2, \tag{3.5}$$

with respect to β_l, $\boldsymbol{\alpha}_l$, $l \in \mathcal{M}_1$ and β_j, $\boldsymbol{\alpha}_j$, $j \in \mathcal{M}_1^c$. This regression reflects the additional contribution of the jth covariate conditioning on the existence of the variable set $\mathcal{M}_1$. After marginally screening similar as in Step 1 by ranking the RSS of model (3.5), we choose a set of covariates $\mathcal{S}_2 \in \mathcal{M}_1^c$. The NNG procedure is then applied on the set $\mathcal{M}_1 \cup \mathcal{S}_2$ to select a subset $\mathcal{M}_2$.
4. Repeat Step 3 until $\mathcal{M}_k = \mathcal{M}_{k+1}$, or the size of $\mathcal{M}_k$ reaches a prespecified threshold d^*. The set of selected covariates is $\mathcal{M}_K$, where $K = \arg\min_k \mathrm{BIC}(k)$.

In the marginal screening step, the data-driven threshold ξ_n estimated from the random permutation can be computed as follows. Suppose that covariates in $\mathcal{M}_k$ have been recruited after k iterations. We randomly permute the rows of $\{X_j\}$ to yield $\{\tilde{X}_j\}$, $j \in \mathcal{M}_k^c$ and then minimize the null model

$$\sum_{i=1}^{n}[y_i - \sum_{l \in \mathcal{M}_k} \beta_l f_l(x_{il}, \boldsymbol{\alpha}_l) - \beta_j f_j(\tilde{x}_{ij}, \boldsymbol{\alpha}_j)]^2, \tag{3.6}$$

with respect to β_l, $\boldsymbol{\alpha}_l$, $l \in \mathcal{M}_k$, and β_j, $\boldsymbol{\alpha}_j$, $j \in \mathcal{M}_k^c$. The threshold ξ_n is chosen to be the q-th quantile of the RSS of the null models (3.6). In the simulation examples, we use $q = 0$, so ξ_n is the minimum value of the permuted RSS. An alternative screening scheme is to choose d covariates with the smallest marginal RSS. This strategy is computationally more efficient than the permutation-based screening. In addition, we show in Section 4 that this alternative thresholding scheme produces smaller numbers of false positives and also smaller prediction errors.

The proposed method can be extended to models with multiple possible parametric forms for covariates. For example, the additive components f_j in model (1.2) can be either linear or nonlinear. The marginal screening step in the above algorithm can be modified to include both the marginal nonlinear fit and the marginal linear fit for each covariate and the marginal RSS is then selected as the minimum value of the RSS of the linear and nonlinear marginal models.

The top d covariates with the smallest RSS_j are selected, denoted as $\mathcal{S}_1$. We then apply the NNG for nonlinear additive models to the model

$$y_i = \sum_{j\in\mathcal{S}_1} \beta_j^{(nlin)} f_j(x_{ij}, \boldsymbol{\alpha}_j) + \sum_{j\in\mathcal{S}_1} \beta_j^{(lin)} x_{ij} + \epsilon, \tag{3.7}$$

where $\beta_j^{(lin)}$ and $\beta_j^{(nlin)}$ are the coefficients for the linear and nonlinear components, respectively. In the implementation, the initial estimates of $(\beta_j^{(nlin)}, \boldsymbol{\alpha}_j^T, \beta_j^{(lin)})$ need to be computed carefully, as the variable selection procedure can be very sensitive to these initial values. One may add some restrictions on the parameter searching space, such as $\left|\beta_j^{(nlin)}\beta_j^{(lin)}\right| = 0$, or use multiple initial estimates to improve the efficiency and accuracy of the nonlinear optimization. A simulation example in which the additive regression model contains both linear and nonlinear components is included in Section 4.

4. Simulation Studies

In this section, we report on the numerical performance of our method in simulations. The simulation data were generated from the nonlinear additive model (1.2) without intercept. The covariates X_j were simulated according to the random effects model

$$X_j = 5\frac{W_j + tU}{1+t} - 2.5,\ j = 1, \dots, p_n,$$

where $W_1, \dots, W_p$ and U were i.i.d. random variables from Unif$(0,1)$. We used $t = 1$, so the covariates X_j were in $(-2.5, 2.5)$ with pairwise correlation equal to 0.5. The sample size was set to be $n = 100$.

In the first example, all f_j's were nonlinear functions, chosen to be the centered sigmoid function:

$$f_j(x, \alpha) = \frac{1}{1+e^{-\alpha x}} - 0.5.$$

We considered the following scenarios:

S1. $\beta_j = 3$, $1 \le j \le 4$ and $\beta_j = 0$, $j > 4$; α_j's chosen equidistantly on $[1.5, 4.2]$, $1 \le j \le 4$; $\epsilon_i \sim \mathcal{N}(0, 0.82^2)$;

S2. $p = 50$; $\beta_j = 3$, $1 \le j \le 10$ and $\beta_j = 0$, $j > 10$; α_j's chosen equidistantly on $[1.5, 4.2]$, $1 \le j \le 10$; $\epsilon_i \sim \mathcal{N}(0, 1.31^2)$;

The error term was chosen to give a signal-to-noise ratio (SNR) 3:1, defined as the square root of the ratio of the sum of nonzero components squared divided by the sum of residual squared. The total number of covariates p was 50 or 200. For scenario S1, we also considered $p = 1{,}000$ in order to illustrate the performance of our method under the ultra-high dimensional case.

Table 1. Average values of the numbers of true positives (TP) and false positives (FP) and the medians of the prediction errors (PE) for the first example. c-NLIS-NNG and p-NLIS-NNG refer to our proposed method with the direct cut-off screening scheme and the permutation-based screening scheme, respectively, and g-INIS refers to the iterative nonparametric independence screening method developed by Fan, Feng, and Song (2011). Robust standard deviations are given in parentheses.

Simulation Setting		Method	TP	FP	PE
S1	$p = 50$	c-NLIS-NNG	3.99 (0.02)	0.11 (0.19)	0.7070 (0.0848)
		p-NLIS-NNG	3.99 (0.02)	1.82 (1.11)	0.7443 (0.1017)
		g-INIS	4.00 (0.00)	0.93 (0.78)	0.9905 (0.1915)
	$p = 200$	c-NLIS-NNG	3.95 (0.09)	0.27 (0.40)	0.7646 (0.1087)
		p-NLIS-NNG	3.97 (0.05)	1.91 (0.97)	0.7753 (0.0946)
		g-INIS	3.98 (0.04)	1.17 (0.97)	1.0648 (0.1722)
	$p = 1000$	c-NLIS-NNG	3.90 (0.18)	0.52 (0.56)	0.7806 (0.1347)
		p-NLIS-NNG	3.89 (0.20)	2.67 (1.35)	0.8080 (0.1244)
		g-INIS	3.92 (0.15)	0.84 (0.85)	0.9906(0.2211)
S2	$p = 50$	c-NLIS-NNG	9.80 (0.35)	0.54 (0.64)	2.0932 (0.3186)
		p-NLIS-NNG	9.77 (0.37)	2.35 (1.31)	2.2965 (0.3428)
		g-INIS	7.90 (1.38)	1.23 (0.92)	4.0681 (1.0882)
	$p = 200$	c-NLIS-NNG	8.90 (1.26)	2.22 (1.42)	2.3630 (0.5701)
		p-NLIS-NNG	8.49 (1.39)	5.64 (2.30)	2.8248 (0.7689)
		g-INIS	4.58 (1.56)	2.81 (1.63)	9.0936 (2.5190)

We considered a second example in which the additive regression model contained both linear and nonlinear components:

$$Y = \sum_{j=1}^{3} \beta_j \left(\frac{1}{1+e^{-\alpha_j X_j}} - 0.5 \right) + \sum_{j=4}^{6} \beta_j X_j + \epsilon.$$

In this simulation, we not only needed to choose the important covariates for predicting the response, but also needed to determine the parametric form of each selected covariate, which in this case is either linear or sigmoid function. We set $p = 50$, $\beta_j = 3$, $\alpha_j = j + 1$, $1 \leq j \leq 3$, $\beta_j = 1$, $4 \leq j \leq 6$, $\beta_j = 0$, $j > 6$, and took $\epsilon_i \sim \mathcal{N}(0, 1)$. The SNR of this example was about 4:1.

We used cross-validation to select the tuning parameter λ in the nonnegative garrote and each of the above simulation settings was repeated 100 runs. The true positives (TP), false positives (FP) and prediction errors (PE) of the first example are summarized in Table 1. The prediction error was calculated on an independent test dataset of size $n/2$. The method c-NLIS-NNG in Table 1 refers to our proposed method with the direct cut-off screening scheme, where the top d covariates with the smallest marginal RSS are selected. Here we set

Table 2. Median of parameter estimates with median absolute deviation (MAD) in parentheses for simulation setting S1 using the method c-NLIS-NNG.

Parameters	$p = 50$	$p = 200$	$p = 1{,}000$
$\beta_1 = 3$	3.1195 (0.5488)	2.7985 (0.4935)	3.0162 (0.5555)
$\beta_2 = 3$	3.0299 (0.3982)	3.0381 (0.3947)	2.8314 (0.3595)
$\beta_3 = 3$	3.0567 (0.3249)	2.9814 (0.2179)	2.9052 (0.2567)
$\beta_4 = 3$	3.0495 (0.2105)	3.0743 (0.2385)	2.9783 (0.2343)
$\alpha_1 = 1.5$	1.5039 (0.4945)	1.6768 (0.5351)	1.3106 (0.3106)
$\alpha_2 = 2.4$	2.3375 (0.6541)	2.2749 (0.4777)	2.5577 (0.5984)
$\alpha_3 = 3.3$	3.4061 (0.7139)	3.3341 (0.6510)	3.5423 (0.7391)
$\alpha_4 = 4.2$	4.3842 (0.6158)	4.0815 (0.6841)	4.1484 (0.8510)

Table 3. Median of parameter estimates with median absolute deviation (MAD) in parentheses for simulation setting S2 using the method c-NLIS-NNG.

Parameters	$p = 50$	$p = 200$	Parameters	$p = 50$	$p = 200$
$\beta_1 = 3$	3.1313 (0.7505)	2.8141 (0.7785)	$\alpha_1 = 1.5$	1.5785 (0.5785)	1.6189 (0.6189)
$\beta_2 = 3$	3.1759 (0.8522)	2.9269 (0.7809)	$\alpha_2 = 1.8$	1.7264 (0.7208)	2.0048 (0.9166)
$\beta_3 = 3$	3.1636 (0.5707)	2.8909 (0.6961)	$\alpha_3 = 2.1$	1.9644 (0.7856)	2.4203 (1.1869)
$\beta_4 = 3$	2.7876 (0.4590)	2.9831 (0.7133)	$\alpha_4 = 2.4$	2.483 (1.1055)	2.6863 (1.0092)
$\beta_5 = 3$	3.0042 (0.5454)	3.1175 (0.6674)	$\alpha_5 = 2.7$	2.7308 (0.9203)	2.8353 (1.1193)
$\beta_6 = 3$	3.004 (0.3352)	2.9959 (0.5895)	$\alpha_6 = 3.0$	3.1168 (0.9187)	3.3602 (1.3157)
$\beta_7 = 3$	3.0353 (0.4776)	3.1148 (0.6541)	$\alpha_7 = 3.3$	3.6671 (1.0256)	3.0564 (1.3226)
$\beta_8 = 3$	2.9735 (0.4068)	3.1066 (0.3664)	$\alpha_8 = 3.6$	3.6248 (1.0440)	3.5123 (1.3032)
$\beta_9 = 3$	3.0312 (0.4187)	3.112 (0.4006)	$\alpha_9 = 3.9$	3.8169 (1.1831)	3.9751 (1.0249)
$\beta_{10} = 3$	3.1164 (0.4113)	2.9493 (0.4248)	$\alpha_{10} = 4.2$	4.6199 (0.3801)	4.6018 (0.3982)

$d = 1$ because it gives the smallest false selection rate. The method p-NLIS-NNG refers to the permutation-based screening scheme, which selects the covariates with RSS smaller than a data-driven threshold ξ_n estimated from the random permutation. We also compared the proposed method with the iterative nonparametric independence screening (INIS) method developed by Fan, Feng, and Song (2011). The INIS method was designed for the nonparametric additive model. It can be applied to the variable selection in the nonlinear additive model (1.2) where the nonlinear functions f_j's are assumed to be unknown and estimated nonparametrically. As suggested by Fan, Feng, and Song (2011), we use the greedy modification of the INIS algorithm (g-INIS) where only one covariate is recruited in each screening step.

For scenario S1, all three methods had comparable performances in terms of selecting the true variables, but the numbers of false positives of both p-NLIS-NNG and g-INIS were larger than that of c-NLIS-NNG. For the more challenging scenario S2, c-NLIS-NNG and p-NLIS-NNG both performed well in selecting the

Table 4. (a) Average values of the numbers of true positives (TP) and false positives (FP) and the medians of the mean squared errors (MSE) for the second example with both linear and nonlinear components. Robust standard deviations are given in parentheses. (b) Median of parameter estimates with median absolute deviation (MAD) in parentheses for the second example with both linear and nonlinear components.

	TP	FP	MSE
Combined	6.00 (0.00)	0.17 (0.40)	0.6564 (0.0994)
Nonlinear	2.37 (0.71)	0.76 (0.78)	
Linear	2.26 (0.77)	0.78 (0.79)	

True	Estimate	True	Estimate	True	Estimate
$\beta_1 = 3$	3.0989 (0.4574)	$\beta_4 = 1$	0.9952 (0.0981)	$\alpha_1 = 2$	1.9741 (0.5687)
$\beta_2 = 3$	2.9050 (0.3134)	$\beta_5 = 1$	1.0009 (0.0607)	$\alpha_2 = 3$	3.2921 (0.7273)
$\beta_3 = 3$	2.9623 (0.2646)	$\beta_6 = 1$	0.9709 (0.0706)	$\alpha_3 = 4$	4.2079 (0.9409)

true variables. However, g-INIS tended to miss about two important variables when $p = 50$ and miss about half of the important variables when $p = 200$. From the perspective of the prediction error, c-NLIS-NNG outperformed the other two methods in all cases. In Tables 2 and 3, we report the parameter estimations of the method c-NLIS-NNG conditioned on the parameters being selected. We conclude from this simulation study that ignoring the information of knowing the nonlinear functions f_j and estimating them nonparametrically can lead to inefficient predictions and sometimes even seriously biased variable selection results. This simulation also indicates that the direct cut-off screening scheme works better than the permutation-based screening scheme in keeping down the false positives. We used the direct cut-off screening for the data application in Section 5.

For the second example with both linear and nonlinear components, we can see from Table 4 (a) that the selected models by our method always contain all the nonzero components. The average numbers of true positives and false positives under the nonlinear and linear categories indicate both the selection accuracy as well as the accuracy of determining the corresponding parametric forms. Table 4 (b) displays the parameter estimations conditioned on the parameters being selected.

For the more challenging scenarios, S3 and S4, our method tends to miss some important covariates, but the selection accuracy is still within the reasonable range. For Example 2 with both linear and nonlinear components, the selected models by our method always contain all the nonzero components. The average numbers of true positives and false positives under the nonlinear and linear categories indicate both the selection accuracy as well as the accuracy of determining the corresponding parametric forms. Tables 2 and 3 display the parameter estimations conditioning on that the parameters are selected.

5. Application to Identify Gene Regulations

The regulation of gene expressions depends on the recognition of specific promoter sequences by transcriptional regulatory proteins and it often occurs through the coordinated action of multiple transcriptional regulators. Chen et al. (2004) used the nonlinear additive model (1.2) to model the transcription rate of a target gene as a combination of a set of regulatory functions from relevant regulators. Here the response Y is the transcription rate of the target gene; the covariate X_j is the gene expression of the j-th regulator; the intercept β_0 is the basal expression of the target gene; and the coefficient β_j is the regulatory capability from the j-th regulator. The regulatory function of the j-th regulator is a sigmoid function:

$$f_j\{X_j(t), \alpha_j\} = \frac{1}{1 + \exp\{-\alpha_j[X_j(t) - M_j]\}}, \tag{5.1}$$

where α_j is the transition rate and M_j is the mean expression level of regulator j. The goal is to select the important regulators from the potential set of regulators with $\beta_j \neq 0$.

We illustrate our proposed method using the dataset reported by Rabani et al. (2011). Using the short metabolic labeling of RNA with 4-thiouridine (4sU), the authors were able to distinguish recently transcribed RNA from the overall RNA and therefore obtained direct measurements of RNA transcription rate. The expression of RNA-total and RNA-4sU were measured for 254 representative signature genes during the response of mouse dendritic cells to lipopolysaccharide (LPS). There are 13 measurements for each gene, obtained at 15-min intervals over the first 3 hours after LPS stimulation. We focused on 44 genes that have been identified to have varying degradation rates and were interested in finding genes that related to these genes using model (1.2). The response is the RNA-4sU expression of a gene from these 44 genes and the predictors are the RNA-total expressions of 254 signature genes. The data were normalized by the mean of the 8 control genes (Ppp2r1a, Ndufs5, Psma7, Tomm7, Psmb4, Ndufa7, Eif4h, Capza1) and then standardized to have mean 0 and variance 1. So $M_j = 0$ in equation (5.1). We also centered the sigmoid function (5.1) as in the simulation studies so there is no intercept in model (1.2).

We applied the proposed method to this dataset, where $n = 13$ and $p = 254$. The numbers of selected predictors for the 44 genes were between 1 and 3. Figure 1 displays the model fits for 6 randomly chosen genes. The selected regulators for these 6 genes and the corresponding parameter estimates and leave-one-out cross validation prediction errors are listed in Table 5. We found some of our

Table 5. Selected regulators and the corresponding parameter estimates and leave-one-out cross validation prediction errors (CVPE) for six randomly chosen genes.

Target genes	Regulators $(\hat{\beta}_j, \hat{\alpha}_j)$				CVPE
Il6	Il17ra	(2.0400, 1.1731)	Dusp2	(4.1435, 0.7378)	0.3314
Il7r	Cebpb	(-2.5906, 3.2332)	Hmga1	(2.1319, 2.1482)	1.3638
Zfp36	Zfp36	(-0.3415, 1.6778)	Btg2	(9.3860, 0.4893)	0.0410
Icosl	Icosl	(6.8418, 1.8038)	tgif1	(-3.0583, 3.3271)	0.2570
Jun	Btg2	(3.3999, 1.3200)	Rilpl1	(-0.4657, 3.7681)	0.2509
Ifnb1	Nr4a1	(-2.5196, 2.4291)	Myd116	(8.2476, 1.0578)	0.0929

data-driven implications of gene regulations consistent with previous biological findings. For example, Zfp36 is a known regulator of RNA stability that destabilizes its mRNA targets by binding AU-rich elements in their 3'UTR and it is known to autoregulate the stability of its own mRNA (Lai et al. (2006)), which is in line with with our finding that Zfp36 negatively regulates itself. Sadik et al. (2011) found in their studies that the expression of several pro-inflammatory mediators, including Il6 were decreased in Il17ra$^{-/-}$ mice (mice lacking Il17ra) compared to wild-type mice, consistent with our result that Il17ra positively regulates Il6. We also identified some interesting genes, like Cebpb and Icosl, both of which are important regulators in immune and anti-inflammatory processes. Evidently our method can provide biological investigators with a targeted list of genes, which could be very useful in subsequent studies.

6. Discussion

In the mstatistical literature, the commonly adopted approach to nonlinearity is through nonparametric modeling. However, the parametric approach should not be neglected. The parameters in the nonlinear regression model can offer important biological implications. For example, the transition rate α_j in the sigmoid function (5.1) describes how fast the effect of a gene saturates. For small α_j, the regulatory effect is close to linear, while for large α_j, function (5.1) is close to a step function, indicating that a small deviation from the mean expression value of the gene leads to a large effect on the response, but this effect quickly stabilizes as the deviation increases. In practice, one can refine the parameter estimates using nonlinear least squares once the important covariates are determined, given the estimates from the nonnegative garrote as initial values. The inference of the selected model, such as the confidence intervals of the parameter estimates, can be carried out following the theories of nonlinear regression models (Jennrich (1969); Malinvaud (1970); Wu (1981)).

In our nonlinear regression model (1.2), the nonlinear functions $\boldsymbol{f}$ are assumed to be known, which requires some prior knowledge about the relations

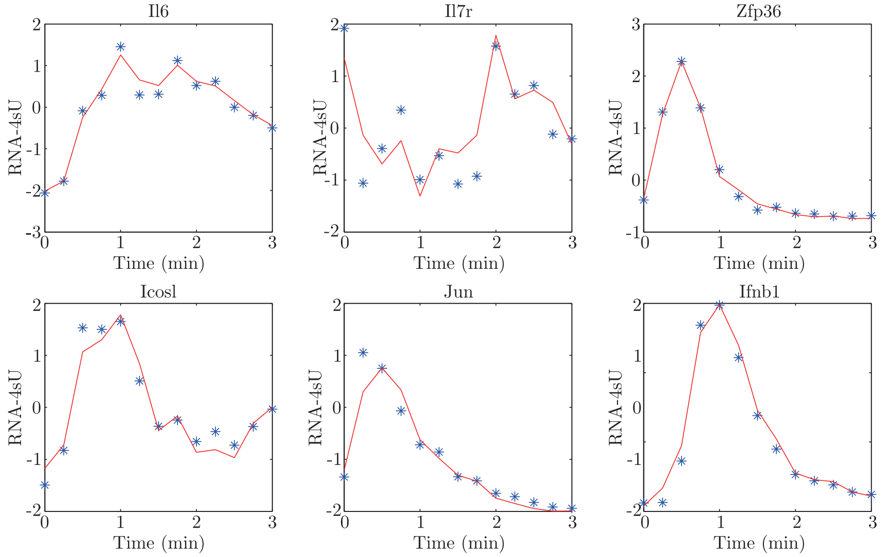

Figure 1. Standardized measurements (star) overlaid with the model fits (solid line) for six randomly chosen genes.

between the covariates and the response. Here we chose the sigmoid function that is widely used in gene regulatory networks, to illustrate our method. However, this is a general methodology that is also applicable to other nonlinear functions, such as the hill function and the gamma function. An important extension of our method is to include interactions of the covariates in the regression model. Another interesting problem is to relax the additive assumption in model (1.2) and allow more complicated nonlinear relationships. These topics are to be studied in the future.

References

Antoniadis, A. and Fan, J. (2001). Regularization of wavelet approximations. *J. Amer. Statist. Assoc.* **96**, 939-967.

Biedermann, S. and Woods, D. C. (2011). Optimal designs for generalized non-linear models with application to second-harmonic generation experiments. *J. Roy. Statist. Soc. C* **60**, 281-299.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37**, 373-384.

Candes, E. and Tao, T. (2007). The dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.* **35**, 2313-2351.

Chen, H.-C., Lee, H.-C. Lin, T.-Y. Li, W.-H. and Chen, B.-S. (2004). Quantitative characterization of the trancriptional regulatory network in the yeast cell cycle. *Bioinformatics* **20**, 1914-1927.

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. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32**, 928-961.

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

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with np-dimensionality. *Ann. Statist.* **38**, 3567-3604.

Frank, I. and Friedman, J. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109-148.

Golub, G. and Pereyra, V. (2003). Separable nonlinear least squares: the variable projection method and its applications. *Inverse Problems* **19**, R1-R26.

Gorst-Rasmussen, A. and Scheike, T. (2013). Independent screeing for single-index hazard rate models with ultrahigh dimensional features. *J. Roy. Statist. Soc. Ser. B* **75**, 217-245.

Hall, P. and Miller, H. (2009). Using generalized correlation to effect variable selection in very high dimensional problems. *J. Comput. Graph. Statist.* **18**, 533-550.

Hall, P., Titterington, D. M. and Xue, J.-H. (2009). Tilting methods for assessing the influence of components in a classifier. *J. Roy. Statist. Soc. Ser. B* **71**, 783-803.

Huang, J., Horowitz, J. L. and Ma, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.* **36**, 587-613.

Huang, J., Horowitz, J. L. and Wei, F. (2010). Variable selection in nonparametric additive models. *Ann. Statist.* **38**, 2282-2313.

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Ann. Math. Statist.* **40**, 633-643.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. and Barabási, A.-L. (2000). The large-scale organization of metabolic networks. *Nature* **407**, 651.

Jiang, X., Jiang, J. and Song, X. (2012). Oracle model selection for nonlinear models based on weighted composite quantile regression. *Statist. Sinica* **22**, 1479-1506.

Kim, Y., Choi, H. and Oh, H.-S. (2008). Smoothily clipped absolute deviation on high dimensions. *J. Amer. Statist. Assoc.* **103**, 1665-1673.

Kosmidis, I. and Firth, D. (2009). Bias reduction in exponential family nonlinear models. *Biometrika* **96**, 793-804.

Lai, W., J. Parker, S. Grissom, D. stumpo, and P. Blackshear (2006). Novel mrna targets for tristetraprolin (ttp) identified by global analysis of stabilized transcripts in ttp-deficient fibroblasts. *Molecular and Cellular Biology* **26**, 9196-9208.

Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *J. Amer. Statist. Assoc.* **107**, 1129-1139.

Lin, Y. and Zhang, H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.* **34**, 2272-2297.

Malinvaud, E. (1970). The consistancy of nonlinear regressions. *Ann. Math. Statist.* **41**, 956-969.

Meier, L., V. Geer, and P. Bühlmann (2009). High-dimensional additive modeling. *Ann. Statist.* **37**, 3779-3821.

Mestl, T., Plahte, E. and Omholt, S. (1995). A mathematical framework for describing and analysing gene regulatory networks. *J. Theoret. Biol.* **176**, 291-300.

Rabani, M., Levin, J., Fan, L., Adiconis, X., Raychowdhury, R. Garber, M., Gnirke, A., Nusbaum, C., Hacohen, N., Friedman, N., Amit, I. and Regev, A. (2011). Metabolic labeling of rna uncovers principles of rna production and degradation dynamics in mammalian cells. *Nature Biotechnology* **29**, 436-442.

Ravikumar, R., Liu, H., Lafferty, J. and Wasserman, L. (2009). Sparse additive models. *J. Roy. Statist. Soc. Ser. B* **71**, 1009-1030.

Ruhe, A. and P.-Å. Wedin (1980). Algorithm for separable nonlinear least squares problems. *SIAM Rev.* **22**, 318-337.

Sadik, C., Kim, N., Alekseeva, E. and Luster, A. (2011). Il-17ra signaling amplifies antibody-induced arthritis. *PLoS Biology* **6**, e26342.

Seber, G. A. F. and C. J. Wild (2003). *Nonlinear Regression.* Wiley, Hoboken, New Jersey.

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, H. (2012). Factor profiled sure independence screening. *Biometrika* **99**, 15-28.

Wei, B. C. (1998). *Exponential Family Nonlinear Models.* Springer-Verlag, Singapore.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1-26.

Wu, C.-F. (1981). Asymptotic theory of nonlinear least squares estimation. *Ann. Statist.* **9**, 501-513.

Xue, H., Miao, H. and Wu, H. (2010). Sieve estimation of constant and time-varying coefficients in nonlinear ordinary differential equation models by considering both numerical error and measurement error. *Ann. Statist.* **38**, 2351-2387.

Xue, L. and Zou, H. (2011). Sure independence screening and compressed random sensing. *Biometrika* **98**, 371-380.

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

Yuan, M. and Lin, Y. (2007). On the nonnegative garrote estimator. *J. Roy. Statist. Soc. Ser. B* **69**, 143-161.

Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.* **36**, 1567-1594.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.* **7**, 2541-2567.

Zhu, L.-P., Li, L., Li, R. and Zhu, L.-X. (2012). Model-free feature screening for ultrahigh-dimensional data. *J. Amer. Statist. Assoc.* **106**, 1464-1475.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14642, USA.

E-mail: shuang_wu@urmc.rochester.edu

Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14642, USA.

E-mail: hongqi_xue@urmc.rochester.edu

Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA.

E-mail: ywu11@ncsu.edu

Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14642, USA.

E-mail: hulin_wu@urmc.rochester.edu

(Received November 2012; accepted September 2013)